IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 14 Issue 11
November 2023

ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

SAI
www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

Thank you for Sharing Wisdom!

Kohei Arai
Editor-in-Chief
IJACSA
Volume 14 Issue 11 November 2023
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

Sentiment-Driven Forecasting LSTM Neural Networks for Stock Prediction-Case of China Bank Sector

Shangshang Jin
Department of Art and Science, Johns Hopkins University, Washington, D.C., United States

Abstract*—This study explores the predictive analysis of public sentiment in China's financial market, focusing on the banking sector, through the application of machine learning techniques. Specifically, it utilizes the Baidu Index and Long Short-Term Memory (LSTM) networks. The Baidu Index, akin to China's version of Google Trends, serves as a sentiment barometer, while LSTM networks excel in analyzing sequential data, making them apt for stock price forecasting. Our model integrates sentiment indices from Baidu with historical stock data of significant Chinese banks, aiming to unveil how digital sentiment influences stock price movements. The model's forecasting prowess is rigorously evaluated using metrics such as R-squared (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and confusion matrices, the latter being instrumental in assessing the model's capability in correctly predicting stock up or down movements. Our findings predominantly showcase superior prediction performance of the sentiment-based LSTM model compared to a standard LSTM model. However, effectiveness varies across different banks, indicating that sentiment integration enhances prediction capabilities, yet individual stock characteristics significantly contribute to the prediction accuracy. This inquiry not only underscores the importance of integrating public sentiment in financial forecasting models but also provides a pioneering framework for leveraging digital sentiment in financial markets. Through this endeavor, we offer a robust analytical tool for investors, policymakers, and financial institutions, aiding in better navigation through the intricate financial market dynamics, thereby potentially leading to more informed decision-making in the digital age.

Keywords*—Machine learning; LSTM; sentiment; forecasting; banking sector

I. Introduction

The digital epoch ushered in a plethora of tools and platforms, with search engines emerging as crucial conduits for gauging public sentiment. Shiller [1] accentuates the importance of grounding research in actual human behavior to comprehend how individuals genuinely think and act. This sentiment concurs with the dynamics of the digital realm, where search query fluctuations concerning financial entities can unveil early signs of shifting public interest or potential concerns. These subtle shifts, characteristic of the digital age, harbor the potential to impact stock market trends [2].

This study endeavors to tap into public sentiment within China's financial market, probing how sentiments harvested from digital platforms can act as a predictive lens for scrutinizing and foreseeing market dynamics. Central to this exploration is the Baidu index, a tool often drawn parallel to China's version of Google, emblematic of the digital transition. Baidu, surpassing its initial function as a search engine, has evolved into a dynamic digital nexus, reflecting the collective sentiment of its extensive user base. This discernment elevates tools like the Baidu index to a significant pedestal in contemporary financial prediction frameworks, particularly against the backdrop of digital transformation sweeping through banks and financial institutions [3].

Banks are pivotal players in China’s economic narrative, entrusted with capital provisioning, trade facilitation, and financial stability maintenance [4]. The banking sphere, headlined by colossal entities like the Industrial and Commercial Bank of China (ICBC) and the Agricultural Bank of China (ABC), commands substantial influence over China's economic fabric. In the digital era, their roles and impact have evolved, with stock performance mirroring not only the health of individual institutions but also the broader economic vitality [5]. Observing the stock trends of these institutions can significantly inform both domestic economic strategies and global financial outlooks. Traditional stock prediction methodologies, shackled by inherent constraints, often grapple with accuracy challenges, paving the path for Neural Networks—specifically, the Long Short-Term Memory (LSTM) networks—renowned for their prowess in time series predictions owing to their capability to process long-term data dependencies.

Nonetheless, the nuanced realm of stock market predictions often finds the exclusive reliance on raw financial metrics inadequate. The digital epoch grants stakeholders a platform to voice their opinions, assumptions, and concerns. Tapping into this vast sentiment reservoir, particularly reflected in search patterns on platforms like Baidu, can unveil precious insights. The fusion of sentiment analysis with LSTM models ushers in a novel frontier for stock prediction, especially within crucial sectors like China's banking domain.

This inquiry embarks on an innovative journey, integrating sentiment indices from the Baidu platform with LSTM neural networks to refine the prediction framework for China's banking stocks. Our scrutiny especially shines a light on stalwarts like Industrial and Commercial Bank of China (ICBC), Agricultural Bank of China (ABC), Bank of Communications (BoCom), CCB (China Construction Bank),

China Merchants Bank (CMB), Shanghai Pudong Development Bank (SPDB), and Industrial Bank (IB). Overlaying sentiment metrics with historic stock data, we endeavor to ascertain the predictive potency of sentiments, gauged via the Baidu index, in molding future stock valuations within the banking sector.

This endeavor distinctively enriches the literature by elucidating the intertwined roles of banks in financial stability, regulatory landscapes, investor interests, and the digital transformation of the banking sector. Accurate forecasting of bank stock prices enhances our comprehension of impending financial scenarios and vulnerabilities. By leveraging sentiment analytics, we aim to foresee shifts in the continually evolving regulatory labyrinth banks traverse, equipping stakeholders with insights into potential policy alterations. Recognizing the magnetic appeal the banking sector exudes for investors, our model strives to equip them with precise forecasting tools. Additionally, as the banking sector navigates through transformative digital innovations, our sentiment-driven approach seeks to decipher public perceptions, offering insights into how fintech advancements might reshape bank stock trajectories.

Our exposition is organized as follows: Section II delves into pertinent literature, mapping the journey of sentiment-imbued stock predictions. Section III unravels our methodology, elucidating the integration of Baidu's sentiment indicators with LSTM networks. In Section IV, we disseminate our empirical findings, accentuating the merits and limitations of our model. Finally, Section V encapsulates a reflective summary and contemplates potential directions for future exploration in this domain.

II. Related Work

Research posits that investor sentiment plays a predictive role in cross-sectional stock returns, suggesting that the collective mood of investors can drive stock price fluctuations and impact their anticipated returns [6, 7, 8]. Moreover, an increasing body of literature is exploring the influence of Internet search activities on financial markets, delving into how online information-seeking behaviors may reflect or shape investor sentiment and, in turn, market dynamics.

Li et al. [9] marked one of the initial strides into understanding how financial news articles sway stock prices. Their contention was that while many had mapped the influence of news in a bag-of-words paradigm, the more profound underlying sentiment—bridging the chasm between lexical patterns and stock fluctuations—remained an undercharted territory. By implementing sentiment dictionaries, their work enriched the sentiment space, and the results affirmed that sentiment-infused models yield superior accuracy over the conventional bag-of-words counterparts.

Pivoting to the realm of digital social interactions, Guo et al. [10] embarked on an exploration capitalizing on user comments from Xueqiu, a focal social networking site in the Chinese stock market landscape. Their research sought to dynamically unravel the web of interactions between investor sentiment and the stock market. One salient revelation was that sentiment data's efficacy as a predictive tool is tethered to the degree of public attention the stock garners. This insight underscores the dynamism and contingent nature of sentiment's predictive power.

Smales in [11] delved deeper into the mechanics of how investor sentiment could cause asset prices to deviate from their fundamental equilibrium. He asserted that understanding these deviations is paramount, given their potential ramifications on capital allocation and its costs. Through a comprehensive analysis spanning a quarter-century, he demonstrated that sentiment indicators, exemplified by the Volatility Index (VIX), forge a robust bond with stock returns. Further granularity revealed that some stocks, especially those in tech or smaller capitalization brackets, display heightened susceptibility to sentiment fluctuations. Intriguingly, recessional periods amplify this sentiment-driven volatility.

Bringing sophisticated machine learning into the fold, Xing et al. [12] championed the sentiment-aware volatility forecasting (SAVING) model. Their venture sought to elucidate the bidirectional dance between asset prices and market sentiment. This synthesis of deep learning and sentiment analysis manifested in results that eclipsed traditional forecasting tools and rival neural network architectures in terms of precision.

Lastly, the ubiquity of social media platforms in shaping stock predictions has witnessed academic spotlight. Swathi et al. [13] echoed the notion that platforms like Twitter, given their massive user engagement, have metamorphosed into sentiment barometers. By fusing the Teaching and Learning Based Optimization (TLBO) technique with Long Short-Term Memory (LSTM) frameworks, they demonstrated an unprecedented accuracy in forecasting stock price movements based on social media sentiment.

III. Methodology

This section elucidates the systematic approach employed to forecast the stock prices of significant Chinese banks, leveraging sentiment analysis and LSTM Neural Networks. We initiate with an introduction to LSTM and its aptness for time series forecasting and subsequently incorporate sentiment data derived from the Baidu Index. Acknowledging the latent impact of sentiment on stock market behavior, the Baidu Index is introduced as a lagged variable. Employing grid search, the model's hyperparameters are optimized. Evaluation metrics such as R^2, RMSE, MAE, and MAPE are then utilized to determine the model's predictive accuracy.

A. Long Short-Term Memory (LSTM) Networks

LSTM (Long Short-Term Memory) networks, proposed by Hochreiter and Schmidhuber [14], are a specialized subset of Recurrent Neural Networks (RNN), meticulously crafted to grasp long-term dependencies within sequential datasets. Their intrinsic capability to retain patterns across extended sequences renders them particularly proficient for tasks such as stock price prediction.

Refer to Fig. 1 for a visualization of the LSTM model, structured as follows:

- 1 Input Layer: Accepts the sequence of stock prices and sentiment indices.

- 2 Hidden Layers: Composed of LSTM units that can remember patterns in data over long periods.
- 1 Output Layer: Produces the predicted stock price.

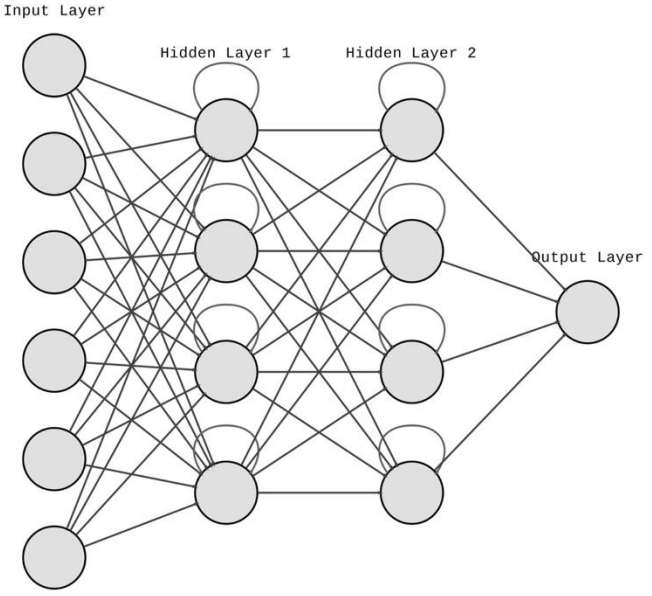


Fig. 1. LSTM structure.

Mathematically, the operations of an LSTM unit are encapsulated as:

$$\begin{aligned} f_t &= \sigma\big(W_f \cdot [h_{t-1}, X_t] + b_f\big) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \\ \tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, X_t] + b_C) \\ C_t &= f_t \times C_{t-1} + i_t \times \tilde{C}_t \\ o_t &= \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \\ h_t &= o_t \times \tanh(C_t) \end{aligned} \tag{1}$$

where:

- $f_t, i_t, \tilde{C}_t, C_t, o_t, h_t$ are the forget gate, input gate, cell candidate, cell state, output gate, and hidden state at time t respectively.
- W_f, W_i, W_C, W_o and b_f, b_i, b_C, b_o are the weights and biases for the respective gates.
- σ is the sigmoid function.

B. Sentiment Index

In this paper, the Baidu index, a popular sentiment index in China, gauges the public sentiment by analyzing search query frequencies and user behaviors on the Baidu search engine. To harmonize the scale between stock prices and the Baidu Index during pre-processing, Min-Max scaling is applied. Recognizing the potential latency in sentiment influence on stock prices, a time-lagged version of the Baidu Index is integrated. This latency captures the inherent delay in investor reactions, attributable to the time taken to interpret and act upon sentiment-driven information, thereby encapsulating market inertia.

For model input augmentation, at each time step t, our LSTM model receives a concatenated input vector, represented as:

$$X_t = [P_t, B_t] \tag{2}$$

Where:

- P_t is the stock price at time t.
- B_t represents the Baidu Sentiment Index at time t.

The LSTM, characterized by its inherent memory, updates its cell state c, factoring in both the stock prices and sentiment index, thereby accounting for sentiment influence across sequences. During the LSTM's forward propagation, this merged input not only impacts immediate outputs but also modifies the hidden states. Consequently, the sentiment information guides predictions and tweaks the LSTM's internal memory. The chosen loss function, Mean Squared Error (MSE), benchmarks model forecasts against real stock prices. During backpropagation, the gradients, inherently influenced by the Baidu Index, fine-tune the model's weightings throughout its training phase.

C. Grid Search Optimization

Grid search is an exhaustive search approach employed for hyperparameter tuning, ensuring optimal model performance. By systematically working through multiple combinations of hyperparameter sets, it evaluates which combination gives the best performance based on a scoring technique [15, 16].

In the optimization process of the LSTM model, several pivotal hyperparameters are meticulously evaluated. The batch size, dictating the number of samples processed before a model update, emerges as a key factor under consideration. The efficacy of various optimizers, notably RMSprop and Adam, is also examined [17]. These algorithms adjust the model's weights in alignment with the data and its corresponding loss function. To fortify the model against overfitting, dropout layers are integrated within the LSTM's architecture [18]. The dropout rate specifies the percentage of neurons randomly nullified during each training phase, enhancing model generalizability. A detailed visualization of the grid search, illustrating the synergy of these hyperparameters and their bearing on model performance, is provided in Fig. 2.

D. Forecasting Evaluation Criterias

To gauge the accuracy and reliability of the model's forecasts, followed by Paudel et al, [19], several statistical metrics are employed:

R^2 (Coefficient of Determination):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \tag{3}$$

RMSE (Root Mean Squared Error):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2} \tag{4}$$

MAE (Mean Absolute Error):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \left|Y_i - \hat{Y}_i\right| \tag{5}$$

MAPE (Mean Absolute Percentage Error):

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \quad (6)$$

where Y_t is the actual value at time $t, \hat{Y}_t$ is the predicted value at time , and $\bar{Y}$ represents the mean of the actual values.

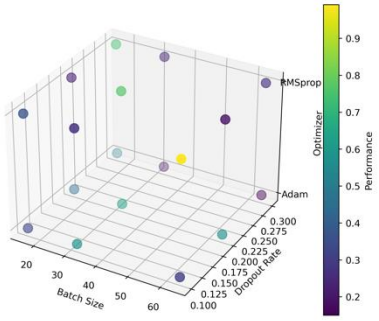


Fig. 2. 3D visualization of grid search.

E. Summary

The proposed technique for predicting stock prices by incorporating the Sentiment Index employs a systematic process, seamlessly integrating the power of Long Short-Term Memory (LSTM) networks and the insights from the Sentiment Index.

Initially, input data comprising of stock prices and a one-day lagged Sentiment Index is collected. This data undergoes normalization to ensure that the range of values is consistent, thereby facilitating effective training. A time-series dataset is then constructed to ensure sequences of data are fed to the LSTM network, aiding in the capture of temporal dependencies.

The dataset gets bifurcated into training and test sets. The LSTM model, designed with two LSTM layers and a dense output layer, is structured to efficiently incorporate and process this information. The first LSTM layer, activated by the tanh function and equipped with dropout, returns sequences which are fed into the second LSTM layer. This layer not only processes the sequences but also integrates the Sentiment Index data. The dense layer at the end aids in producing the final prediction.

Upon constructing this model, it undergoes compilation wherein various aspects like optimizer, loss function, and evaluation metrics are defined. To determine the most optimal parameters for training, a grid search technique is employed. This process considers parameters such as optimizer choice, dropout rate, and batch size, ensuring the model's robustness and accuracy. The model, trained with the best parameters identified by the grid search, is then used to make predictions on the test set.

Once predictions are made, they are de-normalized to revert them to their original scale. Finally, these predictions are evaluated using several metrics to determine the model's efficacy. The metrics include RMSE, MAE, MAPE, and R^2.

The entire technique, encapsulated in the flowchart (see Fig. 3), presents a comprehensive approach to effectively predict stock prices by harnessing the dual power of LSTM networks and sentiment indices.

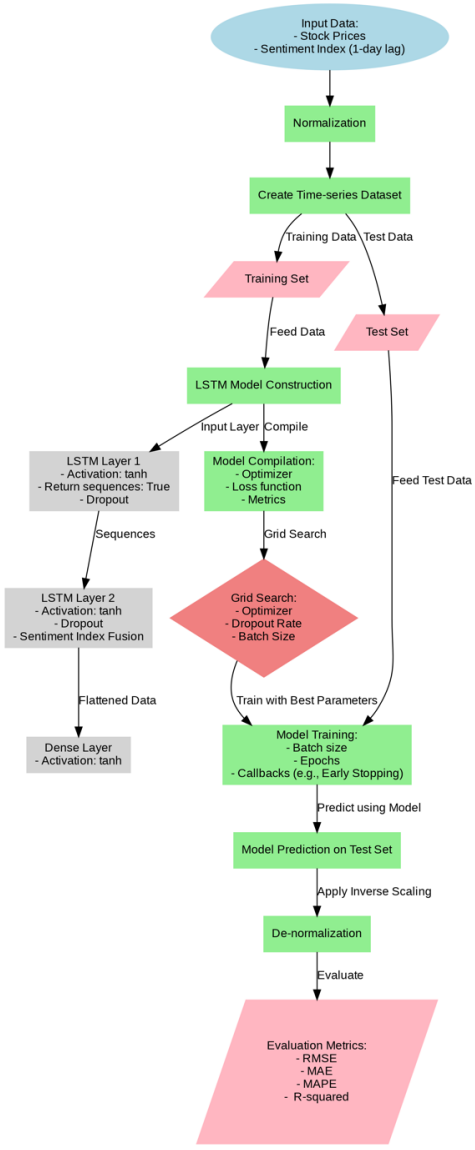


Fig. 3. Flowchart of the proposed technique.

IV. Numerical Results

A. Data Description

The study employs daily stock prices from August 1, 2022, to August 1, 2023, of seven eminent banks listed on the Shanghai Stock Exchange, including ICBC, ABC, BoCom, CCB, CMB, SPDB, and IB (see Fig. 4). This data, sourced from Yahoo Finance, assures precision. Considering Baidu's significant presence in China, its index is a robust reflection of the public's sentiment towards these banks, sentiment indicators is derived from the Baidu index specific to each bank depicted in Fig. 5. For modeling purposes, the data is split: 80% for training and 20% for testing. Preprocessing steps involve data normalization to ensure consistency.

Fig. 4. Stock price.

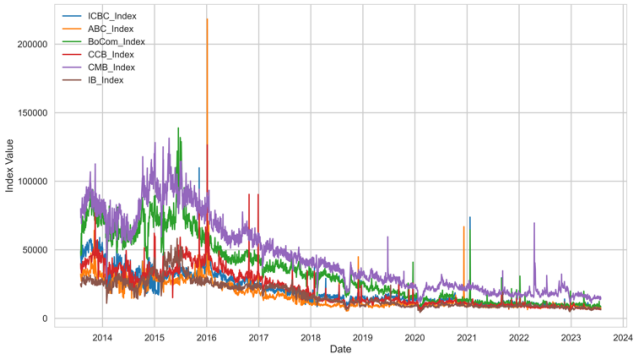


Fig. 5. Stock sentiment index.

B. Hyperparameter Determination

Using grid search, various combinations of batch size, dropout rate, and optimizer type are examined. Each bank's data undergoes this analysis individually, catering to the distinct behavior of each bank's stock prices. The optimal hyperparameters obtained are delineated below in Table I.

C. Forecasting Performance

Table II showcases the forecasting performance of both the LSTM and Sentiment-integrated LSTM (SB-LSTM) models for each bank. For ICBC, the SB-LSTM model exhibits a remarkable increase in R^2 by 6.67% compared to the standard LSTM, and the RMSE also drastically reduces from 0.0510 to 0.0107. In the case of ABC, the R^2 in the SB-LSTM model increases by 2.01% and sees a significant reduction in RMSE from 0.0411 to 0.0185. BoCom's results indicate that the SB-LSTM has an RMSE of 0.0338, an improvement from 0.0553 as observed in the standard LSTM.

Interestingly, for CCB, the standard LSTM outperforms the SB-LSTM in terms of R^2 and RMSE metrics. This suggests that the integration of sentiment does not consistently lead to improved forecasts for every stock. For CMB, the SB-LSTM model dramatically reduces the RMSE from 0.7648 to 0.2316. Similarly, IB's forecasting with the SB-LSTM exhibits a substantial enhancement in R^2, marking an increase of 7.29% compared to the LSTM.

TABLE I. OPTIMAL HYPERPARAMETERS FOR EACH BANK

Stock	Hyperparameter	LSTM	SB-LSTM
ICBC	Batch Size	16	16
	Dropout rate	0.3	0.1
	Optimizer	Adam	Adam
ABC	Batch Size	16	16
	Dropout rate	0.1	0.2
	Optimizer	Adam	Adam
BoCom	Batch Size	16	16
	Dropout rate	0.3	0.3
	Optimizer	Adam	Adam
CCB	Batch Size	16	16
	Dropout rate	0.1	0.3
	Optimizer	Adam	Adam
CMB	Batch Size	16	16
	Dropout rate	0.1	0.3
	Optimizer	Adam	Adam
IB	Batch Size	16	16
	Dropout rate	0.1	0.2
	Optimizer	Adam	Adam

TABLE II. MODEL PERFORMANCE

Stock	Model	R^2	RMSE	MAE	MAPE
ICBC	LSTM	93.03%	0.0510	0.0410	0.9805%
	SB-LSTM	99.70%	0.0107	0.0091	0.2197%
ABC	LSTM	97.48%	0.0411	0.0295	1.0268%
	SB-LSTM	99.49%	0.0185	0.0171	0.6091%
BoCom	LSTM	98.69%	0.0553	0.0379	0.8234%
	SB-LSTM	99.51%	0.0338	0.0320	0.7278%
CCB	LSTM	95.99%	0.0598	0.0419	0.7625%
	SB-LSTM	93.35%	0.0771	0.0767	1.4282%
CMB	LSTM	98.55%	0.7648	0.5662	1.5184%
	SB-LSTM	99.87%	0.2316	0.1973	0.5328%
IB	LSTM	92.66%	0.3340	0.2675	1.6014%
	SB-LSTM	99.95%	0.0274	0.0229	0.1372%

In essence, while the SB-LSTM model generally showcases superior accuracy and precision in forecasting for most stocks, there are notable exceptions, such as CCB. This underscores the idea that sentiment integration typically enhances prediction capabilities, but the individual characteristics of stocks must be taken into account.

D. Classification Analysis

Fig. 6 delineates the confusion matrices for the LSTM and SB-LSTM models across six prominent banks: ICBC, BoCom, ABC, CCB, CMB, and IB. For the ICBC stock, the LSTM model yielded 138 true negatives and 84 true positives but was marred by 131 false positives and 132 false negatives.

Conversely, the SB-LSTM showed a marked improvement, boasting 247 true negatives, 212 true positives, and significantly fewer false positives and negatives. A similar trend was observed for BoCom, where the SB-LSTM displayed enhanced accuracy with 232 true negatives and 227 true positives, overshadowing the LSTM's 134 true negatives and 102 true positives. ABC's LSTM model presented 166 true negatives and 58 true positives, but the SB-LSTM again outperformed by producing 242 true negatives and 189 true positives. This pattern of SB-LSTM superiority persisted with CCB, where the LSTM's 144 true negatives and 86 true positives were eclipsed by the SB-LSTM's 251 true negatives and 210 true positives. For CMB, while the LSTM model recorded 144 true negatives and 111 true positives, the SB-LSTM surged ahead with 257 true negatives and 220 true positives. Lastly, IB's results from the LSTM, comprising 132 true negatives and 96 true positives, were also surpassed by the SB-LSTM's impressive 257 true negatives and 225 true positives.

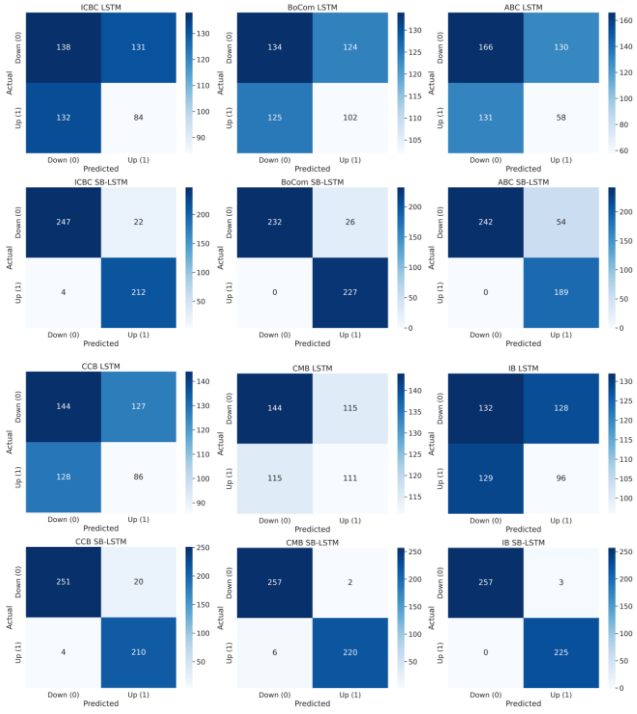


Fig. 6. Confusion matrix.

In summary, the SB-LSTM models consistently demonstrated higher classification accuracy across all stocks when juxtaposed with the LSTM models, emphasizing their superior capability in predicting stock price movements.

V. Conclusion

This study ventured into the realm of sentiment-driven stock prediction, specifically within China's banking sector, utilizing a Sentiment-integrated LSTM (SB-LSTM) model. Our findings predominantly showcased the SB-LSTM model's superior accuracy over traditional LSTM in forecasting stock prices, echoing the potential of sentiment analysis in financial forecasting as suggested by previous studies [20, 21, 22]. Particularly, the use of the Baidu Index highlighted how sentiment data can enhance predictive accuracy. The consistent enhancement in classification accuracy across most stocks by the SB-LSTM models further corroborates the promise of integrating sentiment analysis with LSTM networks for more precise stock price forecasting. While the use of the Baidu Index effectively demonstrated how sentiment data can enhance predictive accuracy, a key limitation lies in the reliance on a single sentiment data source, which may not capture the full spectrum of market sentiments. Additionally, the model's efficacy outside the Chinese banking sector remains untested, suggesting the need for further research to evaluate its applicability across different sectors and geographical regions. This endeavor not only enriches the literature but also paves the way for future explorations in leveraging unstructured sentiment data for financial forecasting amidst the digital transformation sweeping across the financial sector. This transformation is increasingly recognized as crucial in the evolving landscape of financial analytics.

References

[1] Shiller, R. J. (2014). Speculative Asset Prices. *American Economic Review*, *104*(6), 1486–1517. https://doi.org/10.1257/aer.104.6.1486

[2] Li, X., Xie, H., Lau, R. Y. K., Wong, T.-L., & Wang, F.-L. (2018). Stock Prediction via Sentimental Transfer Learning. *IEEE Access*, *6*, 73110–73118. https://doi.org/10.1109/ACCESS.2018.2881689

[3] Lang, Q., Wang, J., Ma, F., Huang, D., & Mohamed Ismail, M. W. (2023). Is Baidu index really powerful to predict the Chinese stock market volatility? New evidence from the internet information. *China Finance Review International*, *13*(2), 263–284.

[4] Fang, J., Lau, C.-K. M., Lu, Z., Tan, Y., & Zhang, H. (2019). Bank performance in China: A perspective from bank efficiency, risk-taking and market competition. *Pacific-Basin Finance Journal*, *56*, 290–309. https://doi.org/10.1016/j.pacfin.2019.06.011

[5] Peng, Z.-Y., & Guo, P.-C. (2022). A Data Organization Method for LSTM and transformer when predicting Chinese banking stock prices. *Discrete Dynamics in Nature and Society*, *2022*, 1–8. https://doi.org/10.1155/2022/7119678

[6] Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications*, *42*(1), 306–324. https://doi.org/10.1016/j.eswa.2014.08.004

[7] Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, *114*, 128–147. https://doi.org/10.1016/j.knosys.2016.10.003

[8] Li, Q., Wang, J., Wang, F., Li, P., Liu, L., & Chen, Y. (2017). The role of social sentiment in stock markets: A view from joint effects of multiple information sources. *Multimedia Tools and Applications*, *76*(10), 12315–12345. https://doi.org/10.1007/s11042-016-3643-4

[9] Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, *69*, 14–23. https://doi.org/10.1016/j.knosys.2014.04.022

[10] Guo, K., Sun, Y., & Qian, X. (2017). Can investor sentiment be used to predict the stock price? Dynamic analysis based on China stock market. *Physica A: Statistical Mechanics and Its Applications*, *469*, 390–396. https://doi.org/10.1016/j.physa.2016.11.114

[11] Smales, L. A. (2017). The importance of fear: Investor sentiment and stock market returns. *Applied Economics*, *49*(34), 3395–3421. https://doi.org/10.1080/00036846.2016.1259754

[12] Xing, F. Z., Cambria, E., & Zhang, Y. (2019). Sentiment-aware volatility forecasting. *Knowledge-Based Systems*, *176*, 68–76. https://doi.org/10.1016/j.knosys.2019.03.029

[13] Swathi, T., Kasiviswanath, N., & Rao, A. A. (2022). An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis. *Applied Intelligence*, *52*(12), 13675–13688. https://doi.org/10.1007/s10489-022-03175-2

[14] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[15] Kao, L.-J., Chiu, C.-C., Lu, C.-J., & Yang, J.-L. (2013). Integration of nonlinear independent component analysis and support vector regression for stock price forecasting. *Neurocomputing*, 99, 534–542. https://doi.org/10.1016/j.neucom.2012.06.037

[16] Luo, S., & Tian, C. (2020). Financial high-frequency time series forecasting based on sub-step grid search long short-term memory network. *IEEE Access*, *8*, 203183–203189. https://doi.org/10.1109/access.2020.3037102

[17] Mahardika T, N. Q., Fuadah, Y. N., Jeong, D. U., & Lim, K. M. (2023). PPG signals-based blood-pressure estimation using grid search in hyperparameter optimization of CNN–LSTM. *Diagnostics*, 13(15), 2566. https://doi.org/10.3390/diagnostics13152566

[18] Priyadarshini, I., & Cotton, C. (2021). A novel LSTM–CNN–grid search-based deep neural network for sentiment analysis. *The Journal of Supercomputing*, *77*(12), 13911–13932. https://doi.org/10.1007/s11227-021-03838-w

[19] Paudel, S., Pudasaini, A., Shrestha, R. K., & Kharel, E. (2023). Compressive strength of concrete material using machine learning techniques. *Cleaner Engineering and Technology*, *15*, 100661. https://doi.org/10.1016/j.clet.2023.100661

[20] Huang, M. Y., Rojas, R. R., & Convery, P. D. (2019). Forecasting stock market movements using Google Trend searches. *Empirical Economics*, *59*(6), 2821–2839. https://doi.org/10.1007/s00181-019-01725-1

[21] Liang, C., Tang, L., Li, Y., & Wei, Y. (2020). Which sentiment index is more informative to forecast stock market volatility? evidence from China. *International Review of Financial Analysis*, *71*, 101552. https://doi.org/10.1016/j.irfa.2020.101552

[22] Ito, T., Masuda, M., Naito, A., & Takeda, F. (2021). Application of google trends-based sentiment index in exchange rate prediction. *Journal of Forecasting*, *40*(7), 1154–1178. https://doi.org/10.1002/for.2762

Semantic Sampling: Enhancing Recommendation Diversity and User Engagement in the Headspace Meditation App

Rohan Singh Rajput*, Christabelle Pabalan*, Akhil Chaturvedi*, Prathamesh Kulkarni, Adam Brownell
Headspace
Los Angeles, USA

***Abstract*—In this paper, we present a clever approach to enhance the performance of sequential recommendation systems, specifically in the context of meditation recommendations within the Headspace app. Our method, termed "Semantic Sampling", leverages the power of language embeddings and clustering techniques to introduce diversity and novelty in the recommendations. We augment the Time Interval Aware Self-Attention for Sequential Recommendation (TiSASRec) model with semantic sampling, where the next recommended item is randomly sampled from a cluster of semantically similar items. Our empirical evaluation, conducted on a sample set of 276,700 users, reveals a statistically significant increase of 2.26 % in content start rate for the treatment group (TiSASRec with semantic sampling) compared to the control group (TiSASRec alone). Furthermore, our approach demonstrates improved coverage and rarity, indicating a broader range of recommendations and higher novelty. The results underscore the potential of Semantic Sampling in enhancing user engagement and satisfaction in recommendation systems.**

***Keywords*—Information retrieval; machine learning; recommender system.**

I. Introduction

Recommendation systems have become an indispensable part of modern digital applications, providing users with personalized content and product suggestions [1]. These systems are especially important in the realm of wellness applications, where personalization can enhance user engagement and satisfaction significantly. Headspace, for instance, offers a rich variety of content covering meditation, sleep, focus, and music, among others [2].

However, a common challenge for these systems is the 'long tail problem'. A significant number of items are rarely recommended, resulting in limited diversity in the recommendations [3]. This lack of diversity can curb user exploration and engagement, especially in areas where exploring new content is beneficial. In the context of wellness applications like Headspace, users gain exposure to a diverse range of content, aiding them in discovering new techniques, sustaining interest, and deepening their practice.

Unlike platforms like TikTok or YouTube, where content consumption can be sporadic and unplanned, Headspace users often embark on sequential mindfulness journeys through the content. Recognizing this unique behavior, we found TiSASRec, a sequential recommendation system built on the transformer architecture, to be an apt baseline for our approach [4]. While TiSASRec outperformed earlier models, it started to loop back to repetitive content recommendations after extended training. This tendency to lean towards shorter, frequently accessed content over longer, seldom accessed content limited the scope of recommendations and potentially restricted user exploration.

To amplify recommendation diversity while retaining relevance, we integrated semantic sampling into the TiSASRec model. Semantic sampling involves extracting language embeddings from the titles and teasers of content using sentence transformers. After these embeddings are secured, we compute their cosine similarities. Instead of merely relying on TiSASRec's recommendations, we select content from the cluster of the N most semantically similar items [5]. This approach not only broadens the array of recommendations but ensures they remain pertinent to the user, enhancing the discovery experience.

By melding sequential recommendations with semantic sampling, we present a robust solution to the long tail problem. This strategy augments the diversity of recommendations while ensuring their relevance to individual users. By blending the strengths of both sequential recommendation systems and semantic sampling, our goal is to elevate the user experience in wellness platforms like Headspace.

II. Method

To tackle the long-tail problem in recommendation systems, we introduced a technique called semantic sampling. This approach leverages language embeddings and cosine similarity to offer diverse and engaging content recommendations. We conducted extensive online A/B testing, evaluating various metrics to validate the efficacy of our approach. The results from the A/B test provided crucial insights into the real-world performance of our recommendation system.

A. Headspace App

Headspace is a popular mindfulness and meditation app that offers a wide variety of content to its users. The content is organized into different modules, each focusing on a specific topic or theme. The app provides recommendations to users on the "Today" tab, which is the main landing page of the app. The recommendations are personalized based on the user's past interactions and preferences [6].

*Corresponding authors.

The app also provides recommendations in specific modules. For example, in the "Sleep" module, the app recommends sleep-focused content, while in the "Meditation" module, it recommends meditation-focused content. The goal of these recommendations is to provide users with content that is relevant and engaging, enhancing their overall experience with the app. This is further exemplified in Fig. 1.

B. Semantic Sampling

Semantic sampling is an approach we introduced to enhance the diversity of recommendations. The method employs the extraction of language embeddings from content titles and teasers using the Language-agnostic BERT Sentence Embedding (LaBSE) transformer model [7].

LaBSE is a multilingual sentence encoder, trained on a broad corpus of bilingual sentence pairs. It produces language-agnostic sentence embeddings, ensuring that sentences with equivalent meanings across different languages are proximal in the embedding space [7]. Such a property becomes invaluable for our use case, enabling the comparison and identification of similarities between content, irrespective of language barriers.

Our choice of LaBSE was driven by its excellent performance in paraphrasing similarity tasks, vital for our objective of discerning semantically similar content pieces. The illustration of Semantic Sampling workflow is explained in the Fig. 2.

Semantic sampling proceeds through the following stages:

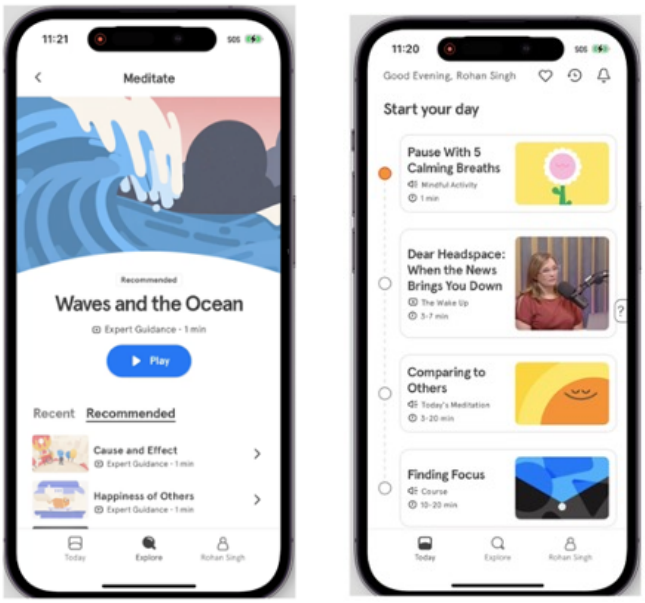
Content Embedding: We commence by extracting the language embeddings from content titles and teasers using LaBSE. This yields a high-dimensional vector representation for every content item.

Similarity Calculation: Cosine similarity between these embeddings is computed. This metric quantifies the cosine of the angle between two vectors, representing a measure of their alignment and, by extension, their similarity. Its value spectrum ranges from -1 (entirely dissimilar) to 1 (perfectly similar).

Recommendation Refinement: Instead of directly using the recommendation produced by the TiSASRec model, we assess the similarity of this chosen content with all other content items. Content pieces falling below a cosine similarity threshold of 0.75 to the chosen content are filtered out. From the remaining, more similar content pieces, one is selected randomly. For this process, there's an 80% likelihood that the content will be resampled using semantic sampling.

The TiSASRec model underpins our recommendation system. A transformer-centric model, it has been empirically validated to adeptly capture users' sequential behaviors [8]. The model utilizes the transformer framework, rooted in self-attention mechanisms, to emulate the sequential tendencies of users. This endows it with the capability to understand both immediate and extended user preferences, rendering it apt for our application.

For our endeavors, the TiSASRec model was the source of initial recommendations, which subsequently received enhancement through our semantic sampling technique.



(a) Recommended Meditation (b) Today Tab

Fig. 1. Screenshots of the Headspace app showcasing our recommendation system. (a) The 'Recommended for You' section in the Meditation module displays personalized suggestions. (b) The 'Today Tab' features dynamic content shelves, each filled with diverse recommendations from our system for an engaging experience.

$$\text{Semantic Sampling}(c) = \arg\max_{c' \in C} \frac{1}{N} \sum_{i=1}^{N} \text{sim}(c, c'_i) \quad (1)$$

Here, c symbolizes the content piece put forth by the TiSASRec model, C embodies the collection of all content items, N stands for the count of the most similar items taken into account, and $\text{sim}(c, c'_i)$ signifies the cosine similarity between the embeddings of content pieces c and c'_i.

By leveraging this methodology, we facilitate a richer spectrum of recommendations. This not only bolsters the user's exploratory experience but also ensures the continued relevance of recommendations to individual users.

C. Metrics for Evaluation

In the context of recommender systems, research has traditionally focused on the precision of the recommendations. However, it has been recognized that other recommendation qualities—such as whether the list of recommendations is diverse and whether it contains novel items—may have a significant impact on the overall quality of a recommender system. Consequently, the focus of recommender systems research has shifted to include a wider range of 'beyond accuracy' objectives [9]. These metrics include coverage, entropy, rarity, and intra-list diversity (ILD).

1) Coverage: Coverage is a measure of the proportion of items in the catalog that the recommender system can suggest. It provides an understanding of how well the recommendations cover the available items. A higher coverage indicates that the recommender system is capable of suggesting a wider

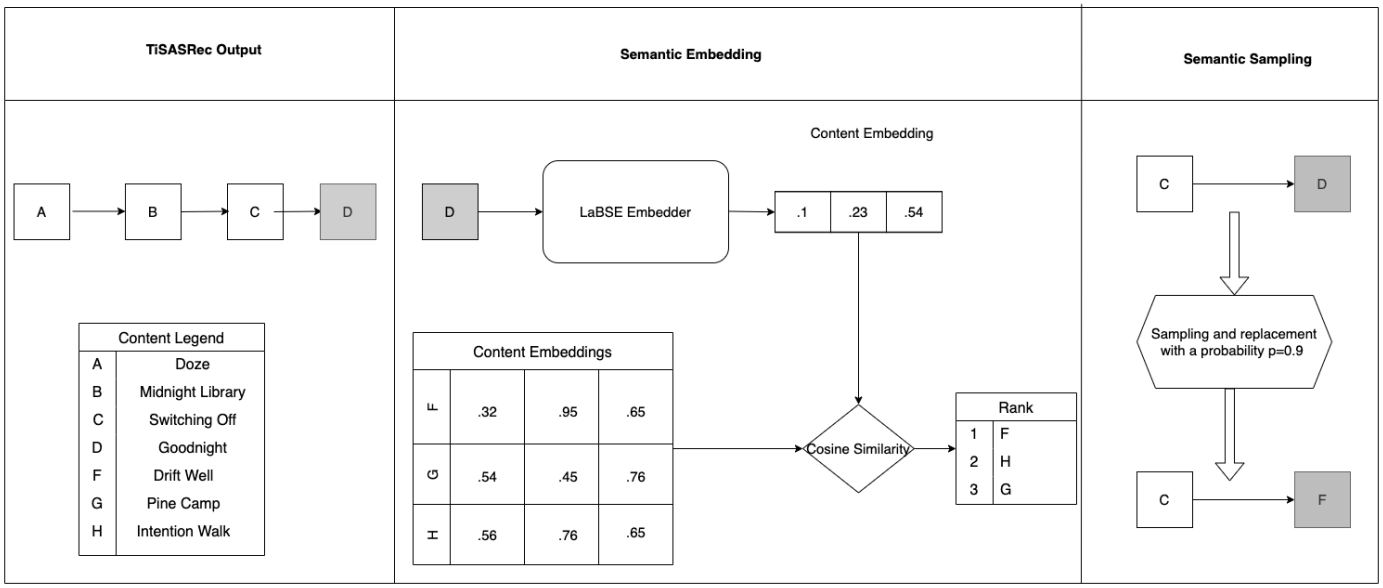


Fig. 2. Semantic Sampling Workflow. (a) The left side illustrates the input-output sequence of the TiSASRec model. (b) The central block depicts the generation of the LaBSE Embedding. (c) The right block demonstrates the replacement of the TiSASRec output with content that has been semantically embedded and sampled.

variety of items, which can contribute to a more diverse and personalized user experience [10]. It can be calculated as follows:

$$Coverage = \frac{|I_{rec}|}{|I|} \tag{2}$$

where, I_{rec} is the set of items recommended and I is the total set of items.

2) Entropy: Entropy is a measure of the unpredictability or randomness of the recommendations. It is derived from information theory, where it is used to quantify the amount of information contained in a set of data. In the context of recommender systems, a higher entropy indicates a more diverse set of recommendations, as it suggests that the recommendations are spread out over a larger number of different items. Optimal entropy is achieved when the recommendation distribution is uniformly distributed, therefore, an increase in entropy signifies an improvement for the long-tail problem [11]. It can be calculated as follows:

$$Entropy = -\sum_{i=1}^{|I|} p(i) \log p(i) \tag{3}$$

where, $p(i)$ is the probability of item i being recommended.

3) Rarity: Rarity is a measure of how uncommon or unique the recommended items are. It is defined as the inverse of normalized popularity, with 0 being our most viewed content and 1 being our least viewed content. A higher rarity score indicates that the recommender system is suggesting more unique or less popular items, which can contribute to a more diverse set of recommendations [12]. It can be calculated as follows:

$$Rarity = 1 - \frac{pop(i)}{max_{j \in I} pop(j)} \tag{4}$$

where, $pop(i)$ is the popularity of item i, and $max_{j \in I} pop(j)$ is the popularity of the most popular item.

4) Intra-List Diversity (ILD): Intra-List Diversity (ILD) is a metric that measures the average dissimilarity between all pairs of items within a recommendation list. It is a measure of the diversity of the recommendations and is defined as follows:

$$ILD(L) = \frac{2}{|L| \cdot (|L| - 1)} \sum_{i=1}^{|L|} \sum_{j=i+1}^{|L|} d(i, j) \tag{5}$$

where, L is the list of recommended items, $|L|$ is the number of items in the list, and $d(i, j)$ is the dissimilarity between items i and j. The dissimilarity between items can be calculated using various methods, such as cosine distance in the embedding space. A higher ILD value indicates a more diverse set of recommendations [13].

5) Content Click-Through Rate: Content Click-Through Rate (CTR) is a measure of user engagement with the recommended content. It is defined as the number of times users initiate interaction with the content divided by the number of times the content is displayed to the users. This metric directly reflects the user's interaction with the recommended content, providing a clear measure of the effectiveness of the recommendations. An increase in this rate is indicative of users finding the recommendations more engaging and relevant [14]. It can be calculated as follows:

$$CTR = \frac{Clicks}{Impressions} \tag{6}$$

where $Clicks$ is the number of times users initiate interaction with the content, and $Impressions$ is the number of times the content is displayed to the users.

D. Online Analysis

For our online analysis, we conducted an A/B test on 138.4K control users (TiSASRec only) and 138.3K treatment users using the StatSig platform. StatSig is a platform that provides robust statistical analysis for A/B testing, ensuring that our results are statistically significant and reliable. This A/B test was run for a period of 42 days.

A key metric we focused on was the change in average content starts across the entire Headspace App. The content start rate, a widely accepted measure of user engagement, is defined as the number of times users initiate interaction with the content divided by the number of times the content is displayed to the users. This metric was chosen because it directly reflects the user's interaction with the recommended content, providing a clear measure of the effectiveness of the recommendations. An increase in this rate is indicative of users finding the recommendations more engaging and relevant [14].

To gain a deeper understanding of the changes in diversity brought about by our semantic sampling approach, we also calculated the entropy, ILD, coverage, and rarity for the clicks originating from both TiSASRec and the treatment model. These metrics were computed for the entire app, providing a comprehensive evaluation of the impact of semantic sampling on the quality of recommendations as described above.

This comprehensive online analysis allowed us to assess the real-world performance of our semantic sampling approach, providing valuable insights into its effectiveness in enhancing the diversity of recommendations while maintaining user engagement which is our primary goal.

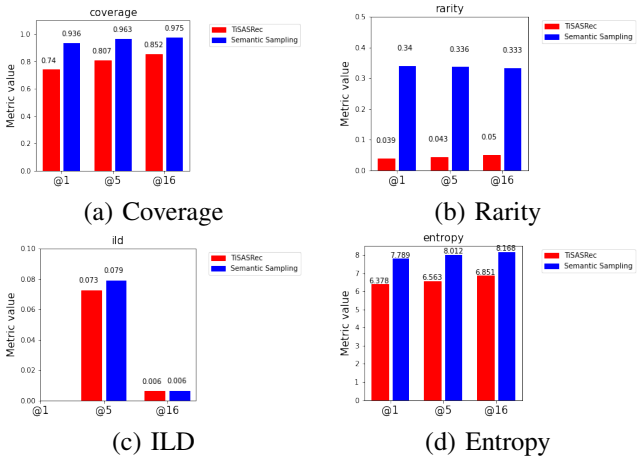
III. Results

In this section, we present the results of our study comparing the performance of the baseline TiSASRec model with the enhanced approach using semantic sampling. We evaluated the two models across various metrics to assess the impact on recommendation diversity and user engagement. Additionally, we discuss the findings from the online A/B test, which provided insights into the real-world effectiveness of our semantic sampling approach to solve the long tail problem.

A. Semantic Sampling Improves Recommendation Diversity

We conducted extensive evaluations to compare the performance of TiSASRec with our semantic sampling approach based on results from the online experiment. The results of these comparisons across different metrics are exemplified in Table I and Fig. 3.

1) Coverage: We first examined the coverage metric, which measures the proportion of items in the catalog that the recommender system is able to suggest. The results showed that the semantic sampling approach significantly outperformed TiSASRec across all top-k rankings (Coverage@1, Coverage@5, and Coverage@16). For example, at Coverage@1, semantic sampling achieved a coverage of 0.936, representing a substantial increase of 26.67% compared to TiSASRec's coverage



(a) Coverage (b) Rarity

(c) ILD (d) Entropy

Fig. 3. Four figures depicting various recommender system metrics. (a) Coverage illustrates the algorithm's range of potential recommendations. (b) Rarity indicates the uniqueness of recommendations. (c) ILD represents the average dissimilarity between recommended items. (d) Entropy quantifies the information in a stochastic process.

of 0.740. This indicates that the semantic sampling approach recommended a wider variety of items to users, enhancing their opportunity for content discovery and engagement.

2) Entropy: The entropy metric measures the unpredictability or randomness of the recommendations. Higher entropy values suggest a more diverse set of recommendations. Our results revealed that the semantic sampling approach significantly increased the entropy of recommendations compared to TiSASRec. For instance, at Entropy@1, the mean entropy of semantic sampling was 7.789, which was 22.09% higher than TiSASRec's mean entropy of 6.378. Similarly, at Entropy@5 and Entropy@16, semantic sampling demonstrated improvements of 22.09% and 19.23%, respectively. These results indicate that the semantic sampling approach generated more diverse and less predictable recommendations, fostering a richer and more engaging user experience.

3) Rarity: The rarity metric measures how uncommon or unique the recommended items are. A higher rarity score indicates that the recommender system suggests more unique or less popular items, contributing to a more diverse set of recommendations. The semantic sampling approach significantly increased the rarity scores compared to TiSASRec for all top-k rankings (Rarity@1, Rarity@5, and Rarity@16). For example, at Rarity@1, semantic sampling achieved a score of 0.340, representing a remarkable increase of 765.88% compared to TiSASRec's score of 0.039. These results further confirm that semantic sampling effectively promotes the discovery of less frequently recommended content items.

4) Intra-List Diversity (ILD): Intra-List Diversity (ILD) quantifies the average dissimilarity between all pairs of items within a recommendation list. Our semantic sampling approach achieved a modest improvement in ILD at top-k ranking values of ILD@5 and ILD@16, with a 8.82% and 2.84% increase, respectively. These results indicate a positive impact on the diversity of recommendations.

TABLE I. COMPARISONS BETWEEN TISASREC AND SEMANTIC SAMPLING ACROSS DIFFERENT METRICS

	TiSASRec		Semantic Sampling		% Change (Mean)
Metric	Mean	Std	Mean	Std	Mean
Coverage@1	0.740	0.015	**0.936**	0.008	**+26.670%**
Coverage@5	0.807	0.010	**0.963**	0.004	**+18.520%**
Coverage@16	0.852	0.008	**0.975**	0.002	**+13.550%**
Entropy@1	6.378	0.039	**7.789**	0.026	**+22.090%**
Entropy@5	6.563	0.026	**8.012**	0.019	**+22.090%**
Entropy@16	6.851	0.019	**8.168**	0.013	**+19.230%**
Rarity@1	0.039	0.001	**0.340**	0.007	**+765.880%**
Rarity@5	0.043	0.001	**0.336**	0.006	**+688.120%**
Rarity@16	0.050	0.001	**0.333**	0.006	**+562.830%**
ILD@1	NaN	NaN	NaN	NaN	NaN
ILD@5	0.073	0.000	**0.079**	0.000	**+8.820%**
ILD@16	0.006	0.000	**0.006**	0.000	**+2.840%**

TABLE II. A/B TESTING EXPERIMENT RESULTS

Metric	**Control**	**Treatment**
Members	138.4K	138.3K
Average Content Starts	15.57	15.92
Delta % (Content Starts)	**+2.26%**	
p-value	**0.004**	

B. Online A/B Test Results

To validate the real-world impact of our semantic sampling approach, we conducted an A/B test involving a substantial user base of 276,700 members. The test compared the control group, which used the baseline TiSASRec model, with the treatment group, which experienced the enhanced model with semantic sampling. These results are shown in Table II.

1) Average Content Starts: As a media-oriented app, one of the primary metrics we focused on in the A/B test was the average content starts per user during the experimental period. This metric measures how frequently users initiated interactions with the recommended content. The treatment group, which used the semantic sampling approach, exhibited an average content start rate of 15.92 starts per user, while the control group, using the baseline TiSASRec model, had an average content start rate of 15.57 starts per user.

The 2.26% lift in content starts for the treatment group compared to the control group was statistically significant (p-value ¡ 0.05), indicating that the increase in user engagement with the recommended content was not due to random chance. This result highlights the effectiveness of the semantic sampling approach in encouraging users to interact more frequently with the content suggested by the recommender system.

IV. DISCUSSION

The results of our study demonstrate the effectiveness of the semantic sampling approach in enhancing the diversity of recommendations and increasing user engagement in the Headspace app. The evaluation of diversity metrics from the online experiment showed significant improvements in coverage, entropy, rarity, and intra-list diversity, indicating that the semantic sampling approach successfully addressed the long tail problem in recommendation systems. By suggesting more diverse, unique, and less predictable content to users, the semantic sampling approach enriches users' discovery experience, encouraging them to explore a wider range of content.

The online A/B test further validated the real-world impact of the semantic sampling approach, showing a statistically significant lift of 2.26% in average content starts for the treatment group. This indicates that users in the treatment group found the recommendations generated using semantic sampling to be more engaging and relevant, leading to increased interactions with the recommended content.

A. Explanation of Increased Diversity and Relevance

The success of the semantic sampling approach in enhancing diversity and relevance can be attributed to two key factors:

Semantic Understanding: The use of language embeddings allowed the system to better understand the semantic meaning of content items. By capturing the inherent relationships between content titles and teasers, the approach could identify and group together semantically similar pieces. This understanding enabled the recommendation system to present a broader range of content options to users, encompassing items that share similar themes or topics.

Random Sampling: The introduction of random sampling from the cluster of semantically similar items injected diversity into the recommendation process. Instead of being confined to a fixed set of items, users were presented with randomly selected content pieces with similar meanings. This randomness allowed for serendipitous discoveries and introduced novelty, making the user experience more exciting and diverse.

Enhanced Relevance: Despite the introduction of diversity through random sampling, the semantic sampling approach ensured that the recommended items remained highly relevant to each individual user. By selecting items from the cluster of semantically similar content, the approach ensured that the recommendations retained a certain level of thematic coherence and alignment with users' preferences. This balance between diversity and relevance led to a more personalized and engaging experience for users, as they received a mix of both familiar and novel content that resonated with their interests.

Overall, the semantic sampling approach struck a delicate balance between increasing diversity and maintaining relevance, making it a powerful tool in addressing the long-tail

problem in recommendation systems. By leveraging semantic understanding and random sampling, the approach provided users with a diverse and personalized set of recommendations that enriched their discovery experience while ensuring the content remained highly relevant to their individual tastes and preferences.

B. Limitations

While the semantic sampling approach offers a promising solution to the long-tail problem and has demonstrated significant improvements in diversity and user engagement, it does have certain limitations that should be considered:

Limited Diversity Boost for Extremely Niche Content: The semantic sampling approach relies on identifying semantically similar content items to enhance diversity. However, for extremely niche or specialized content items that have limited semantic similarities with other items in the system, the approach may have limitations in boosting their diversity. This could lead to less diverse recommendations for such niche content.

Language Embedding Quality: The effectiveness of the semantic sampling approach is highly dependent on the quality of language embeddings obtained from models like LaBSE. Any limitations or biases present in the language embedding model can impact the accuracy of semantic similarities and, consequently, the diversity of recommendations. Ensuring the high quality and representativeness of language embeddings is critical for the success of the approach.

Impact of Sampling Parameters: The semantic sampling approach involves selecting a certain number of most semantically similar items (N) from which to sample recommendations. The choice of N can influence the level of diversity and relevance of the recommendations. Suboptimal values of N may lead to underemphasizing or overemphasizing certain content clusters, affecting the overall quality of recommendations. Careful experimentation and tuning of the sampling parameters are necessary for optimal results.

The semantic sampling approach represents a notable advancement in recommendation systems, with the potential to enhance diversity and user engagement in the wellness content domain. Ongoing research and refinement can address its limitations and further amplify its impact.

V. Conclusion

In this study, we introduced a semantic sampling method tailored for wellness recommendation systems, with an emphasis on the Headspace app. Online A/B testing involving over 276,700 users revealed that our method yielded a 2.26% uptick in the average content start rate—a clear indication of elevated user engagement.

In terms of diversity metrics, our method surpassed the TiSASRec baseline consistently. There were marked gains in metrics such as coverage, entropy, and rarity. Notably, the semantic sampling method ensured a broader range of engaging content recommendations without sacrificing user-specific relevance. Furthermore, increased content diversity did not detract from the start rate, underscoring the method's capability to harmonize between tailored recommendations and content variety.

From a practical standpoint, semantic sampling amplifies the value proposition of content-rich platforms like Headspace. It fosters an environment where users are more inclined to explore diverse content, leading to a dynamic and individualized user journey. Future investigations might delve into refining similarity clustering methodologies and probing the long-term user satisfaction for lasting effects.

In summation, our research underscores semantic sampling's potential in augmenting both diversity and engagement in wellness recommendation systems, solidifying user satisfaction, and ensuring sustained app interaction.

Acknowledgment

This work was funded through Headspace's internal research and development budget. We thank Rachel Stevenson and Setu Shah all of whom are current employees of Headspace Health, for their feedback and contributions to this paper.

Data Availability

For this study, we used a dataset of Headspace App's user analytics events, specifically pertinent to content starts, plays, and completes on recommended content on all tabs and subsections of the app. This data is not publicly available due to privacy reasons and to safeguard Headspace's in-house analytics data.

References

[1] L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou, "Recommender systems," *Physics reports*, vol. 519, no. 1, pp. 1–49, 2012.

[2] M. Economides, J. Martman, M. J. Bell, and B. Sanderson, "Improvements in stress, affect, and irritability following brief use of a mindfulness-based smartphone app: a randomized controlled trial," *Mindfulness*, vol. 9, no. 5, pp. 1584–1593, 2018.

[3] M. Kunaver and T. Požrl, "Diversity in recommender systems–a survey," *Knowledge-based systems*, vol. 123, pp. 154–162, 2017.

[4] J. Li, Y. Wang, and J. McAuley, "Time interval aware self-attention for sequential recommendation," in *Proceedings of the 13th international conference on web search and data mining*, 2020, pp. 322–330.

[5] Z. Xu, C. Chen, T. Lukasiewicz, Y. Miao, and X. Meng, "Tag-aware personalized recommendation using a deep-semantic similarity model with negative sampling," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016, pp. 1921–1924.

[6] R. L. Lee, "Review of headspace: Meditation and sleep." 2023.

[7] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," *arXiv preprint arXiv:2007.01852*, 2020.

[8] J. Li, Y. Wang, and J. McAuley, "Time interval aware self-attention for sequential recommendation," *Proceedings of the Thirteenth ACM International Conference on Web Search and Data Mining*, p. 9, 2020.

[9] P. Castells, N. Hurley, and S. Vargas, "Novelty and diversity in recommender systems," in *Recommender systems handbook*. Springer, 2021, pp. 603–646.

[10] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang, "Solving the apparent diversity-accuracy dilemma of recommender systems," *Proceedings of the National Academy of Sciences*, vol. 107, no. 10, pp. 4511–4515, 2010.

[11] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in artificial intelligence*, vol. 2009, 2009.

[12] Y. C. Zhang, D. Ó. Séaghdha, D. Quercia, and T. Jambor, "Auralist: introducing serendipity into music recommendation," in *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012, pp. 13–22.

[13] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *Proceedings of the 14th international conference on World Wide Web*, 2005, pp. 22–32.

[14] J. Liu, P. Dolan, and E. R. Pedersen, "Personalized news recommendation based on click behavior," in *Proceedings of the 15th international conference on Intelligent user interfaces*, 2010, pp. 31–40.

State of the Art in Intent Detection and Slot Filling for Question Answering System: A Systematic Literature Review

Anis Syafiqah Mat Zailan[1], Noor Hasimah Ibrahim Teo[2], Nur Atiqah Sia Abdullah[3], Mike Joy[4]
School of Computing Sciences-College of Computing, Informatics, and Media,
Universiti Teknologi MARA (UiTM), Melaka Branch, Merlimau, Melaka, Malaysia[1, 2]
School of Computing Sciences-College of Computing, Informatics, and Media,
Universiti Teknologi MARA (UiTM), Shah Alam Branch, Shah Alam, Selangor, Malaysia[3]
Department of Computer Science, University of Warwick, Coventry, CV4 7AL, United Kingdom[4]

***Abstract*—A Question Answering System (QAS), also known as a chatbot, is a Natural Language Processing (NLP) application that automatically provides accurate responses to questions posed by humans in natural language. Intent Detection and Classification are crucial elements in NLP, especially in a task-oriented dialogue system. In this paper, we conduct a systematic literature review that will perform a comparative analysis of different techniques or algorithms that are being implemented for intent detection and classification with slot filling. The goals of this paper are to identify the distribution, methodology, techniques or algorithms, and evaluation methods, that can be used to develop and construct a model of intent detection and classification with slot filling. This paper also reviews academic documents that have been published from 2019 to 2023, based on a four-step selection process of identification, screening, eligibility, and inclusion, for the selection process. In order to examine these documents, a systematic review was conducted and four main research questions were answered. The results discuss the methodology that can be used for the implementation of intent detection and classification with slot filling, along with the techniques, algorithms and evaluation methods that are widely used and currently implemented by other researchers.**

Keywords*—Intent detection; intent classification; slot filling; question answering system

I. Introduction

A Question Answering System (QAS) is a group of natural language texts or a pre-structured database which are used to automatically provide correct and accurate responses to questions posed by humans in human natural language [1, 2]. A QAS is more capable and more efficient in answering the user query than most search engines such as Google, Yahoo, Live.com, Ask, YouTube, Facebook, and Microsoft Bing [3].

Search engines have a remarkable capability; however, the engines provide lists of related websites and resources including documents that match the user's query [2] without regard to their real intention or what the real question asked is. Hence, instead of the user having to search for the most relevant website or result to their query provided by the search engines, a QAS makes it easier for the user since it just displays the necessary information and results by detecting, recognizing, and classifying the intents of the user via their human natural language, thus providing them with a corresponding yet accurate response and result directly [2, 4]. Natural Language Processing (NLP) is one of the branches of Artificial Intelligence (AI) that studies human-computer interactions [5]. AI has two subsets, which are Natural Language Understanding (NLU) and Natural Language Generation (NLG). Being able to detect the intent of an utterance has been a keen issue in NLU [6] since it uses syntactic and semantic analysis of the text and speech to determine the meaning of a sentence.

Intent Detection (ID) and Classification (IC) act as critical elements in human language or NLU, especially in a task-oriented dialogue system [1, 6, 7, 8]. In particular, ID is usually related to the keywords of utterance, entities or slots, and intent. ID mainly aims to identify the user's true intent from a given utterance [8] in which, according to [9], the user's true intention sometimes does not settle the meaning of their utterance, and vice versa.

Entities or slots refer to the extraction of the associated arguments of the utterance in which every word in the utterance has its entity or slot. The most usual way of its representation is the IOB representation [10], where the initial "B" in the label indicates the beginning of the slot, the "I" indicates an extension of the "B", and lastly "O" refers to a null label, and it commonly labels the other general words in an utterance. The slots or entities that are constructed are identified by the researchers, hence it is meaningful and easy to be understood by humans.

The intent that is detected by the ID model is another crucial part that defines the context of the text, whereby these intents are the ones that classify and capture the true intention of the user, what they are trying to search for or what they are trying to convey [11] to the application or the system. There are numerous possible intents, depending on the utterance, and for an example of an utterance related to books, some of its intent might be SearchBook, PurchaseBook, BookAuthor, BookGenre, and so on. Many researchers have investigated intent recognition in the English language with very good accuracy. However, to the best of our knowledge, there has been little work done on the Malay language for question answering, with exceptions such as [12].

Within this paper, several methods and methodologies have been identified and proposed for developing ID and IC with a slot-filling model. For instance, [11] has performed several experiments on ID and IC by using various techniques such as Seq2seq, Slot-Gated, Capsule NLU, SF-ID, StackPropagation, SlotRefine, GL-GIN, and Joint-BERT. In addition, a model of ID and IC has been proposed with slot filling of JointBERT and Conditional Random Forest (CRF), called JointIDSF, whereby both XLM-R and PhoBERT were used as the utterance encoder [1]. This paper will contribute to better understanding, and thus developing and implementing, the methods of ID and IC with slot filling in various fields.

In addition, we review the existing literature on the development and implementation of ID and IC with slot filling in this paper in which our main objective is to explore which methods or techniques are the most suitable and can achieve higher accuracy in terms of IC performance. The following are the contributions of our review.

- The previous research that has been published regarding ID and IC with slot filling is investigated.
- The methodologies or approaches proposed by previous research are identified and discussed.
- The techniques or algorithms that can be implemented on ID and IC with slot filling are identified.
- A discussion on the validation process and main results of the previous research is conducted.

Furthermore, many previous research articles have been published and have discussed the researcher's implementation on the topic of ID and IC with slot filling. Papers [1, 8, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37] have discussed and presented their approach or implementation on developing ID and IC with slot filling using various frameworks, methodologies, and techniques. Various frameworks, methodologies, approaches, techniques and algorithms have been implemented by these researchers, whereas for instance, [19, 22, 24, 26, 27, 28, 31, 34, 35, 36] have implemented techniques of Bidirectional Long Short-Term Memory ID (BiLSTM) with Conditional Random Forest (CRF) in their implementation of ID and IC with slot filling.

On the other hand, [8, 15, 17, 20, 24, 26, 28, 31, 33] have implemented integration techniques of Bidirectional Encoder Representations from Transformers (BERT) with various other techniques such as CRF, LSTM, BiLSTM, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), or Regular Expressions (RE). On top of that, there are also published papers with comprehensive reviews [50-51, 53, 55]. The aim of [50-51] is to conduct a literature review about Intents, Intention Mining, and IC and focuses on the review of algorithms, models, and tools that have been implemented in Intention Mining. A review paper on the ID methods in the human-machine dialogue system seeks to advance the study of multi-intent detection methods based on Recurrent Neural Networks [52] and deep neural networks [53]. This paper primarily analyses compares, and summarizes the deep learning methods used in the research of ID in recent years. It also considers how to apply deep learning models to multi-intent detection tasks. The third paper [55] introduces the methods of two tasks ranging from the independent model to the joint model. It focuses on joint modelling methods based on deep neural networks and analyzes current problems and future development trends of two sub-tasks.

Therefore, one of the aims of this paper is to focus on conducting a systematic literature review about ID and IC with slot filling in which we are not focusing on the implementation, but we are going deeper into the investigation and discussion of the framework, methodology, and techniques or algorithms that can be implemented in ID and IC with slot filling. Furthermore, what we are reviewing differs from [51, 53, 55], which reviews the topic of intention mining, ID methods in the human-machine dialogue system, and methods of two tasks from the independent model to the joint model, respectively.

This paper is divided as follows: Section II presents the conducted systematic review methodology that consists of the definition of research questions, search phases, inclusion and exclusion criteria, and paper eligibility screening. Section III presents the results of the systematic review consisting of answers to the research questions, and finally, the conclusion of the systematic review and our thoughts on directions for future work are presented in Section IV.

II. Research Method

The systematic review (SR) in this paper was conducted by implementing the preferred reporting items for an SR and meta-analysis (PRISMA) approach as done in [13], in which PRISMA is an evidence-based minimum set of items used to guide the development and structure of SRs and other meta-analysis. According to [14], PRISMA is designed to help researchers to perform literature reviews systematically and transparently. and report how the review was done in a manner which leads to the findings. Thus, by implementing the PRISMA approach in our paper, the reviewing protocol includes three steps, namely definition of the research questions, search phases, and specification of inclusion and exclusion criteria. The specification of these steps is described in the following sections.

A. Definition of Research Question

This SR is organized to encompass the range of research examined by classifying and evaluating previous related articles. To correctly explain the coverage rate of existing works, the research questions must first be defined. We can gain various insights by examining comparable works, which can subsequently assist researchers in coming up with new insights. Table I lists the research questions that were considered in our SR.

B. Search Phase

Defining the information sources is the initial stage in conducting our SR. As shown in Table II, several academic databases, digital libraries, and open-access search engines have been consulted. In order to locate publications that are pertinent to our setting, the following stage entails creating procedures for examining the scientific and technical documentation that these searches produced. The process is built around these two steps: (i) it is necessary to first

determine the search phrases from the earlier research questions in order to create a list of keywords; (ii) it is necessary to create the queries that will be used to locate and gather all connected results, thus the Boolean operators AND/OR will be used to find and gather all related results in accordance to ID and IC with slot filling. About 221 papers were found overall in the first phase with search terms that might be most pertinent in the title. The search terms used for this paper are shown in Table III.

TABLE I. RESEARCH QUESTION

Research question	Motivation
RQ1. What is the distribution per year, domain application, and publisher of the published papers related to ID and IC with slot filling?	The answer to this question allows us to identify the work's domain, when, and where the research studies have been conducted.
RQ2. What is the methodology used to develop ID and IC with slot filling?	The answer to this question illustrates the steps and phases in developing ID and IC with slot filling.
RQ3. What are the techniques or algorithms that can be used to implement ID and IC with slot filling?	The answer to this question helps to identify the most suitable techniques or algorithms that can be adopted into the implementation of ID and IC with slot filling.
RQ4. Which evaluation method was used and what are the main results that have been drawn based on the evaluation method used?	The answer to this question identifies the methods used to evaluate the performance of ID and IC with slot filling and presents the main results or outcomes of the studied works.

TABLE II. SEARCH SOURCES

Source	Type	URL
Science Direct-Elsevier	Digital Library	https://www.sciencedirect.com/
Scopus	Search Engine	https://www.scopus.com/home.uri
IEEE Explore	Digital Library	https://ieeexplore.ieee.org/Xplore/home.jsp
ACM Digital Library	Digital Library	https://dl.acm.org/
Web of Science	Search Engine	https://www.webofknowledge.com/
SpringerLink	Digital Library	https://link.springer.com/
Google Scholar	Search Engine	https://scholar.google.com/

TABLE III. RESEARCH QUERIES

TITLE-ABS-KEY
S1 (intent AND detection OR recognition AND classification)
S2 (intent AND detection AND classification)
S3 (intent AND detection AND recognition AND classification)
S4 (intent AND detection AND slot filling)
S5 (intent AND classification AND slot filling)
S6 (intent AND detection AND classification AND slot filling)

C. Inclusion and Exclusion Criteria

We employed a set of inclusion criteria and exclusion criteria to identify pertinent papers and to narrow search results (see Table IV). Papers that do not address the exclusion criteria are disregarded, and a screening procedure is used to identify papers that are pertinent to our setting. The following three inclusion criteria phases form the basis of the screening procedure:

1) An abstract-based step: It using details and keywords from publication abstracts, we eliminate irrelevant results. Articles were kept for additional screening if their abstracts met at least 60% of the inclusion criteria.

2) Full-text-based step: We eliminate any results that did not address or make reference to the research terms listed in Table III, i.e., any publications that only cover a small portion of the search terms mentioned in their abstracts.

3) Quality-analysis-based step: We perform quality analysis on the remaining results and discard any that do not meet the requirements listed below:

- C1: The paper discusses a comprehensive approach and methodology to ID and IC with slot filling.
- C2: The paper includes the technical implementation of the proposed solution.
- C3: The paper references additional works.
- C4: The paper discusses the outcomes that were found.

TABLE IV. LIST OF IC AND EC

Inclusion Criteria	Exclusion Criteria
Studies published within the period 2019-2023.	Studies that are not written in English.
Studies should meet at least one of the search terms.	Duplicated papers.
Studies should be published/in-press at a journal or conference.	Studies with missing full text.
Studies should provide answers to the research questions.	Papers that are not directly relevant to the ID and IC with slot-filling topics.
The search is performed based on the title, abstract, and full text.	

D. Systematic Search Strategy Procedure

The PRISMA systematic search-strategy procedure includes four core processes which are identification, screening, eligibility, and inclusion were used to choose the pertinent publications for this review. In the identification procedure initially got 221 records (n=221). After the screening procedure, the results were pared down to 149 (n= 149) after duplicates were eliminated. Next subjected to eligibility criteria based on the title and abstract, acquiring 69 (n-69); finally, eligibility criteria based on the complete text allowed us to acquire 25 pertinent studies. To glean the findings reported in the next part, a thorough analysis of these 25 studies was conducted.

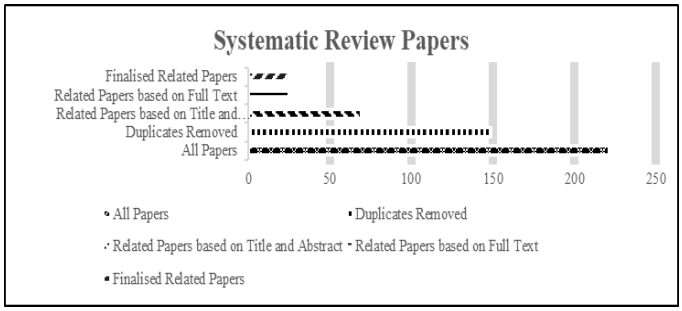


Fig. 1. Related papers reviewed.

Of the 25 papers that had been screened from the review process, all of these papers discussed the ID and IC with slot-filling topics by using various techniques or algorithms. Fig. 1 illustrates the number of papers from each stage in the SR process in a graphical representation of the papers that have been reviewed.

III. Results And Discussion

This section includes a discussion of the review's findings about the earlier proposed research questions. This review is made up of 25 publications that were carefully chosen to address the topic of ID and IC with slot filling. These answers help us to know the related recent literature, methodologies used, techniques or algorithms that can be used to implement ID and IC with slot filling, as well as the methods that can be used to evaluate the performance of the model.

Answer to research question RQ1: What is the distribution per year, domain application, and publisher of the published papers related to ID and IC with slot filling?

The papers that have been gathered and which relate to ID and IC with slot filling originated from several domains, including medical, education, electrical, music, economy, and airline. Fig. 2 and Fig. 3 display the distribution of the selected papers by publication year and source respectively.

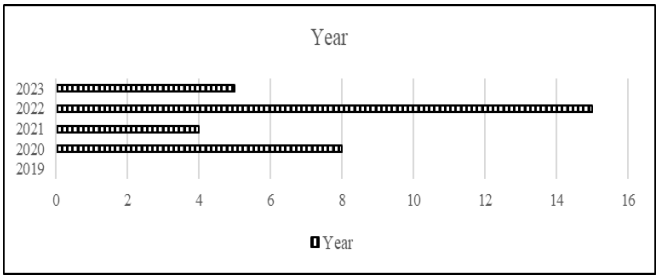


Fig. 2. Distribution of selected papers by publication year.

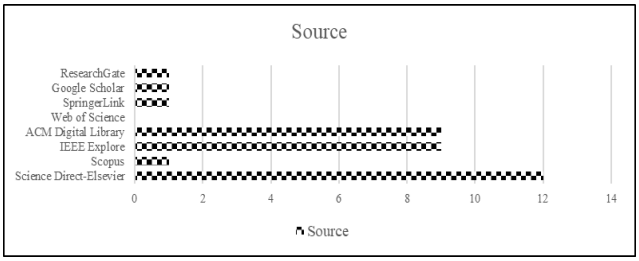


Fig. 3. Distribution of selected papers by source.

Answer to research question RQ2: What is the methodology used to develop ID and IC with slot filling?

Numerous methodologies were used for developing ID and IC with slot-filling phases in which Table V displays the list of methodology names and the phases that are involved in developing ID and IC with slot-filling. After reviewing and analyzing all of the papers, we identified that the majority of the papers integrate several methods resulting in a new methodology or approach. Papers 1, 22, and 25 have integrated Bidirectional Long Short-Term Memory (BiLSTM), Bidirectional Encoder Representations from Transformers (BERT), and Joint BERT with Conditional Random Forest (CRF) respectively. In addition, since there are various and multiple ways in which ID and IC with slot filling have been implemented, almost all of the papers' implementations differ from one another, even though they might seem similar.

The second paper implements a novel non-aggressive joint model, and the fourth paper implements a GloVe approach in their implementation. Some papers implement a unique methodology or approach such as paper 9 which implements a Capsule Network or Capsule-NLU, paper 13 implements a Multi-level Shared-private Framework, paper 14 implements a Deep Concurrent Multi-Task Paradigm, a Dual pseudo-labelling and dual learning methods approach by Paper 16, a Generative and Classification-based approach is implemented by paper 17, and paper 23 implements an Attention-based RNN and Slot-Gated mechanism. Meanwhile, the remaining methodologies implemented by the remaining papers may be further viewed in Table V and Table VI.

In addition, we removed papers [38-47, 54] and [2-15, 49-50, 56-70] from the list due to their related survey paper status. Furthermore, papers [2-15, 49, 56-70, 81] were removed due to their unrelatedness to ID and IC with slot filling or QAS topic relevance.

Answer to research question RQ3: What are the techniques or algorithms that can be used to implement ID and IC with slot filling?

There are several techniques or algorithms that can be implemented for ID and IC with slot filling. According to Table V, these 25 papers have implemented various techniques. Some papers have similarities in their approach with one another but still differ and report unique results.

Initially, Long Short-Term Memory often referred to as LSTM networks are an extension of Recurrent Neural Network (RNN) whereas RNN is a type of neural network that is specially designed for sequence prediction problems since it imposes an order on the observations that must be preserved when training models and making predictions [72]. The LSTM architecture consists of a set of recurrently connected subnets that are also known as memory blocks, which can be thought of as a differentiable version of memory chips in a digital computer [73]. Not only that, LSTM has been primarily implemented for problems such as speech modelling and language translation [72] and it has achieved state-of-the-art performance for IC and slot filling [74]. In addition, an LSTM network is a special variant of an RNN such that it overcomes stability bottlenecks encountered in traditional RNNs, thus

enabling its practical application [72]. Furthermore, an LSTM can also utilize its internal memory in such a way that its predictions are conditional on the recent context in the input sequence, not what has just been presented as the current input to the network. As an example, an LSTM model can show one observation at a time sequentially, and it can learn what observations it has seen previously and which are relevant. From there, it will think and train on how prediction can be done based on their observations made earlier [72].

Furthermore, Bidirectional Long Short-Term Memory (BiLSTM) is a further refinement of LSTM [24] which integrates the forward hidden layer and the backward hidden layer [71], which can acquire and access both the previous and subsequent contexts. Since LSTM exclusively exploits the historical context, unlike BiLSTM, as a result, BiLSTM is better than LSTM at resolving the sequential modelling problem [71]. LSTM and BiLSTM have been used to classify texts and have achieved some progress [19, 22, 24, 26-28].

According to our review, the majority of the papers implement BiLSTM whereas this technique was implemented by 10 papers - 7, 10, 12, 14, 15, 16, 19, 22, 23, and 24. In addition, these papers have integrated BiLSTM with various other techniques which include Conditional Random Forest (CRF), Concurrent Neural Network (CNN), and Bidirectional Encoder Representations from Transformers (BERT) to perform both ID and IC with slot-filling tasks. Some of these papers only implement BiLSTM such as papers 15 and 23 as their main technique. In addition, papers 7, 22, and 24 have integrated BiLSTM with CRF, papers 10 and 14 have integrated BiLSTM with CNN, and are followed by papers 12 and 16 which have integrated BiLSTM with BERT respectively. However, there is also a paper that integrates BiLSTM with two more techniques, and paper 19 has integrated BiLSTM with both CNN and BERT to perform ID and IC with slot filling. The integration between two to three techniques may help to improve the proposed model's performance.

TABLE V. LIST OF PROPOSED FRAMEWORK AND TECHNIQUES

Reference	Proposed Framework	Technique(s) Used
Paper 1 [1]	JointIDSF	JointBERT+CRF
Paper 2 [8]	SlotRefine	BERT+CRF
Paper 3 [15]	Neural Network-Regular Expressions (NN-RE)	RE+CNN+RNN+ BERT+LSTM
Paper 4 [16]	Computational Linguistics with Deep-Learning-Based Intent Detection and Classification (CL-DLBIDC)	Deep Learning Modified Neural Network (DLMNN) +Mayfly Optimization (MFO)
Paper 5 [17]	SLIM	BERT
Paper 6 [18]	BART+MS+I	BART
Paper 7 [19]	Bi-confidence-frequency cross-lingual transfer framework (BiCF)	BiLSTM+CRF
Paper 8 [20]	MultiLingual MultiTask (MLMT)	BERT+CRF
Paper 9 [21]	Capsule ISNP	NER
Paper 10 [22]	Real-time Pilot-controller Voice Communications (PCVC)	Convolutional Neural Network (CNN) + BiLSTM
Paper 11 [23]	PIL Chatbot with Angular, Flask and RASA framework	JointBERT
Paper 12 [24]	Context-aware Graph Convolutional Network with Adaptive Fusion Layer (CGCN-AF)	BiLSTM+BERT
Paper 13 [25]	No specific framework name mentioned	RoBERTa
Paper 14 [26]	CIDIS - Concurrent Intelligent Model for Dialogue Act Classification, Intent Detection and Slot Filling	BERT+Character Embedding Formulation (CharCNN)+BiLSTM
Paper 15 [27]	No specific framework name mentioned	BiLSTM
Paper 16 [28]	Dual semi-supervised NLU with Semantic-to-sentence Generation (SSG)	BiLSTM+CRF+BERT
Paper 17 [29]	No specific framework name mentioned	JointBERT+XLM-Roberta
Paper 18 [30]	No specific framework name mentioned	KoBERT, KLUE-RoBERTa, mBERT
Paper 19 [31]	Multitask Learning with Knowledge Base for Joint Slot-Filling and Intent-Detection (MTL)	BERT+BiLSTM +CNN
Paper 20 [32]	No specific framework name mentioned	CNN+LSTM+Rules
Paper 21 [33]	Tagger and Classifier	LSTM+BERT
Paper 22 [34]	MTL-Fully Shared Network (MTL-FSN) and Hierarchical-MTL (H-MTL)	BiLSTM+CRF
Paper 23 [35]	Slot-Gated Modeling	BiLSTM
Paper 24 [36]	SF-ID Network	BiLSTM+CRF
Paper 25 [37]	No specific framework name mentioned	JointBERT+CRF

TABLE VI. LIST OF METHODOLOGIES AND THEIR RESULT

Reference	Methodology	Methodology Phases
Paper 1 [1]	Extension of JointBERT + CRF	Encoding Layer, Intermediate Intent-Slot Attention Layer, Decoding Layers of Intent Detection and Slot Filling
Paper 2 [8]	Novel Non-Aggressive Joint Model	BERT Layer, Multi-Head Self Attention Layer, Multi-Head Positional Attention Layer, CRF Layer, Iteration Mechanism, Joint Optimization
Paper 3 [15]	No specific methodology name mentioned	Memory network encoder, Intent determination and slot filling module, Joint optimization
Paper 4 [16]	GloVe approach	Data Preprocessing, Feature Extraction Process using Word Embedding Technique, Intent Detection and Classification Process using Modified Neural Network Model, Parameter Tuning using Mayfly Optimization Algorithm
Paper 5 [17]	Multi-intent Spoken Language Understanding (SLU)	Encoder, Intent Classifier and Slot Classifier, Slot-Intent Classifier
Paper 6 [18]	No specific methodology name mentioned	Intent Classification and Slot Filling, Hate Speech Taxonomy, Counter Speech, BART+MS+I Joint Optimization
Paper 7 [19]	No specific methodology name mentioned	BiCF Mixing, Latent Space Refinement, BiLSTM+CRF Joint Encoder
Paper 8 [20]	CNN and RNN with LSTM and GRU	Embeddings, Dense Layer, Task Specific Layer, BERT+CRF Layer - Slot Filling and Attention
Paper 9 [21]	Capsule Network / Capsule-NLU	POS Tagging, Named Entity Recognition (NER), Word Embeddings
Paper 10 [22]	Deep learning-based	Automatic Speech Recognition (ASR) - Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), Fully Connected (FC) Layer, Controlling Intent Inference (CII), Control Safety Monitoring (CSM)
Paper 11 [23]	Retrieval-based	Conversation Manager, Domain Knowledge Handler, Natural Language Interpreter - Intent Classification And Slot Filling, Joint BERT-Based Model
Paper 12 [24]	No specific methodology name mentioned	Hierarchical Encoder, Context-Aware Graph Convolutional Network, Intent and Dialogue Act Classification, Adaptive Fusion Layer for Slot Filling, Multi-Task Training
Paper 13 [25]	Multi-level Shared-private Framework	Shared-Private Syntactic Encoder, Domain-Aware Sentence-Level Transfer for Intent Detection, Task-Aware Token-Level Transfer for Slot Filling, Joint Training
Paper 14 [26]	Deep concurrent multi-task paradigm	Intelligent Word Embedding Formulation, Contextual Sentence Representation Formulation, Dialogue Act Classification, Intent Detection, Slot Filling, Query Response Retrieval
Paper 15 [27]	Encoder-decoder model	Encoding Layer, Word Embedding, Bi-LSTM Layer, Slot Filling, Intent Detection Layer
Paper 16 [28]	Dual pseudo-labeling and dual learning methods	Sentence Encoding, Intent Classification, Slot Tagging, Semantic-to-Sentence Generation: Encoder, Decoder, Dual Semi-Supervised NLU, Dual Learning Model
Paper 17 [29]	Generative and Classification-based	Encoder, Slot Filling, Intent Detection, Joint Training, Decoder
Paper 18 [30]	No specific methodology name mentioned	Intent Classifier, Slot Classifier, Slot Value Predictor, Value Refiner
Paper 19 [31]	Multitasking Learning	LSTM-CNN Layer, Bi-LSTM + Attention Layer, WordNet Knowledge Base, Joint Optimization, Adaptive Moment Estimation
Paper 20 [32]	Rule-based and Model-based	Text Matching, Encoding, Intent Detection, Slot Filling, Rules Fusing, Joint Optimization
Paper 21 [33]	Weakly-Supervised Dual-Model Learning	Encoder, Word Embedding, Intent Detection-LSTM Layer, Attention Layer, Linear Layer, Joint Optimization
Paper 22 [34]	BiLSTM and CRF Model	Word Embedding, Named Entity Recognition (NER), Semantic Tagging (SemTag), Bi-LSTM+CRF Layer
Paper 23 [35]	Attention-based RNN and Slot-Gated Mechanism	BiLSTM Layer, Intent Attention Layer, Slot Attention Layer, Slot Sequence, Joint Optimization
Paper 24 [36]	Novel Bi-directional Interrelated Model	BiLSTM Layer, Slot Attention, Intent Attention, SF Subnet, SF Iteration Mechanism, ID Subnet, ID Iteration Mechanism, CRF Layer
Paper 25 [37]	JointBERT and CRF Model	BERT Layer, Softmax Layer, WordPiece Tokenizer, CRF Layer, Joint Optimization

In recent times, the BERT framework has been investigated for jointly identifying the intent and slots of an utterance [74, 75]. The model architecture of BERT is a multi-layer bidirectional transformer encoder that is based on the original Transformer model [74, 76-77] whereby it jointly conditions both left and right contexts in the Transformer [75, 77]. Furthermore, the input representation for BERT is a concatenation of WordPiece embeddings [8], positional embeddings, and segment embedding [74].

The BERT model is pre-trained with two strategies on large-scale unlabeled text which refers to Masked Language Model (MLM) and Next Sentence Prediction (NSP) [74-75, 77]. In addition, MLM's function is to randomly mask in order to avoid a token observing itself in a multi-layered context and on the other hand, NSP aims to capture useful information for sentence pair-oriented tasks [75]. Nonetheless, BERT models greatly contribute to enhancing NLP and it is the most commonly used transformer architecture [78]. Furthermore, the BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks such as in QAS and language inference, without substantial task-specific architecture modifications [77-78,80].

The most implemented technique other than BiLSTM is Bidirectional Encoder Representations from Transformers known as BERT in which BERT has been implemented by nine papers - 2, 3, 5, 8, 12, 14, 16, 19, and 21. These papers

have also integrated BERT with CRF, LSTM, BiLSTM, CNN, RNN, or Regular Expressions (RE). Firstly, paper 5 has implemented BERT as its sole technique for ID and IC with slot filling. However, of the remaining papers, papers 2 and 8 have integrated BERT with CRF, paper 21 has integrated BERT with LSTM, and paper 12 has integrated BERT with BiLSTM. Papers 14 and 19 have integrated BERT with BiLSTM and CNN meanwhile paper 16 has integrated BERT with BiLSTM and CRF. There is also a paper that has integrated BERT with four more techniques, making a total of 5 techniques including BERT. Paper 3 has integrated BERT with LSTM, CNN, RNN, and Regular Expressions (RE).

In addition, BERT can be extended into several extensions such as JointBERT, RoBERTa, KoBERT, mBERT, and KLUE-RoBERTa. Even though these techniques are extensions of BERT, they still differ and are different in terms of performance, whereas JointBERT has been implemented by papers 1, 11, 17, and 25. Paper 11 only implements JointBERT in its implementation, meanwhile, papers 1 and 25 have integrated JointBERT with CRF, and paper 17 has integrated JointBERT with RoBERTa. This is followed by RoBERTa which has been implemented by papers 13 and 17 and finally, paper 18 has integrated KoBERT, KLUE-RoBERTa, and mBERT in their ID and IC with slot filling implementation.

The remaining papers (3, 4, 6, 9, and 20) have implemented a technique that is not vastly and widely implemented such as some of the papers have implemented Concurrent Neural Networks (CNN), Recurrent Neural Network (RNN), Deep Learning Modified Neural Network (DLMNN), Mayfly Optimization (MFO), Named Entity Recognition (NER) in their implementation of ID and IC with slot filling. More details on each of these paper's implementation of techniques or algorithms can be further viewed in Table V.

Answer to research question RQ4: Which evaluation method was used and what are the main results that have been drawn based on the evaluation method used?

There are various ways of evaluating the performance of a model after it has been implemented by using various techniques. However, there are several main evaluation metrics or methods that are generally used to evaluate the performance of an ID and IC with a filling model. Such methods include accuracy, F1-score, precision, and recall which is mainly applied for both ID and slot filling task, respectively. Besides that, there are also other evaluations conducted within the 25 papers such as sentence accuracy, semantic error, G-measure, and so on. Technically, model evaluation is crucial to check and validate whether the model is well-functioning and works correctly according to the specifications and the requirements or not.

The majority of these papers evaluate the performance of their ID and IC with slot filling model by using accuracy and F1-score for both intent, sentence and slot performance.

Initially, accuracy is the most used empirical measure and it can be defined as a ratio of accurately classified data items to the total number of observations [48, 80], making it one of the most suitable methods to evaluate the performance of ID and IC. Despite it being the most used technique for testing, however, accuracy does not distinguish between the number of correct labels of different classes [79] and is valid only when the evaluation of classification is well-balanced and is not skewed, and there is no class imbalance. Hence, it is not the most appropriate performance metric in some situations especially in a case where the target variable classes in the dataset are unbalanced. This is because, when the model predicts that each point belongs to the majority class label, the accuracy will be high but the model is not accurate because of the imbalances.

Next is a precision technique which is a measure of correctness that is achieved in true prediction, or it also means how many predictions are positive out of all the total of positive predictions. Precision is calculated by the ratio of the total number of correctly classified positive classes divided by the total number of predicted positive classes [68, 80]. However, precision can only work properly and accurately when the FP is higher than the FN [48, 68, 79] and is usually suitable for a system that has a yes or no, true or false result such as e-mail spam detection.

The following evaluation technique is a recall technique which is measured as the actual observations which are predicted correctly. Recall can also be called sensitivity and is the most suitable evaluation technique to be used when the researcher wants to capture as many positives as possible [48, 68, 80]. Recall is a ratio of the total number of correctly classified positive classes divided by the total number of positive classes, or in other words, out of the observations that are positive, how many of them have been predicted by the algorithm [48].

On the other hand, the F1-score, also known as the F-measure, uses both precision and recall and is one of the best evaluation techniques to calculate the performance of an algorithm [48, 80], especially in the evaluation of text classification and identification tasks [59, 80], because it balances out the precision and recall, whereby if the precision is low, the F1 is low and if the recall is low, then the F1 is also low. Plus, the F1score evaluation method has been generally used to evaluate the performance of the slot-filling task.

Therefore, in this paper, we have identified the evaluation methods used in these papers, and Table VII depicts the list of evaluation methods and their results.

TABLE VII. LIST OF EVALUATION METHODS AND THEIR RESULTS

Reference	Evaluation Methods	Results			
		Encoder	Intent Accuracy (%)	Slot F1 (%)	Sentence Accuracy (%)
Paper 1 [1]	Accuracy, F1-Score	XLM-R	97.56	94.95	86.17

<table>
<tr><td></td><td></td><td>PhoBERT</td><td>97.62</td><td>94.98</td><td>86.25</td></tr>
<tr><td>Paper 2 [8]</td><td>Accuracy</td><td colspan="4">Slot Accuracy: 96.22%
Intent Accuracy: 97.11%
Sentence Accuracy: 86.96%</td></tr>
<tr><td>Paper 3 [15]</td><td>Precision,
Recall,
F1-Score</td><td colspan="4"><table><tr><th>Dataset</th><th>Cat</th><th>Prec(%)</th><th>Recall (%)</th><th>F1-Score (%)</th></tr><tr><td rowspan="2">KVRET</td><td>Intent</td><td>98.50</td><td>98.46</td><td>98.48</td></tr><tr><td>Slot</td><td>74.32</td><td>79.18</td><td>76.91</td></tr><tr><td rowspan="2">Frames</td><td>Intent</td><td>95.32</td><td>92.95</td><td>94.17</td></tr><tr><td>Slot</td><td>74.23</td><td>73.15</td><td>73.53</td></tr></table></td></tr>
<tr><td>Paper 4 [16]</td><td>Accuracy,
Recall,
Specificity,
F1-Score,
MCC,
G-Measure</td><td colspan="4"><table><tr><th>Methods</th><th>Training Set (%)</th><th>Testing Set (%)</th></tr><tr><td>Accuracy</td><td>99.29</td><td>99.51</td></tr><tr><td>Recall</td><td>97.50</td><td>98.26</td></tr><tr><td>Specificity</td><td>99.58</td><td>99.71</td></tr><tr><td>F1-Score</td><td>97.50</td><td>98.29</td></tr><tr><td>MCC</td><td>97.09</td><td>98.01</td></tr><tr><td>G-Measure</td><td>97.51</td><td>98.30</td></tr></table></td></tr>
<tr><td>Paper 5 [17]</td><td>F1-Score,
Accuracy</td><td colspan="4"><table><tr><th>Dataset</th><th>Slot F1 (%)</th><th>Intent Accuracy (%)</th><th>SeFr Accuracy (%)</th></tr><tr><td>MixATIS</td><td>88.5</td><td>78.30</td><td>47.60</td></tr><tr><td>MixSNIPS</td><td>96.5</td><td>97.20</td><td>84.00</td></tr></table></td></tr>
<tr><td>Paper 6 [18]</td><td>F1-Score</td><td colspan="4">F1-Score on Full Parse Tree: 52.96%
F1-Score on Top Level Parse Tree: 56.29%
F1-Score on Lower Level Parse Tree: 62.04%
F1-Score on Intent Classification: 57.17%</td></tr>
<tr><td>Paper 7 [19]</td><td>Accuracy</td><td colspan="4"><table><tr><th>Domain</th><th>Intent Accuracy (%)</th><th>Slots Accuracy (%)</th></tr><tr><td>Restaurant</td><td>93.02</td><td>82.91</td></tr><tr><td>Hotel</td><td>94.73</td><td>77.15</td></tr><tr><td>Taxi</td><td>92.73</td><td>91.03</td></tr><tr><td>Attraction</td><td>94.88</td><td>90.74</td></tr></table></td></tr>
<tr><td>Paper 8 [20]</td><td>Accuracy,
F1-Score</td><td colspan="4"><table><tr><th>Dataset</th><th>Intent Accuracy (%)</th><th>Slots F1-Score (%)</th><th>Template Accuracy</th></tr><tr><td>ATIS</td><td>99.18</td><td>97.93</td><td>90.05</td></tr><tr><td>Trains</td><td>86.45</td><td>99.01</td><td>89.56</td></tr><tr><td>Frames</td><td>80.91</td><td>91.67</td><td>83.56</td></tr><tr><td>Snips</td><td>99.11</td><td>97.08</td><td>91.20</td></tr></table></td></tr>
<tr><td>Paper 9 [21]</td><td>Accuracy,
F1-Score,
Semantic Error</td><td colspan="4"><table><tr><th>Dataset</th><th>Intent Accuracy (%)</th><th>Slots F1-Score (%)</th><th>Semantic Error (%)</th></tr><tr><td>ATIS</td><td>89.00</td><td>94.40</td><td>78.10</td></tr><tr><td>Snips</td><td>98.00</td><td>92.90</td><td>85.00</td></tr></table></td></tr>
<tr><td>Paper 10 [22]</td><td>Precision,
F1-Score</td><td colspan="4">- Precision: 99.4%
- F1-Score: 98.7%</td></tr>
<tr><td>Paper 11 [23]</td><td>Accuracy,
SUS score,
Understanding score,
Navigation score,
Intelligence score</td><td colspan="4">- Accuracy: 80.06%
- SUS score: 85.42%
- Understanding score: 83.59%
- Navigation score: 79.93%
- Intelligence score: 92.61%</td></tr>
</table>

<table>
<tr><td>Paper 12 [24]</td><td>Accuracy,
F1-Score</td><td colspan="3">- Intent Accuracy: 99.96%
- Act F1-Score: 97.67%
- Slot F1-Score: 95.06%
- Frame Accuracy: 91.10%</td></tr>
<tr><td>Paper 13 [25]</td><td>Overall Exact,
Accuracy,
Exact</td><td>Dataset / Evaluation</td><td>MTOD</td><td>ASMixed</td></tr>
<tr><td></td><td></td><td>Overall Exact (%)</td><td>91.27</td><td>84.81</td></tr>
<tr><td></td><td></td><td>Slot Accuracy (%)</td><td>95.69</td><td>94.30</td></tr>
<tr><td></td><td></td><td>Intent Accuracy (%)</td><td>99.20</td><td>97.30</td></tr>
<tr><td></td><td></td><td>Exact (%)</td><td>Reminder: 85.62
Alarm: 92.37
Weather: 93.29</td><td>ATIS: 86.53
Snips: 82.62</td></tr>
<tr><td>Paper 14 [26]</td><td>Accuracy,
F1-Score</td><td>Dataset</td><td>Intent Accuracy (%)</td><td>Slots F1-Score (%)</td><td>Dialogue Act Accuracy (%)</td></tr>
<tr><td></td><td></td><td>ATIS</td><td>98.54</td><td>98.89</td><td>99.10</td></tr>
<tr><td></td><td></td><td>Frames</td><td>63.09</td><td>93.52</td><td>48.77</td></tr>
<tr><td></td><td></td><td>Trains</td><td>81.21</td><td>94.07</td><td>78.89</td></tr>
<tr><td>Paper 15 [27]</td><td>Accuracy,
F1-Score</td><td colspan="3">Intent Accuracy: 97.24%
F1-Score: 98.01%</td></tr>
<tr><td>Paper 16 [28]</td><td>Accuracy,
F1-Score</td><td>Dataset</td><td>Intent Accuracy (%)</td><td>Slots F1-Score (%)</td></tr>
<tr><td></td><td></td><td>ATIS</td><td>99.10</td><td>96.00</td></tr>
<tr><td></td><td></td><td>Snips</td><td>99.10</td><td>97.10</td></tr>
<tr><td>Paper 17 [29]</td><td>Accuracy,
F1-Score</td><td colspan="3">Intent Accuracy: 96.26%
Slot F1-Score: 94.01%</td></tr>
<tr><td>Paper 18 [30]</td><td>Accuracy,
F1-Score</td><td>Dataset</td><td>Intent Accuracy (%)</td><td>Slots F1-Score (%)</td></tr>
<tr><td></td><td></td><td>KoBERT</td><td>98.90</td><td>99.70</td></tr>
<tr><td></td><td></td><td>KLUE-RoBERTa</td><td>98.98</td><td>99.45</td></tr>
<tr><td></td><td></td><td>mBERT</td><td>98.38</td><td>99.52</td></tr>
<tr><td>Paper 19 [31]</td><td>Accuracy,
F1-Score</td><td>Dataset</td><td>Intent Accuracy (%)</td><td>Slots F1-Score (%)</td></tr>
<tr><td></td><td></td><td>ATIS</td><td>98.83</td><td>97.06</td></tr>
<tr><td></td><td></td><td>Snips</td><td>98.79</td><td>97.31</td></tr>
<tr><td>Paper 20 [32]</td><td>Accuracy,
F1-Score</td><td colspan="3">Accuracy: 97.37%
F1-Score: 86.67%</td></tr>
<tr><td>Paper 21 [33]</td><td>Precision,
Recall,
F1-Score</td><td colspan="3">Precision: 90.5%
Recall: 92.40%
F1-Score: 91.40%</td></tr>
<tr><td>Paper 22 [34]</td><td>F1-Score</td><td colspan="3">F1-Score for NER: 88.76%
F1-Score for Semantic Tagging(SemTag): 88.96%
F1-Score for both NER and SemTag: 88.78%</td></tr>
<tr><td>Paper 23 [35]</td><td>Accuracy,
F1-Score</td><td>Dataset</td><td>Intent Accuracy (%)</td><td>Slots F1-Score (%)</td><td>Sentence Accuracy (%)</td></tr>
<tr><td></td><td></td><td>ATIS</td><td>94.10</td><td>95.20</td><td>82.60</td></tr>
<tr><td></td><td></td><td>Snips</td><td>96.80</td><td>88.30</td><td>74.60</td></tr>
<tr><td>Paper 24 [36]</td><td>Accuracy,
F1-Score</td><td>Dataset</td><td>Intent Accuracy (%)</td><td>Slots F1-Score (%)</td><td>Sentence Accuracy (%)</td></tr>
<tr><td></td><td></td><td>ATIS</td><td>97.09</td><td>95.80</td><td>86.90</td></tr>
<tr><td></td><td></td><td>Snips</td><td>97.29</td><td>92.23</td><td>80.43</td></tr>
</table>

Paper 25 [37]	Accuracy, F1-Score	**Dataset**	**Intent Accuracy (%)**	**Slots F1-Score (%)**	**Sentence Accuracy (%)**
		ATIS	98.40	96.70	92.30
		Snips	97.90	96.00	88.60

TABLE VIII. LIST OF DATASET, DOMAIN AND LANGUAGE

Reference	Dataset	Domain	Language
Paper 1 [1]	ATIS	Airline	Vietnamese
Paper 2 [8]	ATIS, SNIPS	Airline, General	English
Paper 3 [15]	KVRET, FRAMES	In-Car Assistant, Hotel and Travel-Booking	English
Paper 4 [16]	SNIPS	General	English
Paper 5 [17]	ATIS, SNIPS	Airline, General	English
Paper 6 [18]	Policy-aware Explainable Abuse Detection (PLEAD)	Human Abuse	English
Paper 7 [19]	Self-established ID-WOZ	Restaurant, Hotel, Taxi, Attraction	Indonesian
Paper 8 [20]	ATIS, TRAINS, SNIPS, and FRAMES	Airline, Trains, General, In-Car Assistant, Hotel and Travel-Booking	Hindi, Bengali
Paper 9 [21]	ATIS, SNIPS	Airline, General	English
Paper 10 [22]	Self-Collected Pilot-Controller Voice Communications (PCVC)	Airline	Chinese, English
Paper 11 [23]	PILs Model Repository	Drugs / Medicine	Italian
Paper 12 [24]	ATIS	Airline	English
Paper 13 [25]	MTOD, ASMixed (ATIS, SNIPS)	Alarm, Reminder, Weather, Airline, General	English
Paper 14 [26]	ATIS, TRAINS, FRAMES	Airline, Trains, Hotel and Travel-Booking	English
Paper 15 [27]	ATIS, DSTC5 (The Fifth Dialog State Tracking Challenge)	Airline, State	English
Paper 16 [28]	ATIS, SNIPS	Airline, General	English
Paper 17 [29]	ATIS	Airline	Tamil
Paper 18 [30]	In-Vehicle Domain Dialogue Data	Vehicle	Korean
Paper 19 [31]	ATIS, SNIPS, FRAMES, TRAINS	Airline, General, Hotel and Travel-Booking, Trains	English, Hindi, Bengali
Paper 20 [32]	Self-Collected Music and Non-Music Field of the Human-Machine Dialogue System	Music	English
Paper 21 [33]	ATIS, SNIPS, MIT Restaurant	Airline, General, Restaurant	English
Paper 22 [34]	ATIS, MIT Restaurant, MIT Movie	Airline, Restaurant, Movie	English
Paper 23 [35]	ATIS, SNIPS	Airline, General	English
Paper 24 [36]	ATIS, SNIPS	Airline, General	English
Paper 25 [37]	ATIS, SNIPS	Airline, General	English

In addition, each work's dataset, domain, and language used for the implementation are also written in Table VIII whereby most of the works are developed in English. Therefore, few papers have developed this topic on low-resource languages other than English, such as Bengali [20, 31], Chinese [22], Hindi [20, 31], Indonesian [19], Italian [23], Korean [30], Tamil [29] and Vietnamese [1].

IV. CONCLUSION

This paper summarizes the distribution of papers per year, domain application, and source of the published papers related to ID and IC with slot filling along with the proposed frameworks, techniques or algorithms used, the methodology that has been implemented research, as well as the methodology phases, in each paper. In addition, a systematic review has been conducted using the PRISMA approach, and its selection process of identification, screening, eligibility, and inclusion was reported in detail. A total of 25 works were selected from the 221 works that have been initially extracted, based on their relevance to the four main research questions we have developed. In addition, from the review, we discovered that the techniques and algorithms that are generally and widely used to implement ID and IC with slot filling are Bidirectional Long Short-Term Memory (BiLSTM), Bidirectional Encoder Representations from Transformers (BERT), and Conditional Random Forest (CRF). Not only that, for the evaluation methods, we have looked at various evaluation techniques for ID and IC with slot filling, and we have identified that the majority of the past works' models have been evaluated by the accuracy and F1-score evaluation methods. Therefore, our review on this topic has led us to conclude that ID and IC with slot filling are still crucial and indeed still in need of evolution especially for the low-resource languages other than English such as Malay, Chinese, Tamil, or Vietnamese. Hence, the development of ID and IC with slot filling for low-resource languages requires further studies, implementation, and optimization, in order to provide timely future work opportunities for researchers who are interested in this integrative field.

In the future, this research on intent and slot-filling recognition aims to make these systems more accurate, adaptable to different domains, and responsive in real-time interactions. The focus will be on combining various data sources, like text, speech, and images, to better understand user intents, making conversations more natural. Personalizing models for individual user preferences and ensuring ethical considerations, such as minimizing biases, will be crucial. Additionally, efforts will be directed toward making models interpretable and capable of handling multiple languages seamlessly.

ACKNOWLEDGMENT

The authors express their sincere appreciation to the reviewers for their valuable feedback. The funding for this research was provided by Universiti Teknologi MARA under the Young Talent Research Grant (600-RMC/YTR/5/3 (023/2020).

REFERENCES

[1] Dao, M. H., Truong, T. H., & Nguyen, D. Q. (2021). Intent detection and slot filling for Vietnamese. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 5, 3916–3920. https://doi.org/10.21437/Interspeech.2021-618.

[2] Sweta P. Lende, M. M. R. (2021). Closed domain question answering system using NLP techniques. January 2016. https://doi.org/10.5281/zenodo.49808.

[3] Murphy E P, Fenelon C, Murphy F, et al. (2019). Does Google Have the Answers? The Internet-based Information on Pelvic and Acetabular Fractures. Cureus 11(10): e5952. https://doi.org/10.7759/cureus.5952.

[4] Tang, Y., Han, H., Yu, X., Zhao, J., Liu, G., & Wei, L. (2021). An Intelligent Question Answering System based on the Power Knowledge Graph. 1–5. https://doi.org/https://doi.org/10.48550/arXiv.2106.09013

[5] Dowlagar, S., & Mamidi, R. (2023). A code-mixed task-oriented dialogue dataset for the medical domain. Computer Speech and Language, 78(September 2022), 101449. https://doi.org/10.1016/j.csl.2022.101449.

[6] Zhang, C., Li, Y., Du, N., Fan, W., & Yu, P. S. (2020). Joint slot filling and intent detection via capsule neural networks. ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 5259–5267. https://doi.org/10.18653/v1/p19-1519.

[7] Alshahrani, H. J., Tarmissi, K., Alshahrani, H., Ahmed Elfaki, M., Yafoz, A., Alsini, R., Alghushairy, O., & Ahmed Hamza, M. (2022). Computational Linguistics with Deep-Learning-Based Intent Detection for Natural Language Understanding. Applied Sciences (Switzerland), 12(17). https://doi.org/10.3390/app12178633.

[8] Wu, D., Ding, L., Lu, F., & Xie, J. (2020). SlotRefine: A fast non-autoregressive model for joint intent detection and slot filling. EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 1932–1937. https://doi.org/10.18653/v1/2020.emnlp-main.152.

[9] Leth, P. (2021). Utterance Interpretation and Actual Intentions. Axiomathes, 31(3), 279–298. https://doi.org/10.1007/s10516-019-09462-x.

[10] D. Guo, G. Tur, W. -t. Yih and G. Zweig, "Joint semantic utterance classification and slot filling with recursive neural networks," 2014 IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, NV, USA, 2014, pp. 554-559, doi: 10.1109/SLT.2014.7078634.

[11] Gangadharaiah, R., & Narayanaswamy, B. (2019). Joint Multiple Intent Detection and Slot Labeling for Goal-Oriented Dialog. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1, 564–569. https://doi.org/10.18653/v1/n19-1055.

[12] Puteh, N., Husin, M. Z., Tahir, H. M., & Hussain, A. (2019). Building a Question Classification Model for a Malay Question Answering System. International Journal of Innovative Technology and Exploring Engineering, 8(5s), 184–190.

[13] Zulkipli, Z. Z., Maskat, R., & Teo, N. H. I. (2022). A systematic literature review of automatic ontology construction. Indonesian Journal of Electrical Engineering and Computer Science, 28(2), 878–889. https://doi.org/10.11591/ijeecs.v28.i2.pp878-889.

[14] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.

[15] Abro, W. A., Qi, G., Ali, Z., Feng, Y., & Aamir, M. (2020). Multi-turn intent determination and slot filling with neural networks and regular expressions. Knowledge-Based Systems, 208, 106428. https://doi.org/10.1016/j.knosys.2020.106428.

[16] Alshahrani, H. J., Tarmissi, K., Alshahrani, H., Ahmed Elfaki, M., Yafoz, A., Alsini, R., Alghushairy, O., & Ahmed Hamza, M. (2022). Computational Linguistics with Deep-Learning-Based Intent Detection for Natural Language Understanding. Applied Sciences (Switzerland), 12(17). https://doi.org/10.3390/app12178633.

[17] Cai, F., Zhou, W., Mi, F., & Faltings, B. (2022). Slim: Explicit Slot-Intent Mapping with Bert for Joint Multi-Intent Detection and Slot Filling. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2022-May, 7607–7611. https://doi.org/10.1109/ICASSP43922.2022.9747477.

[18] Calabrese, A., Ross, B., & Lapata, M. (2022). Explainable Abuse Detection as Intent Classification and Slot Filling. Transactions of the Association for Computational Linguistics, 10, 1440–1454. https://doi.org/10.1162/tacl_a_00527.

[19] Di, D., Song, X., Zhang, W., Zhang, Y., & Wang, F. (2022). Building Dialogue Understanding Models for Low-resource Language Indonesian from Scratch. ACM Transactions on Asian and Low-Resource Language Information Processing. https://doi.org/10.1145/3575803.

[20] Firdaus, M., Ekbal, A., & Cambria, E. (2023). Multitask learning for multilingual intent detection and slot filling in dialogue systems. Information Fusion, 91(October 2022), 299–315. https://doi.org/10.1016/j.inffus.2022.09.029.

[21] Staliunaite, I., & Iacobacci, I. (2020). Auxiliary Capsules for Natural Language Understanding. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8149–8153. https://doi.org/https://doi.org/10.1109/ICASSP40776.2020.9053899.

[22] Lin, Y., Deng, L., Chen, Z., Wu, X., Zhang, J., & Yang, B. (2020). A Real-Time ATC Safety Monitoring Framework Using a Deep Learning Approach. IEEE Transactions on Intelligent Transportation Systems, 21(11), 4572–4581. https://doi.org/10.1109/TITS.2019.2940992.

[23] Minutolo, A., Damiano, E., De Pietro, G., Fujita, H., & Esposito, M. (2022). A conversational agent for querying Italian Patient Information Leaflets and improving health literacy. Computers in Biology and Medicine, 141(August 2021), 105004. https://doi.org/10.1016/j.compbiomed.2021.105004.

[24] Qin, L., Che, W., Ni, M., Li, Y., & Liu, T. (2021). Knowing Where to Leverage: Context-Aware Graph Convolutional Network with an Adaptive Fusion Layer for Contextual Spoken Language Understanding. IEEE/ACM Transactions on Audio Speech and Language Processing, 29, 1280–1289. https://doi.org/10.1109/TASLP.2021.3053400.

[25] Qin, L., Wei, F., Ni, M., Zhang, Y., Che, W., Li, Y., & Liu, T. (2022). Multi-domain Spoken Language Understanding Using Domain-And Task-Aware Parameterization. ACM Transactions on Asian and Low-Resource Language Information Processing, 21(4), 1–17. https://doi.org/10.1145/3502198.

[26] Sunilkumar, G., Srihari, S., Gilbert, S. F., & Chitrakala, S. (2022). A Concurrent Intelligent Natural Language Understanding Model for an Automated Inquiry System. Proceedings - 2022 IEEE World Conference on Applied Intelligence and Computing, AIC 2022, 124–129. https://doi.org/10.1109/AIC55036.2022.9848883.

[27] Xu, C., Li, Q., Zhang, D., Cui, J., Sun, Z., & Zhou, H. (2020). A model with length-variable attention for spoken language understanding. Neurocomputing, 379, 197–202. https://doi.org/10.1016/j.neucom.2019.07.112.

[28] Zhu, S., Cao, R., & Yu, K. (2020). Dual Learning for Semi-Supervised Natural Language Understanding. IEEE/ACM Transactions on Audio Speech and Language Processing, 28, 1936–1947. https://doi.org/10.1109/TASLP.2020.3001684.

[29] Ramaneswaran, S., Vijay, S., & Srinivasan, K. (2022). TamilATIS: Dataset for Task-Oriented Dialog in Tamil. Dravidian-LangTech 2022 - 2nd Workshop on Speech and Language Technologies for Dravidian Languages, Proceedings of the Workshop, 25–32. https://doi.org/10.18653/v1/2022.dravidianlangtech-1.4.

[30] Lim, J., Son, S., Lee, S., Chun, C., Park, S., Hur, Y., & Lim, H. (2022). Intent Classification and Slot Filling Model for In-Vehicle Services in Korea. Applied Sciences (Switzerland), 12(23). https://doi.org/10.3390/app122312438.

[31] He, T., Xu, X., Wu, Y., Wang, H., & Chen, J. (2021). Multitask learning with a knowledge base for joint intent detection and slot filling. Applied Sciences (Switzerland), 11(11). https://doi.org/10.3390/app11114887.

[32] Ren, S., Wang, H., Yu, D., Li, Y., & Li, Z. (2018). Joint intent detection and slot filling with rules. CEUR Workshop Proceedings, 2242, 34–40.

[33] Wang, J., Chen, K., Shou, L., Wu, S., & Chen, G. (2021). Effective Slot Filling via Weakly-Supervised Dual-Model Learning. 35th AAAI Conference on Artificial Intelligence, AAAI 2021, 16, 13952–13960. https://doi.org/10.1609/aaai.v35i16.17643.

[34] Louvan, S., & Magnini, B. (2019). Leveraging non-conversational tasks for low resource slot filling: Does it help? SIGDIAL 2019 - 20th Annual Meeting of the Special Interest Group Discourse Dialogue - Proceedings of the Conference, 2003(September), 85–91. https://doi.org/10.18653/v1/w19-5911.

[35] Goo, C. W., Gao, G., Hsu, Y. K., Huo, C. L., Chen, T. C., Hsu, K. W., & Chen, Y. N. (2018). Slot-gated modelling for joint slot filling and intent prediction. NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2, 753–757. https://doi.org/10.18653/v1/n18-2118.

[36] Haihong, E., Niu, P., Chen, Z., & Song, M. (2020). A novel bi-directional interrelated model for joint intent detection and slot filling. ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, September, 5467–5471. https://doi.org/10.18653/v1/p19-1544.

[37] Chen, Q., Zhuo, Z., & Wang, W. (2019). BERT for Joint Intent Classification and Slot Filling. 1–11. http://arxiv.org/abs/1902.10909.

[38] Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2021). Survey on evaluation methods for dialogue systems. In Artificial Intelligence Review (Vol. 54, Issue 1). Springer Netherlands. https://doi.org/10.1007/s10462-020-09866-x.

[39] Jannach, D., Manzoor, A., Cai, W., & Chen, L. (2021). A Survey on Conversational Recommender Systems. ACM Computing Surveys, 54(5). https://doi.org/10.1145/3453154.

[40] Keyvan, K., & Huang, J. X. (2022). How to Approach Ambiguous Queries in Conversational Search: A Survey of Techniques, Approaches, Tools, and Challenges. ACM Computing Surveys, 55(6). https://doi.org/10.1145/3534965.

[41] Han, X., Wang, Y. T., Feng, J. L., Deng, C., Chen, Z. H., Huang, Y. A., Su, H., Hu, L., & Hu, P. W. (2023). A survey of trans-former-based multimodal pre-trained models. Neurocomputing, 515, 89–106. https://doi.org/10.1016/j.neucom.2022.09.136.

[42] Gupta, M., & Agrawal, P. (2022). Compression of Deep Learning Models for Text: A Survey. ACM Transactions on Knowledge Discovery from Data, 16(4), 1–55. https://doi.org/10.1145/3487045.

[43] Liu, S., Mallol-Ragolta, A., Parada-Cabaleiro, E., Qian, K., Jing, X., Kathan, A., Hu, B., & Schuller, B. W. (2022). Audio self-supervised learning: A survey. Patterns, 3(12). https://doi.org/10.1016/j.patter.2022.100616.

[44] Fu, T., Gao, S., Zhao, X., Wen, J. Rong, & Yan, R. (2022). Learning towards conversational AI: A survey. AI Open, 3(February), 14–28. https://doi.org/10.1016/j.aiopen.2022.02.001.

[45] Uc-Cetina, V., Navarro-Guerrero, N., Martin-Gonzalez, A., Weber, C., & Wermter, S. (2022). Survey on reinforcement learning for language processing. In Artificial Intelligence Review (Vol. 56, Issue 2). Springer Netherlands. https://doi.org/10.1007/s10462-022-10205-5.

[46] Liu, K., Chen, Y., Liu, J., Zuo, X., & Zhao, J. (2020). Extracting Events and Their Relations from Texts: A Survey on Recent Research Progress and Challenges. AI Open, 1(February), 22–39. https://doi.org/10.1016/j.aiopen.2021.02.004.

[47] Weld, H., Huang, X., Long, S., Poon, J., & Han, S. C. (2023). A Survey of Joint Intent Detection and Slot Filling Models in Natural Language Understanding. ACM Computing Surveys, 55(8), 1–38. https://doi.org/10.1145/3547138.

[48] Vakili, M., Ghamsari, M., & Rezaei, M. (2020). Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification. https://doi.org/https://doi.org/10.48550/arXiv.2001.09636.

[49] Li, G. (2022). Question Answering System Based on Knowledge Graph in Traditional Chinese Medicine Diagnosis and Treatment of Viral Hepatitis B. 2022.

[50] Weld, H. (2021). A survey of joint intent detection and slot-filling models in natural language understanding. ACM Transactions on Graphics, 37(4). https://doi.org/https://doi.org/10.1145/3547138.

[51] Yanli, H. (2021). Research on Spoken Language Understanding Based on Deep Learning. Scientific Programming, 2021. https://doi.org/10.1155/2021/8900304.

[52] Liu, Z., Feng, Y., & Chen, Z. (2021). DialTest: Automated testing for recurrent neural network-driven dialogue systems. ISSTA 2021 - Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis, 115–126. https://doi.org/10.1145/3460319.3464829.

[53] Liu, J., Li, Y., & Lin, M. (2019). Review of Intent Detection Methods in the Human-Machine Dialogue System. Journal of Physics: Conference Series, 1267(1). https://doi.org/10.1088/1742-6596/1267/1/012059.

[54] Larson, S., & Leach, K. (2022). A Survey of Intent Classification and Slot-Filling Datasets for Task-Oriented Dialog. 1(Section 5), 1–30. https://doi.org/https://doi.org/10.48550/arXiv.2207.13211.

[55] Hou, L., Li, Y., Li, C., & Lin, M. (2019). Review of Research on Task-Oriented Spoken Language Understanding. Journal of Physics: Conference Series, 1267(1). https://doi.org/10.1088/1742-6596/1267/1/012023.

[56] Dowlagar, S., & Mamidi, R. (2023). A code-mixed task-oriented dialog dataset for the medical domain. Computer Speech and Language, 78(September 2022), 101449. https://doi.org/10.1016/j.csl.2022.101449.

[57] Oesterreich, T. D., Anton, E., Schuir, J., Brehm, A., & Teuteberg, F. (2022). How can I help you? Design principles for task-oriented speech dialog systems in customer service. Information Systems and E-Business Management. https://doi.org/10.1007/s10257-022-00570-7.

[58] Jannach, D. (2022). Evaluating conversational recommender systems: A landscape of research. In Artificial Intelligence Review (Vol. 56, Issue 3). Springer Netherlands. https://doi.org/10.1007/s10462-022-10229-x.

[59] Liu, B., Zhang, P., Shu, Y., Guan, Z., Lu, T., Gu, H., & Gu, N. (2022). Building a Personalized Model for Social Media Textual Content Censorship. Proceedings of the ACM on Human-Computer Interaction, 6(2 CSCW). https://doi.org/10.1145/3555657.

[60] Cheng, L., Jia, W., & Yang, W. (2023). Capture Salient Historical Information: A Fast and Accurate Non-autoregressive Model for Multi-turn Spoken Language Understanding. In ACM Transactions on Information Systems (Vol. 41, Issue 2). https://doi.org/10.1145/3545800.

[61] Bhaskaran, S. K., Sreejith, C., & Rafeeque, P. C. (2018). Neural networks and conditional random fields-based approach for effective question processing. Procedia Computer Science, 143, 211–218. https://doi.org/10.1016/j.procs.2018.10.381.

[62] Tang, Y., Han, H., Yu, X., Zhao, J., Liu, G., & Wei, L. (2021). An Intelligent Question Answering System based on Power Knowledge Graph. IEEE Power and Energy Society General Meeting, 2021-July, 1–5. https://doi.org/10.1109/PESGM46819.2021.9638018.

[63] Wang, C., Dai, S., Wang, Y., Yang, F., Qiu, M., Chen, K., Zhou, W., & Huang, J. (2022). ARoBERT: An ASR Robust Pre-Trained Language Model for Spoken Language Understanding. IEEE/ACM Transactions on Audio Speech and Language Processing, 30, 1207–1218. https://doi.org/10.1109/TASLP.2022.3153268.

[64] Fernández-Martínez, F., Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., & Montero, J. M. (2022). Fine-Tuning BERT Models for Intent Recognition Using a Frequency Cut-Off Strategy for Domain-Specific Vocabulary Extension. Applied Sciences (Switzerland), 12(3). https://doi.org/10.3390/app12031610.

[65] Quamar, A., Özcan, F., Miller, D., Moore, R. J., Niehus, R., & Kreulen, J. (2020). Conversational BI: An Ontology-Driven Conversation System for Business Intelligence Applications. Proceedings of the VLDB Endowment, 13(12), 3369–3381. https://doi.org/10.14778/3415478.3415557.

[66] Iovine, A., Narducci, F., Musto, C., de Gemmis, M., & Semeraro, G. (2023). Virtual Customer Assistants in finance: From state of the art and practices to design guidelines. Computer Science Review, 47, 100534. https://doi.org/10.1016/j.cosrev.2023.100534.

[67] Razumovskaia, E., Glavaš, G., Majewska, O., Ponti, E. M., Korhonen, A., & Vulic, I. (2022). Crossing the Conversational Chasm: A Primer on Natural Language Processing for Multilingual Task-Oriented Dialogue Systems. Journal of Artificial Intelligence Research, 74, 1351–1402. https://doi.org/10.1613/JAIR.1.13083.

[68] Ruz, G. A., Henríquez, P. A., & Mascareño, A. (2020). Sentiment analysis of Twitter data during critical events through Bayesian network classifiers. Future Generation Computer Systems, 106, 92–104. https://doi.org/10.1016/j.future.2020.01.005.

[69] Matějů, L., Griol, D., Callejas, Z., Molina, J. M., & Sanchis, A. (2021). An empirical assessment of deep learning approaches to task-oriented dialog management. Neurocomputing, 439, 327–339. https://doi.org/10.1016/j.neucom.2020.01.126.

[70] Anantrasirichai, N., & Bull, D. (2022). Artificial intelligence in the creative industries: a review. In Artificial Intelligence Review (Vol. 55, Issue 1). Springer Netherlands. https://doi.org/10.1007/s10462-021-10039-7.

[71] Bidirectional LSTM with attention mechanism and convolutional layer for text classification.

[72] Mohan, A. T., & Gaitonde, D. V. (2018). A Deep Learning-based Approach to Reduced Order Modeling for Turbulent Flow Control using LSTM Neural Networks. http://arxiv.org/abs/1804.09269.

[73] Graves, A. (2012). Long Short-Term Memory. In: Supervised Sequence Labelling with Recurrent Neural Networks. Studies in Computational Intelligence, vol 385. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-24797-2_4.

[74] Chen, Q., Zhuo, Z., & Wang, W. (2019). Bert for joint intent classification and slot filling. arXiv preprint arXiv:1902.10909.

[75] Castellucci, G., Bellomaria, V., Favalli, A., & Romagnoli, R. (2019). Multi-lingual Intent Detection and Slot Filling in a Joint BERT-based Model. Id. http://arxiv.org/abs/1907.02884.

[76] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 2017-Decem (Nips), 5999–6009. https://doi.org/https://doi.org/10.48550/arXiv.1706.03762.

[77] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 – 2019. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1(Mlm), 4171–4186.

[78] Fernández-Martínez, F., Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., & Montero, J. M. (2022). Fine-Tuning BERT Models for Intent Recognition Using a Frequency CutOff Strategy for Domain-Specific Vocabulary Extension. Applied Sciences (Switzerland), 12(3). https://doi.org/10.3390/app12031610.

[79] Sokolova, M., Japkowicz, N., Szpakowicz, S. (2006). Beyond accuracy, F-score and Roc: A family of discriminant measures for performance evaluation. Lecture Notes in Computer Science, 1015–1021. https://doi.org/10.1007/11941439_114.

[80] Purwandari, K., Cenggoro, T. W., Chanlyn Sigalingging, J. W., & Pardamean, B. (2023). Twitter-based classification for integrated source data of weather observations. IAES International Journal of Artificial Intelligence (IJ-AI), 12(1), 271. https://doi.org/10.11591/ijai.v12.i1.pp271-283.

[81] Palai, P., Agrawal, K., Prasad Mishra, D., & Reddy Salkuti, S. (2023). Text grouping: A comprehensive guide. IAES International Journal of Artificial Intelligence (IJ-AI), 12(3), 1476. https://doi.org/10.11591/ijai.v12.i3.pp1476-1483.

Enhancing Vehicle Safety: A Comprehensive Accident Detection and Alert System

Jamil Abedalrahim Jamil Alsayaydeh[1]*, Mohd Faizal bin Yusof[2], Mohamad Amirul Aliff bin Abdillah[3], Ahmed Jamal Abdullah Al-Gburi[4], Safarudin Gazali Herawan[5], Andrii Oliinyk[6]

Department of Electronics & Computer Engineering Technology, Fakulti Teknologi Kejuruteraan Elektrik & Elektronik (FTKEE), Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia[1, 3, 4]

Department-Homeland Security-Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates[2]

Industrial Engineering Department-Faculty of Engineering, Bina Nusantara University, Jakarta, Indonesia 11480[5]

Department of Software Tools, National University «Zaporizhzhia Polytechnic», 64 Zhukovskoho str., Zaporizhzhya, Ukraine[6]

***Abstract*—This research pioneers a ground-breaking system meticulously engineered to swiftly detect vehicular accidents and dispatch immediate alerts to both emergency services and pre-assigned contacts. This symphony of cutting-edge technologies includes an accelerometer sensor attuned to detect acceleration in any vector, a dynamic Liquid-Crystal Display (LCD) display for rapid alert dissemination, an assertive buzzer for resonant alarms, a Global System for Mobile (GSM) module for the swift transmission of distress messages, and pinpoint location data provided by a Global Positioning System (GPS) module. A user-friendly 'cancel' button acts as an escape hatch from potential false alarms. Orchestrated by the dexterity of an Arduino Uno microcontroller, this ensemble orchestrates a harmonious ballet of safety. This solution boasts cost-effectiveness, steadfastness, and unparalleled efficiency. Rigorous testing across diverse scenarios confirms its precision and robustness. By enhancing accident detection accuracy, expediting emergency responses, and facilitating rapid location dissemination, this innovation serves as a vital lifeline, empowering both passengers and rescue services upon accident initiation. With location data as its guiding star, emergency services gain a swift navigational edge, offering a beacon of hope in the battle against accident-related casualties.**

***Keywords*—Vehicle accident detection; microcontroller-based system; accelerometer sensor; Global Positioning System (GPS) localization; Global System for Mobile (GSM) communication; emergency response; safety innovation**

I. INTRODUCTION

The innovative vehicle accident detection and alert system described in this article addresses a crucial gap in the existing safety mechanisms for vehicles. While traditional safety features like airbags and seatbelt tensioners have been valiant in their efforts, they often fall short in preventing accidents and lack the capability to swiftly relay critical information to emergency services and the victim's loved ones.

The gap lies in the need for a comprehensive solution that not only detects accidents promptly but also initiates an immediate response. The current safety mechanisms, though effective to a certain extent, do not harness advanced technology to redefine accident prevention and response.

To address the existing gap in vehicle safety, our research focuses on three pivotal questions. Firstly, we explore the integration of advanced technology to surpass traditional safety measures and enhance overall vehicle safety. Secondly, we delve into the identification of key components and features essential for a comprehensive vehicle accident detection and alert system. Lastly, we examine methods to ensure the system's capability to provide immediate and accurate alerts to both passengers and emergency services in the event of an accident.

Aligned with these questions, our research objectives are multifaceted. Our primary objective is the design of a cutting-edge vehicle accident detection and alert system. This system will incorporate a range of technologies, including a microcontroller, accelerometer sensor, LCD display, buzzer, GSM module, GPS technology, and a cancel button. Subsequently, our research aims to implement and rigorously test this system across various scenarios, ensuring its accuracy, timeliness, and reliability. In addition, we aspire to provide a cost-effective solution applicable to all types of vehicles, making advanced safety technology more accessible. Finally, our overarching goal is to enable immediate alerts for passengers and emergency services, facilitating a rapid response that holds the potential to save lives. These research objectives collectively contribute to advancing the field of vehicle safety and addressing critical gaps in current safety mechanisms.

The significance of this research lies in its ability to fill the existing gap in vehicle safety measures. By introducing a comprehensive and efficient solution, we aim to redefine accident prevention and response, ensuring a safer and more secure transportation environment. The research contributes to the advancement of technology in vehicle safety, with potential implications for reducing accident-related fatalities and injuries. This innovative system has been designed with a focus on cost-effectiveness and reliability, making it widely applicable and impactful in enhancing overall road safety.

The remainder of this paper is organized as follows: Section II delves into a discussion of related works, presenting an overview of existing research in the field. Section III described the system implementation and testing. Section IV describes the results and analysis meanwhile the conclusion is described in Section V. Lastly; future works is mentioned in Section VI.

II. Related Work

In the pursuit of safer roads and more efficient accident prevention, previous research has primarily focused on conventional safety measures such as airbag deployment and seatbelt tensioners. While these measures have made significant strides in enhancing vehicle safety, they face limitations in effectively detecting and preventing accidents, especially in real-time scenarios. This literature review explores the evolution of accident detection and response systems, leading to the proposed innovative method that leverages advanced technology to address the critical challenge of accurate and timely accident detection.

Historically, vehicle safety measures centered around passive systems like airbags. However, these traditional approaches, while valuable in mitigating accident consequences, fall short in their ability to proactively prevent accidents or provide swift alerts to relevant parties in the event of an incident. Anand Gunadal's pioneering research harnessed MEMS accelerometers and GPS tracking to monitor vehicle behavior, particularly during accidents. These accelerometers, both analog and digital, detected changes in velocity and acceleration. The analog-to-digital conversion process allowed for precise analysis of these changes. Gunadal's work laid the groundwork for advanced accident detection systems, focusing on crucial alterations in vehicle motion for improved safety [1] [2].

The pursuit of smarter road safety gave rise to early innovations, such as the use of MEMS accelerometers and GPS tracking to monitor vehicle behavior and detect accidents. These systems demonstrated the potential to enhance accident detection by assessing acceleration forces and vehicle movements. Giriraj Gurjar's research introduced a discreet accident monitoring system that concealed equipment within vehicles. This system utilized GSM communication to transmit comprehensive accident reports, including temperature, smoke conditions, vehicle speed, and accident time [3]. It featured a MEMS accelerometer for continuous vehicle motion detection. During an accident, the MEMS accelerometer sensed the vehicle's movement and relayed this information to a microcontroller. GPS technology was employed to pinpoint the accident location, with graphical representation displayed on an LCD screen. Gurjar's work enhanced accident detection and reporting capabilities [4].

Purushotham and Kumar's research introduced a novel approach by utilizing GPS technology to track vehicles. Their system compared GPS data with pre-defined checkpoints and mapped the location on platforms like Google Earth [5]. This innovative approach streamlined the task of locating accident sites for rescue teams. Additionally, their system incorporated an ultrasonic wave sensor, which measured echo return time, providing valuable insights for accident detection. By combining GPS tracking, innovative sensing mechanisms, and location mapping, their project offered a comprehensive solution that enhanced both accident detection and efficient rescue operations [6].

Kassem and Jabr's research in [7] explored the diverse benefits of automotive black box systems. These systems, they noted, hold the potential to enhance vehicle design, aid accident victims' treatment, assist insurance providers in collision investigations, and influence traffic conditions to reduce fatalities. Their study emphasized the importance of effectively collecting vehicle data, achieved through a combination of basic components and sensors. Data was presented in real-time graphics and saved in an Excel file for further analysis. The hardware comprised sensors and a black box within the vehicle, recording data, including speed. Additionally, their research underscored the significance of analyzing vehicle lights, such as brake lights and flashers, to gain insights into accident investigations. Their adaptable Black Box system, applicable to various vehicles, initiated data recording upon engine startup, providing comprehensive insights [8]. Buyers received detailed reports containing all relevant data. Kassem and Jabr's work illustrated the potential of automotive black box systems in enhancing vehicle safety, accident investigations, and traffic management [9].

Watthanawisuth, et al.'s study aimed to develop a wireless black box system tailored for monitoring motorcycle accidents. It featured a MEMS accelerometer and GPS tracking, enabling accident type detection, post-crash posture assessment, and GPS ground speed determination using accelerometer signals and a threshold technique. In case of an accident, the system promptly triggered an alarm, sending concise alert data via the GSM network and a text message to a designated contact with GPS location information. Real-time monitoring distinguished between falls and accidents based on motorcycle speed and a limit algorithm. The device also logged track and acceleration data one minute before and after an accident, facilitating comprehensive accident analysis [10] [11] [12].

Ritwik Chinmaya Pandia's study focuses on a project designed to capture vital information such as vehicle speed and position during accidents. The system employs a GSM modem to trigger an accident alarm and send the vehicle's current location to a pre-programmed cell phone upon detecting a crash. It utilizes efficient voltage transformation with transformers, rectifiers, and a microcontroller (ATmega16) for communication and control. A piezo sensor is employed as a key sensory element, and the programming is accomplished using AVR Studio and AVR Dude [13]. This study significantly enhances accident data capture and reporting, blending hardware components and microcontroller technology for improved accident information retrieval [14] [15].

Prabha et al. [16] have made a big contribution to enhancing travel safety, utilizing GSM and GPS technology for real-time accident detection and notification. This project places a strong emphasis on travel security by developing an automatic vehicle accident detection and notification system. It enables immediate SMS alerts in the event of an accident and maintains accurate vehicle positioning through GPS technology [17].

The research by Krishna Kanth et al. [18] Singh focuses on the prompt detection of accidents in any location, enabling swift ambulance response through GPS and GSM networks. Their innovative automotive accident detection module integrates GPS, GSM, and MEMS technology. A standout feature of this project is the capability to send notifications to accident locations when authorized individuals place a missed

call to the GSM module at that site. This approach streamlines vehicle identification in accident scenarios, utilizing robust tracking technology. Overall, this study emphasizes the vital role of GPS, GSM, and MEMS in enhancing accident detection and vehicle tracking for improved road safety and emergency services [19].

Anita Kumari et al. [20] focused on a car tracking system that utilizes a GSM module to capture GPS data and send it to a designated mobile or laptop. It highlights the significance of Vehicle Tracking Systems, which use GPS to locate and monitor vehicles, relaying position data to a central monitoring center. The research emphasizes the increasing popularity of vehicle monitoring systems and describes the process of converting GPS data to RS232 format for transmission to the GSM module via a microcontroller and MAX232. It also mentions the system's potential in mitigating accidents, including collisions and in-vehicle fires [21] [22].

The study conducted by Anil Kumar et al. [23] aims to enhance vehicle security and raise driver awareness through IoT technology. It continuously analyzes vehicle performance and driver behavior using various sensors, including a breath analyzer, accelerometers, and distance sensors between other vehicles. In addition, it features a push and panic button for data input into the vehicle's black box. The system monitors alcohol intake levels and alerts emergency contacts when the limit is exceeded. In the event of an accident, the system uses GSM and GPS to monitor the vehicle's location and communicates this information to hospitals and the police. The IoT-based controller is designed to be power-efficient, enabling real-time applications. The project employs a range of sensors, including breath analyzers, accelerometers, and ultrasonic sensors, to ensure driver and merchant safety. It also detects alcohol gas concentrations between 0.05 mg/L to 10 mg/L. When an accident occurs, the GPS module activates and transmits the location via GSM to local authorities [24].

The project conducted by Ranjitha et al. [25] focuses on using a GPS module to precisely locate road accidents and promptly send this position to a pre-programmed phone number via GSM SMS. The project utilizes an Arduino board, offering easy access to input/output and analog ports, as well as the capability for programmed burning and uploading. The GPS module continuously updates the Arduino with the vehicle's longitude and latitude coordinates. This data is then transmitted through the GSM module, which can send the precise latitude and longitude as an SMS to a pre-programmed phone number. Additionally, the system incorporates a limit switch, which, when subjected to a particular amount of pressure, prompts the Arduino to retrieve the GPS module's latitude and longitude and transmit this information to the GSM module. The system also includes an alcohol detector that alerts the Arduino when alcohol is detected, prompting it to send a message to the GSM module, which in turn transmits the message to the designated recipient [26].

In conclusion, the reviewed literature and studies present commendable advancements in accident detection and response systems. However, a critical analysis reveals potential gaps in scalability, accuracy, privacy, data security, system robustness, response time, AI integration, and pedestrian safety. Further exploration and innovation in these areas could significantly contribute to the development of more comprehensive and effective solutions for road safety and emergency services. Addressing these identified gaps will be crucial for advancing the field and ensuring the broader applicability and reliability of accident prevention and response technologies.

Table I shows a comprehensive comparison between existing works in the field, focusing on the evaluation of fall detection systems across different tasks. In the presented years, the tasks include linear and non-linear falls, normal rides, traversing bumpy surfaces, and sudden braking situations. The test times for each task are specified, and the outcomes are categorized into instances where an alarm was triggered and where it was not. The current work in 2023 demonstrates a fall detection system with a test time of 50 units for each task, consistently triggering alarms for falls and maintaining accuracy for other activities. A comparative analysis with previous works from 2012 to 2018 reveals varying performances, with differences in alarm accuracy for non-linear falls and sudden braking scenarios. This comprehensive overview aids in understanding the evolution and effectiveness of fall detection systems over time, providing valuable insights for further advancements in this critical domain.

TABLE I. COMPARISON BETWEEN EXISTING WORK

Year	Task	Test Time	Alarm	Not Alarm	Reference
2023	Linear fall	50	50	0	This work
	Non-Linear fall	50	50	0	
	Normal Ride	50	0	50	
	Bumpy surface	50	0	50	
	Brake Suddenly	50	0	50	
2012	Linear fall	100	100	0	[3]
	Non-Linear fall	100	99	1	
	Normal Ride	50	0	50	
	Bumpy surface	50	0	50	
	Brake Suddenly	50	2	48	
2012	Linear fall	100	100	0	[5]
	Non-Linear fall	100	99	1	
	Normal Ride	50	0	50	
	Bumpy surface	50	0	50	
	Brake Suddenly	50	2	48	
2014	Linear fall	100	100	0	[7]
	Non-Linear fall	100	99	1	
	Normal Ride	50	0	50	
	Bumpy surface	50	0	50	
	Brake Suddenly	50	1	49	
2016	Linear fall	100	100	0	[9]
	Non-Linear fall	100	100	0	
	Normal Ride	50	0	50	
	Bumpy surface	50	0	50	
	Brake Suddenly	50	0	50	
2018	Linear fall	100	100	0	[11]
	Non-Linear fall	100	100	0	
	Normal Ride	50	0	50	
	Bumpy surface	50	0	50	
	Brake Suddenly	50	0	50	

III. THE SYSTEM IMPLEMENTATION AND TESTING

The Vehicle Accident Detection and Alert System (VADAS) represent a critical innovation in road safety. With the increasing number of vehicles on the roads, ensuring the

safety of drivers and passengers has become a paramount concern. Traditional safety measures have been effective to some extent, but there exists a pressing need for more proactive and responsive systems that can promptly detect accidents and summon assistance. This section provides an overview of the VADAS, which utilizes a combination of cutting-edge technology components to enhance accident detection and emergency response.

These integral elements include the Arduino Uno Microcontroller, which serves as the central control unit, orchestrating module, and sensor interactions. The accelerometer sensor detects abrupt impacts or changes in acceleration, providing critical data to ascertain accident occurrences. An LCD display conveys crucial accident details such as location and time, promptly informing passengers and drivers. An audible alert system in the form of a buzzer plays a pivotal role in notifying vehicle occupants in the event of an accident. The integration of a GSM module enables seamless communication with emergency services and predefined contacts, facilitating message transmission and call requests for assistance. Precise location information is provided by the GPS module, expediting emergency responders' arrival at the accident scene. To minimize false alarms, the system incorporates a cancel button, allowing users to deactivate emergency calls and messages when necessary.

The operational framework of the VADAS is illustrated in Fig. 1. The system is centered around the Arduino Uno microcontroller, which coordinates the activities of the various components. The accelerometer sensor continuously monitors vehicle acceleration and based on this data, alerts the microcontroller to potential accidents. Upon accident detection, the microcontroller activates the alarm, sends emergency messages, and calls, and displays alerts on the LCD screen. Importantly, users have the option to cancel these alerts via the cancel button.

Fig. 2 presents the operational flow of the VADAS. It begins with the microcontroller's vigilant monitoring of the accelerometer sensor for sudden acceleration changes that may signify an accident. If such an event is detected, the microcontroller activates the alarm, alerting vehicle occupants. Simultaneously, it triggers the LCD display to convey the accident notification.

Subsequently, the microcontroller communicates with the GSM module, which sends emergency messages, including the accident's location, to predefined contacts. Notably, users could cancel these messages by utilizing the cancel button. If not cancelled, the system proceeds to notify emergency services, ensuring swift assistance. Following an accident, the microcontroller continues to monitor for any subsequent incidents. The VADAS represents a significant advancement in road safety technology, providing a proactive approach to accident detection and emergency response. In the subsequent sections, we will delve into the technical intricacies of this system and present the outcomes of its testing and validation.

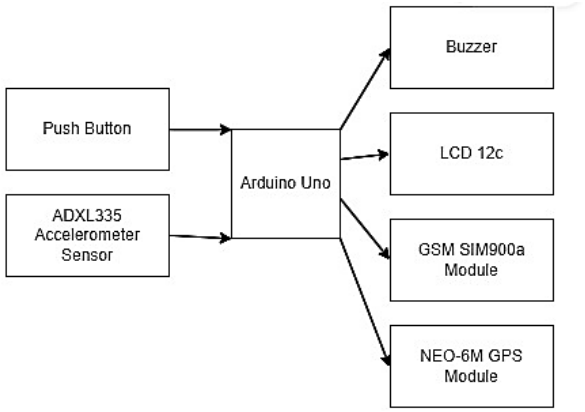


Fig. 1. Block diagram for vehicle accidental tracking system using micro-controller.

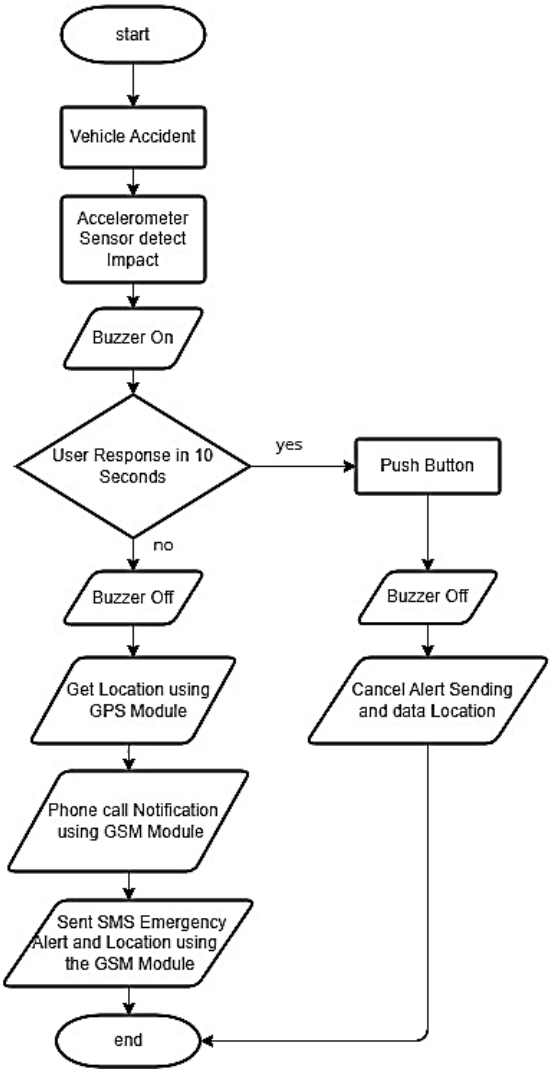


Fig. 2. Flowchart for vehicle accidental tracking system using micro-controller.

A. Hardware Implementation

The hardware implementation of the vehicle accident detection system comprises several key components:

1) Arduino Uno: At the heart of the system, the Arduino Uno takes center stage, featuring the ATmega328P microcontroller. With its array of digital and analog pins, this powerhouse proves its versatility across a spectrum of applications [27]. From LEDs to motors, sensors, and beyond, the Arduino Uno establishes seamless connections, all orchestrated through the user-friendly Arduino IDE. Its innate compatibility with shields amplifies its prowess, transforming it into the ideal candidate for ventures in robotics, home automation, and data logging projects. The Arduino Uno isn't just a board; it's a canvas for creativity — an open-source hardware masterpiece that beckons customization and replication with open arms [28].

2) ADXL 335 Accelerometer Sensor: Enter the realm of motion sensing with Analog Devices' 3-axis accelerometer, a silent guardian detecting shifts in acceleration, a pivotal role in the realm of accident detection. Thriving on low power (3V to 5V), it unveils a measurement prowess with a range of +/- 3g. The ADXL335 doesn't just measure; it brings stability to the temperature dance and hushes into a realm of low noise, making it the unsung hero for delicate applications like airbag deployment systems and the vigilant guardian for vibration monitoring adventures [29] [30].

3) GSM Module (SIM900A): Step into the world of connectivity with the SIM900A GSM/GPRS module from SIMCOM, a communication maestro that bridges the gap with emergency services and predetermined contacts. Boasting a built-in TCP/IP stack for seamless internet connectivity, it not only supports the quintessential SMS and call functionalities but also harmoniously integrates audio capabilities into its repertoire [31]. This wizardry, navigated through the realm of AT commands, finds its place not only in remote-control systems but also dances into the domains of data loggers and the enchanting world of home automation [32].

4) GPS Module (NEO-6M): The NEO-6M GPS module from Ublox provides accurate location, velocity, and time information. It can track up to 22 satellites simultaneously and supports multiple navigation modes, including GPS, GLONASS, and Galileo. With its small form factor and low power consumption, it's used in navigation systems, drones, and robotics.

5) LCD I2C: The LCD I2C is a cost-effective liquid crystal display that uses the I2C communication protocol. It simplifies displaying text and graphics, handling most display functionalities. It comes in various sizes (e.g., 16x2 or 20x4) and is widely used in data logging and temperature monitoring applications.

In the implementation of the vehicle accident detection system, the Arduino Uno serves as the central hub, orchestrating the interactions between the various components. The ADXL335 accelerometer is strategically positioned within the vehicle to continuously monitor acceleration changes. When a significant deviation is detected, suggesting a potential accident, the Arduino Uno triggers the SIM900A GSM module to initiate communication with emergency services and preconfigured contacts, alerting them to the situation. Simultaneously, the NEO-6M GPS module is activated to provide accurate location data, aiding in the quick dispatch of assistance [33]. The LCD I2C display acts as the user interface, conveying relevant information about the accident and the system's status. The modularity of the Arduino Uno allows for seamless integration and communication between these components, creating a cohesive and efficient accident detection and alert system for vehicles [34].

B. Software Implementation

The software implementation of the vehicle accident detection system encompasses various essential components and tools:

1) Arduino IDE: Step into the dynamic realm of creativity with the Arduino IDE (Integrated Development Environment), a software maestro empowering user to craft, upload, and debug code for the Arduino platform. Tailored to streamline the art of programming and device control, from microcontrollers to sensors and actuators, it spreads its wings across Windows, macOS, and Linux, offering a universal embrace. Rooted in Java and nurtured in the fertile grounds of the Processing development environment, the Arduino IDE unfolds its magic through a user-friendly interface adorned with a code editor, serial monitor, and library manager. Speaking the languages of C/C++ and Python, it extends its charm with an extensive library collection, turning every project into a masterpiece [35]. Embraced by hobbyists, educators, and professionals alike, it dances through the realms of robotics, home automation, and the vast expanse of IoT devices. A tool that evolves with the times, adorned with regular updates and a continual infusion of features, it remains a beacon for the creative spirits of our world [36].

2) Proteus: Developed by Labcenter Electronics, Proteus is a comprehensive software package tailored for computer-aided design of electronic circuits. This UK-based software combines schematic capture, simulation, and PCB layout tools, facilitating the design, testing, and layout of electronic circuits before physical construction. Proteus boasts an extensive library of simulation models, covering microcontrollers, microprocessors, sensors, actuators, and communication modules, enabling users to assess circuit behaviour and identify errors through simulation. The software also incorporates a PCB layout editor with features like auto-routing, 3D visualization, and support for surface mount and through-hole components, streamlining circuit board design, error checking, and manufacturing file generation. Proteus is widely embraced in both industry and education, catering to various microcontroller families like AVR, PIC, 8051, ARM, and supporting multiple communication protocols like I2C, SPI, UART, USB, and Ethernet. It is available in student and professional versions and operates within the Windows operating system.

3) SolidWorks: Crafted by the ingenious minds at Dassault Systems, SolidWorks emerges as a luminary in the realm of 3D mechanical CAD (Computer-Aided Design) software, weaving its expertise in solid modeling to breathe life into precise 3D models of parts and assemblies. A virtuoso in simulating their intricate dance of behavior and properties, SolidWorks steps into the limelight across an eclectic array of industries—from mechanical engineering and aerospace to the pulsating realms of automotive and consumer goods. Here, within the digital tapestry it weaves, users find the tools to sculpt 3D masterpieces of parts, intricate assemblies, and complex structures. This digital maestro, adorned with a rich palette of features, not only births vibrant 3D wonders but also unleashes tools for crafting intricate 2D drawings, orchestrating simulations of motion and stress, and orchestrating the symphony of data management. At its core lies the magic of parametric modeling, offering the fluidity of dynamic design changes and the allure of real-time updates. Detailed drawings, pulsating with dimensioning and tolerancing, are born effortlessly, accompanied by simulations donned in the garb of finite element analysis and motion simulations. SolidWorks, the polymath, extends its embrace to robust data management tools—from the meticulous bill of materials to the guardianship of revision control and the spirit of data sharing. In the hands of engineers, designers, and manufacturers, SolidWorks becomes not just a tool but a companion, navigating the intricate landscapes of creation and simulation. An educator at heart, it imparts wisdom to students, arming them with the prowess to sculpt precise designs and orchestrate simulations of grandeur [37]. In its digital wardrobe, SolidWorks adorns itself in multiple versions—Standard, Professional, and Premium—each tailored with its own tapestry of features and functionalities, a harmonious symphony compatible with the Windows operating system [38].

In the implementation phase, the vehicle accident detection system benefits from the synergy of these software tools. The Arduino IDE is utilized to write and upload the code that governs the behaviour of the Arduino Uno, the central control unit of the system. This includes the integration of code for processing data from the ADXL335 accelerometer, communication with the SIM900A GSM module, and handling data from the NEO-6M GPS module. Proteus comes into play by allowing simulation of the electronic circuitry, ensuring that the components interact seamlessly and function as intended. This pre-implementation testing minimizes errors and optimizes the performance of the system. SolidWorks is employed to create accurate 3D models of the physical components, aiding in the design and assembly of the hardware. The software's simulation features assist in predicting potential stress points or motion-related issues. The collective use of these software tools ensures a well-coordinated implementation of the vehicle accident detection system, minimizing errors, optimizing performance, and streamlining the development process [39].

IV. Result and Discussion

A. Hardware Development

Fig. 3 illustrates the system's circuit, offering a visual representation of its hardware components and connections. The hardware development stage is a pivotal component of the vehicle accident detection and alert system's creation.

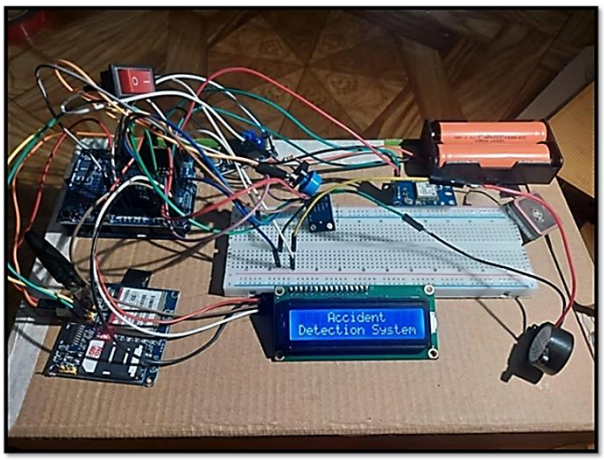


Fig. 3. The hardware connection.

The first crucial step in hardware development involves the careful selection of a microcontroller. The Arduino Uno stands out as the ideal choice due to its renowned ease of use, extensive community support, and rich library ecosystem. This microcontroller serves as the central processing unit for the accident detection and alert system.

To accurately detect accidents, a three-axis accelerometer sensor is integrated into the system. This sensor, capable of measuring acceleration in any direction, seamlessly connects to the Arduino Uno via analog input pins. Its precise measurements form the foundation for identifying potential accidents.

For displaying alerts and critical information, a standard 16x2 character LCD display is chosen. This display interfaces with the microcontroller through digital pins, providing a user-friendly interface to convey important messages related to detected accidents.

An appropriately loud buzzer is incorporated into the system to deliver audible alerts in the event of an accident. The buzzer connects to one of the Arduino Uno's digital pins, ensuring an immediate and attention-grabbing notification.

Facilitating emergency messages and calls is made possible through the integration of a GSM module. Chosen for compatibility with the Arduino Uno and its support for emergency communication, the GSM module interfaces with the microcontroller via serial communication pins.

Accurate location information is achieved through the inclusion of a GPS module. This module, selected based on its compatibility with the Arduino Uno, communicates with the microcontroller through serial communication pins, providing precise location data during emergencies.

A simple push button serves as the emergency cancel mechanism, allowing users to deactivate emergency alerts when necessary. This button connects to one of the microcontroller's input pins, providing a straightforward means to stop alerts and enhance user control.

Ensuring correct wiring of all components is vital to prevent damage. The system is powered by the vehicle's battery, with a voltage regulator employed to maintain a stable power supply. This meticulous approach to wiring and power management ensures the reliability and longevity of the hardware components.

By systematically following these hardware development steps, the vehicle accident detection and alert system's components are assembled cohesively. The resulting system is not only capable of accurately detecting accidents but also excels in promptly initiating emergency messages and calls, providing precise location information, and offering users a convenient means to deactivate emergency alerts when needed. This comprehensive approach enhances the overall usability and reliability of the system in real-world scenarios.

Fig. 4 illustrates the schematic diagram created using Proteus, providing a visual representation of the project's electronic components and their interconnections. This diagram serves as a valuable reference for understanding the system's design and layout.

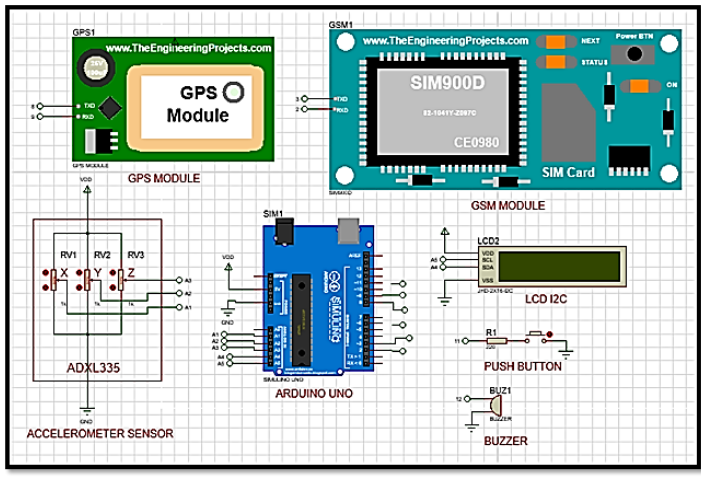


Fig. 4. Schematic diagram using proteus.

B. Prototype Design

The prototype design for the vehicle accident detection and alert system is a critical aspect. It involves careful placement and integration of various components, ensuring they are accessible, functional, and user-friendly. The following describes the design from different views:

Front View: The front view of the prototype prominently features the LCD screen, which serves as the primary interface for displaying alarms and information. The LCD display is strategically positioned within the vehicle to ensure easy visibility by the driver. Adjacent to the display, the cancel button is thoughtfully placed, making it easily reachable for the driver to deactivate emergency alerts if needed.

Side View: The side view showcases the placement of the microcontroller, an Arduino Uno, which serves as the central control unit. The microcontroller is positioned in a convenient location within the vehicle, ensuring accessibility for programming and debugging purposes. Additionally, the side view reveals the placement of the accelerometer sensor, which can be mounted on either the dashboard or the vehicle's floor, enabling accurate accident detection.

Top View: The top view of the prototype highlights the location of the buzzer, responsible for sounding alarms in case of accidents. The buzzer is positioned strategically to ensure that its sound is easily audible to the driver, enhancing the system's effectiveness. Furthermore, the top view reveals the placement of both the GSM module and the GPS module. These modules are situated for easy access within the vehicle, facilitating the rapid sending of emergency messages and calls, along with accurate location data.

Isometric View: The isometric view offers a comprehensive outlook on the overall layout of the system, encompassing all key components such as the microcontroller, accelerometer sensor, LCD display, buzzer, GSM module, GPS module, and the cancel button. The design prioritizes compactness and simplicity to ensure straightforward installation within a vehicle. Consideration is given to the power supply and the establishment of connections between components to maintain system functionality.

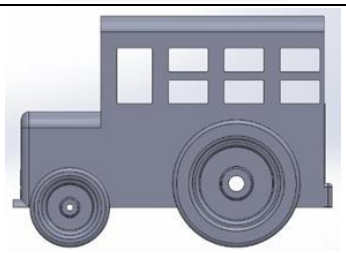
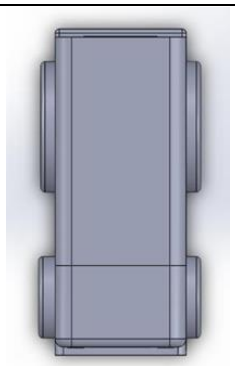
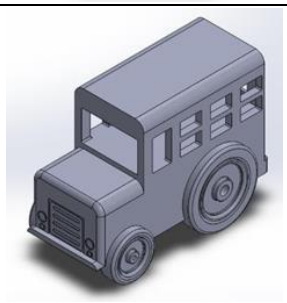
By adhering to this prototype design, the vehicle accident detection and alert system can effectively detect accidents, promptly initiate emergency messages and calls, and provide precise location information. The incorporation of the cancel button offers users a convenient means to deactivate emergency alerts in the event of a false alarm, optimizing usability and accessibility. The design ensures that all components are thoughtfully positioned, easy to access, and user-friendly.

Table II shows the prototype design using SolidWorks encompasses a comprehensive visualization presented through various views, including front, side, top, and isometric perspectives. SolidWorks, a powerful computer-aided design (CAD) software, allows for the creation of detailed and accurate three-dimensional models. The front view provides a frontal representation, the side view offers a profile perspective, the top view illustrates the design from above, and the isometric view presents a three-dimensional, angled depiction. These views collectively provide a holistic understanding of the prototype's geometry, dimensions, and features, aiding in the evaluation and refinement of the design before actual production.

C. The Proposed Work's Integration

Vehicle accident detection systems are vital for rapid response to accidents, minimizing injuries and property damage. This project integrates various components into a cohesive system designed to detect accidents and alert emergency services, with provisions for user control. The integration process involves several key steps:

TABLE II. PROTOTYPE DESIGN USING SOLIDWORKS

View	Image
Front View	
Side View	
Top View	
Isometric View	

1) Hardware connections: The initial step is to establish hardware connections, linking essential components to the Arduino Microcontroller using jumper wires. These components include the accelerometer sensor, LCD display, buzzer, GSM module, and GPS module, each serving a specific function in the system. A button is also integrated, connected to a digital input pin on the microcontroller, enabling user interaction.

2) Accelerometer sensor: The accelerometer sensor continuously provides data to the Arduino Uno, monitoring changes in acceleration. The Arduino program continuously reads and stores accelerometer data, allowing the microcontroller to detect abrupt acceleration changes that could indicate an accident.

3) Accident detection algorithm: An accident detection algorithm is implemented, involving the setting of a threshold value for accelerometer data. If the data surpasses this threshold, it triggers an "accident" event, signifying a sudden acceleration change. The threshold value's adaptability allows customization for different vehicle types and accident detection requirements, minimizing false alarms.

In Fig. 5(a), the display indicates the phrase "Crash Detected." This message is a critical notification to the driver and passengers that the system has detected an accident or a significant impact event. Displaying "Crash Detected" serves as an immediate alert, prompting occupants to take necessary actions and ensuring they are aware of the situation.

In Fig. 5(b), the display provides information related to the magnitude of the impact. This information helps convey the severity of the accident or impact event. Displaying the magnitude of impact can aid emergency responders in assessing the situation and providing appropriate assistance. It can also provide valuable data for post-accident analysis and insurance claims.

(a) (b)

Fig. 5. Display information (a) Crash detected and (b) Magnitude of impact.

These display messages play a crucial role in keeping the vehicle occupants informed about the detected accident and its severity. This real-time feedback contributes to improved situational awareness and facilitates appropriate responses to ensure safety and well-being.

4) Emergency response: Upon detecting an accident event, the Arduino responds by activating the buzzer and initiating SMS messaging via the GSM module. A pre-determined emergency contact receives an SMS containing the vehicle's GPS location and an accident notification. The GPS module determines the vehicle's precise location, facilitating emergency service response to aid potential accident victims.

In Fig. 6(a), the system is initiating an emergency call using the GSM module. This step is crucial for alerting emergency services about the accident or impact event promptly. The emergency call ensures that help is on the way to the location of the incident.

Fig. 6(b) displays an alert message along with location information. The alert message informs the recipient (likely an emergency contact or service) that an accident has occurred. Location information, likely obtained from the GPS module, is included. This information is vital for accurately pinpointing the vehicle's whereabouts. Providing location information helps emergency services quickly locate the vehicle, reducing response times and potentially saving lives.

Fig. 6(c) shows the location information obtained by the GPS module. GPS technology provides precise geographic coordinates, allowing emergency services to pinpoint the exact location of the vehicle involved in the accident. Accurate location data is essential for efficient and effective emergency response. These figures illustrate the critical steps of alerting emergency services, conveying accident information, and providing precise location details. Together, these actions are instrumental in ensuring a rapid and effective response to accidents, ultimately enhancing the safety of vehicle occupants.

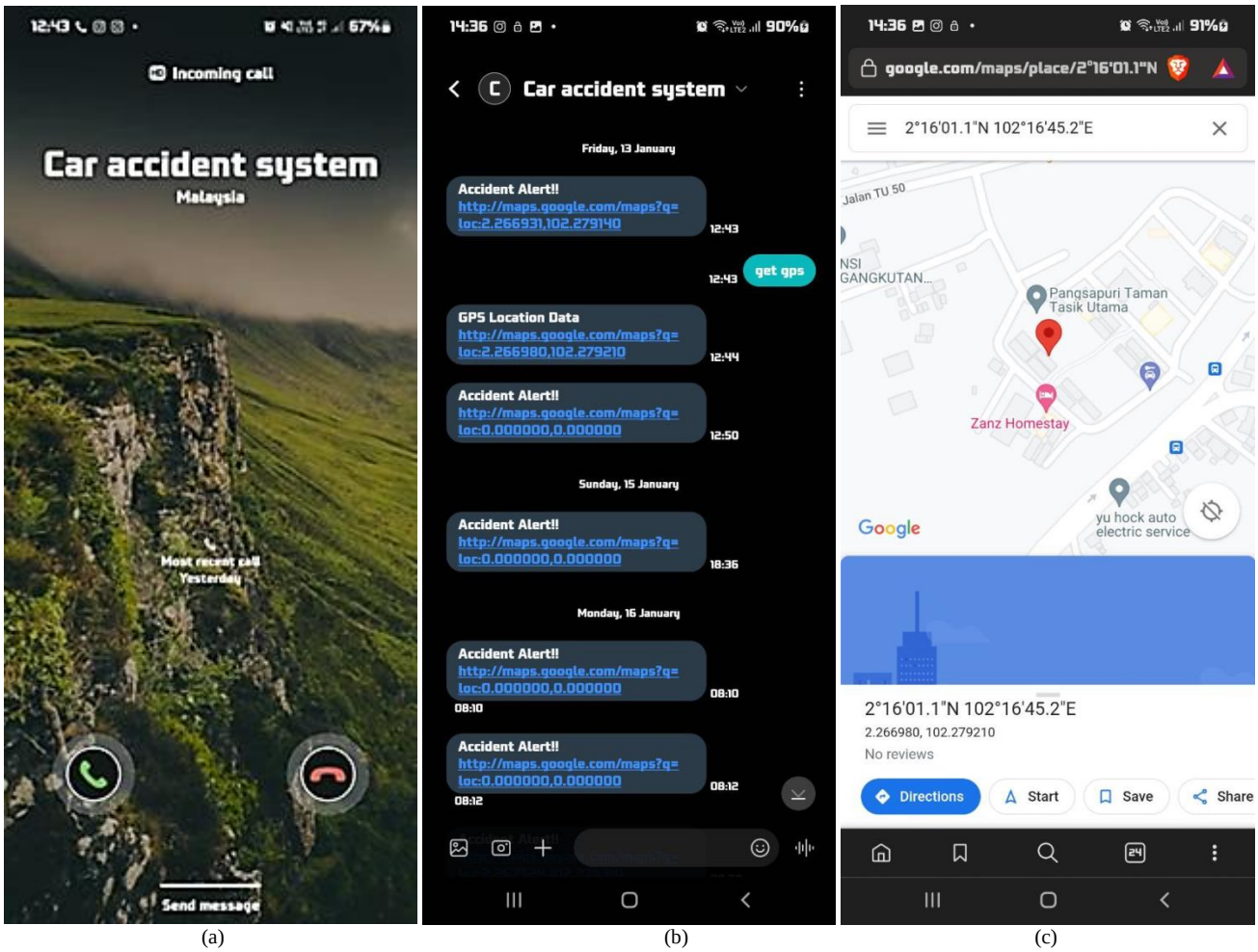
5) LCD Display: The LCD display serves to relay system status messages and information. It communicates events such as "Accident detected" and "Emergency message sent," providing transparency regarding system actions to vehicle occupants.

Fig. 7 shows the LCD display and it is like a small screen in your vehicle that tells you what's happening. It shows messages such as "Accident detected" or "Emergency message sent," so you know what the system is doing. This way, you can stay informed about important events while you're in the car. The display makes it easy for you to understand what's going on and helps you feel more confident and safe.

6) Cancel call and message: A dedicated button allows users to cancel initiated calls and messages in instances of false alarms or resolved emergencies. This feature minimizes unnecessary notifications and reduces false alarms.

Fig. 8 demonstrates the importance of user-friendly features like a dedicated push button for canceling alerts. It provides users with a straightforward way to manage the system and prevent unnecessary emergency calls and messages, contributing to the system's reliability and user satisfaction.

7) System loop: The Arduino program operates in a continuous loop, perpetually monitoring for accident events. The system continually assesses accelerometer data for changes in acceleration. In the event of an accident detection or button press, the system responds accordingly.



(a) (b) (c)

Fig. 6. Emergency call and location information (a) Emergency call using gsm module, (b) Alert message and location information and (c) Location obtained by GPS module.

Fig. 7. LCD display.

(a) (b)

Fig. 8. (a) Alert cancellation and (b) Push button.

D. Analysis of Limitations of the GPS Module and Magnitude of Impact

1) Limitations of the GPS module: The GPS module plays a crucial role in providing accurate location information for the vehicle accident detection and alert system. However, it is essential to understand and analyze its limitations, which can impact the system's performance. The following limitations of the GPS module have been identified:

Table III shows the performance of GPS modules is influenced by the surrounding environment, leading to distinct limitations in various settings. In open areas such as fields, the GPS module excels with clear sightlines, facilitating strong signal reception from satellites. When mounted on moving vehicles, the constant motion aids the antenna in searching for optimal signals, enhancing location accuracy. However, challenges arise in indoor environments where thick walls and building materials can weaken or block GPS signals, making it difficult to obtain reliable location data inside structures. Similarly, natural obstacles like trees and hills in outdoor settings, as well as the presence of tall buildings in urban canyons, can lead to weakened or blocked signals, rendering the GPS module less effective or even nonfunctional in these specific environmental conditions. These limitations highlight the need for alternative positioning technologies in scenarios where GPS signals are compromised.

TABLE III. LIMITATIONS OF THE GPS MODULE

Location	Limitation Signal	Reason
Open Area	Signals Reached	An open area with minimal obstructions, such as trees or buildings, allows the GPS module to receive strong signals from GPS satellites. The clear line of sight enhances signal reception.
Moving Vehicle	Signals Reached	Mounting a GPS module on a moving vehicle can improve reception. The constant movement of the vehicle helps the antenna continuously search for better signals.
Indoor Environment	Signals Weaken/Blocked	GPS signals can be blocked or weakened by thick walls, roofs, and building materials, making it challenging to obtain a reliable signal inside buildings.
Natural Environment	Signals Weaken/Blocked	Natural features like trees, hills, and other obstacles can block or weaken GPS signals, rendering them unreliable or unavailable in specific outdoor environments.
Urban Canyons	Signals Weaken/Blocked	Tall buildings and skyscrapers in urban areas can block or weaken GPS signals, leading to unreliable or unavailable signals in densely populated urban environments.

2) Analysis of magnitude of impact: Table IV show the magnitude of impact is a critical measurement provided by the accelerometer sensor. It quantifies the force experienced by the system in the event of an accident. In a 3-axis accelerometer, which measures acceleration along the x, y, and z axis, calculating the magnitude of acceleration is essential to assess the severity of the impact.

TABLE IV. ANALYSIS MAGNITUDE OF IMPACT

Times of Impact	X-Axis	Y-Axis	Z-Axis	Magnitude
Impact 1	43	86	69	118.35
Impact 2	64	30	42	82.22
Impact 3	38	68	50	92.56
Impact 4	100	64	99	154.59
Impact 5	22	57	59	84.94

The formula to calculate the magnitude of a 3-axis vector (ax, ay, az) is:

$$||(ax, ay, az)|| = \sqrt{}(ax^2 + ay^2 + az^2) \quad (1)$$

This formula calculates the Euclidean norm or magnitude of the acceleration vector by taking the square root of the sum of the squares of the individual components (ax, ay, az). Analyzing the magnitude of impact is of paramount importance for several critical reasons. Firstly, it aids in assessing the severity of accidents, with higher magnitudes indicating more forceful impacts, potentially signaling more severe accidents. Secondly, it plays a pivotal role in determining the need for emergency responses, such as alerting authorities or sending

emergency messages, based on the impact's magnitude. Additionally, understanding the magnitude of impact ensures that appropriate actions are promptly taken to safeguard the well-being of vehicle occupants and minimize injuries. Furthermore, it significantly contributes to enhancing the overall reliability of the accident detection system by ensuring that alerts are triggered when significant impacts occur. In essence, the analysis of impact magnitude empowers the system to make informed decisions about the activation of emergency response measures, thereby bolstering the system's effectiveness in mitigating the consequences of accidents.

E. Arduino Codes

1) Accelerometer integration: Fig. 9 shows the ADXL335 accelerometer, connected to the Arduino Uno, serving as the system's motion sensor. Utilizing the Adafruit_ADXL335 library, the Arduino continuously monitors changes in acceleration. The loop function computes the total acceleration magnitude and, upon surpassing a predefined threshold, triggers the emergency response function. This function, sendEmergencySMS (), initiates communication with the GSM module for alerting emergency services and predefined contacts about a potential accident.

```
#include <Wire.h>
#include <Adafruit_Sensor.h>
#include <Adafruit_ADXL335.h>
Adafruit_ADXL335 accel = Adafruit_ADXL335(12345);
void setup() {
  Serial.begin(9600);
  if(!accel.begin()) {
    Serial.println("Could not find a valid ADXL335 sensor, check
wiring!");
    while(1);
  }
}
void loop() {
  sensors_event_t event;
  accel.getEvent(&event);

  float acceleration = sqrt(event.acceleration.x*event.acceleration.x +
event.acceleration.y*event.acceleration.y +
event.acceleration.z*event.acceleration.z);

  if (acceleration > THRESHOLD_VALUE) {
    // Accident detected, trigger emergency response
    sendEmergencySMS();
  }
  delay(1000);  // Adjust delay based on your application's requirements
}
```

Fig. 9. Accelerometer integration codes.

2) GPS integration: Fig. 10 shows the NEO-6M GPS module interfaces with the Arduino Uno through SoftwareSerial to provide accurate location data. Using the TinyGPS++ library, the Arduino decodes NMEA sentences from the GPS module in the loop function. Upon validating a location fix, the system calls the sendGPSData function to store or transmit the GPS coordinates. This location data is crucial for emergency responders to locate the vehicle involved in the accident. The integration enhances the overall effectiveness of the accident detection system by providing precise location information.

```
#include <TinyGPS++.h>
#include <SoftwareSerial.h>

TinyGPSPlus gps;
SoftwareSerial ss(10, 11); // RX, TX

void setup() {
  Serial.begin(9600);
  ss.begin(9600);
}

void loop() {
  while (ss.available() > 0) {
    if (gps.encode(ss.read())) {
      if (gps.location.isValid()) {
        // GPS location available, store or send for emergency response
        sendGPSData(gps.location.lat(), gps.location.lng());
      }
    }
  }
}
```

Fig. 10. GPS integration codes.

3) GSM module integration: Fig. 11 shows the SIM900A GSM module facilitates communication with emergency services and predefined contacts. Utilizing SoftwareSerial, the Arduino continuously checks for incoming data in the loop function, employing AT commands to process SMS and call functionalities. In the event of an accident, the Arduino triggers the sendEmergencySMS function. This function uses AT commands to format and transmit an emergency SMS to specified contacts, containing information about the detected accident, including the location obtained from the GPS module.

```
#include <SoftwareSerial.h>

SoftwareSerial gsmSerial(7, 8);  // RX, TX

void setup() {
Serial.begin(9600);
gsmSerial.begin(9600);
}

void loop() {
if (gsmSerial.available()) {
// Process incoming SMS or calls
processIncomingData();
}
}

void sendEmergencySMS() {
// Use AT commands to send an emergency SMS
gsmSerial.println("AT+CMGF=1");  // Set SMS mode to text
delay(1000);
gsmSerial.print("AT+CMGS=\"");  // Set the recipient's phone number
gsmSerial.print(EMERGENCY_CONTACT_NUMBER);
gsmSerial.println("\"");
delay(1000);
gsmSerial.print("Accident detected at location: ");  // Send location
information
gsmSerial.println("LAT, LNG");
delay(1000);
gsmSerial.write(26);  // Send Ctrl+Z to indicate the end of the message
delay(1000);
}
```

Fig. 11. GSM module integration codes.

4) LCD display integration: Fig. 12 shows the I2C LCD display, connected to the Arduino Uno, offers a visual interface for real-time information. Using the LiquidCrystal_I2C library, the LCD is initialized in the setup function with a welcome message. In the loop function, the LCD is continuously updated with relevant information, such as current acceleration data and GPS coordinates. This integration provides an on-the-spot visual indication of the system's status, aiding users, and emergency responders in understanding the situation at a glance. The display serves as an essential component for user interface and system monitoring.

```
#include <Wire.h>
#include <LiquidCrystal_I2C.h>

LiquidCrystal_I2C lcd(0x27, 16, 2);  // I2C address 0x27, 16 column and 2
rows

void setup() {
  lcd.begin(16, 2);  // initialize the lcd
  lcd.print("Vehicle Monitor");
}

void loop() {
  // Display relevant information on the LCD
  lcd.setCursor(0, 1);
  lcd.print("Accel: ");
  lcd.print(acceleration);
  lcd.print(" GPS: ");
  lcd.print("LAT, LNG");
  delay(1000);
}
```

Fig. 12. LCD display integration codes.

V. Conclusion

In conclusion, the proposed vehicle accident detection and alert system, employing a microcontroller, accelerometer sensor, LCD display, buzzer, GSM module, GPS module, and cancel button, has proven to be effective in accurately detecting accidents, promptly alerting both passengers and emergency services, and providing vital location information in case of an accident. This system represents a robust and efficient solution, capable of enhancing response times during accidents and augmenting the overall safety and security of vehicle occupants.

To further enhance its capabilities, future development should focus on integration with complementary systems, such as cameras, voice recognition modules, or navigation systems, to provide more comprehensive accident information and assistance. Additionally, incorporating a cancel button for false alarm mitigation demonstrates attention to usability. The system effectively addresses the critical issue of accident detection and rapid alerting of emergency services, which holds great potential for reducing accident-related fatalities and injuries. It stands as a reliable and efficient system with scope for improvement and expansion in various directions. One noteworthy area for improvement lies in enhancing object detection accuracy, potentially through the implementation of Non-Maximum Suppression, leading to more precise region identification. The addition of a larger number of pre-trained models would also enhance performance. Furthermore, transitioning the system to a cloud-based platform would facilitate data storage and accessibility, allowing user guardians to access generated data and potentially integrating localization features for tracking user movement. In summary, the vehicle accident detection and alert system presented here offers a promising solution for enhancing road safety and emergency response, with ample room for future enhancements and broader applications.

VI. Limitation and Future Works

The vehicle accident detection and alert system, utilizing an Arduino microcontroller, accelerometer sensor, LCD display, buzzer, GSM module, GPS module, and cancel button, stands as a reliable and efficient system poised to enhance response times in accident scenarios while bolstering the safety and security of vehicle occupants. However, there exist several promising avenues for future improvements in the system's functionality and performance. One prospective area of development involves the integration of the system with other complementary technologies, such as cameras or navigation systems. This integration would significantly enhance the system's capabilities by providing additional information in the aftermath of an accident. For instance, a camera could capture vital images of the accident scene, while a navigation system could offer directions to the nearest hospital or emergency services. Another promising frontier for advancement is the implementation of machine learning algorithms to elevate the accuracy of accident detection. Machine learning algorithms, proficient in discerning patterns within sensor data, hold the potential to identify anomalous behavior effectively, thereby reducing the occurrence of false alarms and enhancing system precision. Furthermore, there is a possibility to imbue the system with remote monitoring functionality. This would enable real-time tracking of the vehicle's location and status, facilitating the dispatch of emergency alerts to a remote monitoring center. Such a feature would empower vehicle owners and emergency services alike to promptly locate and respond to accidents, potentially saving lives. Finally, the system's power efficiency can be optimized through the adoption of low-power components and the implementation of power-saving techniques like sleep mode and power management. These optimizations would extend the system's battery life, rendering it more practical for sustained use in vehicles. In summation, the vehicle accident detection and alert system already represents a potent and efficient solution for enhancing accident response times and improving the safety of vehicle users. Yet, with continued development and innovation, the system holds the potential to become even more robust and indispensable.

Acknowledgment

The authors would like to thank Centre for Research and Innovation Management (CRIM) for the support given to this research by Universiti Teknikal Malaysia Melaka (UTeM).

References

[1] A. Gunadal, A. Koujalagi, S. Karavade, and S. Konnuri, “Wireless black box using MEMS accelerometer and GPS tracking for accidental monitoring of vehicle,” project reference No. : 37S0430 college branch guide : s.g.balekundri institute of technology,belgaum, 2015.

[2] Chung-ChengChiu, Min-YuKu, Hung-Tsung, Chen Nat, "Motorcycle Detection and Tracking System with Occlusion Segmentation," Image Analysis for Multimedia Interactive Services. Santorini, vol. 2, pp. 32-32, June 2007.

[3] G. Gurjar, N. Chandurkar, R. Nagpure, S. Ganorkar, and J. Shelke, (March 2017) 'Wireless Black Box using MEMS Accelerometer and GPS Tracking for Accidental Monitoring of Vehicles', IJESC, vol 7(No 3).

[4] A. Musa, E. Alozie, S.A. Suleiman, J.A. Ojo, and A.L. Imoize, "A Review of Time-Based Solar Photovoltaic Tracking Systems," Information (Switzerland), vol. 14, no. 4, pp. 211, 2023. [Online]. Available: http://www.mdpi.com/journal/information/. doi: 10.3390/info14040211.

[5] N. Purushotham and J. Madan Kumar, "Wireless black box using MEMS accelerometer and GPS tracking for accidental monitoring of vehicles," International Journal of Scientific Engineering and Technology Research Volume.03, IssueNo.03, March-2014, Pages:0403-0408.

[6] B. S. Shobha and R. Deepu, "Hybrid Deep Learning Signature based Correlation Filter for Vehicle Tracking in Presence of Clutters and Occlusion" International Journal of Advanced Computer Science and Applications(IJACSA), 13(10), 2022, pp. 961-969. DOI:10.14569/IJACSA.2022.01310114.

[7] A. Kassem, R. Jabr, G. Salamouni, and Z. K. Maalouf, "Vehicle black box system," 2008 IEEE Int. Syst. Conf. Proceedings, SysCon 2008, no. May 2008, pp. 463–468, 2008, doi: 10.1109/SYSTEMS.2008.4519050.

[8] S. Sethuraman and S. Santhanalakshmi, "Implementing Vehicle Black Box System by IoT based approach," 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), Tirunelveli, India, 2020, pp. 390-395, doi: 10.1109/ICOEI48184.2020.9142906.

[9] M. Palanisamy, D. Azhagesan, B. Varadharaian and S. Kumar, "Design of Electronic Control Unit with Black Box for Ground Vehicles," 2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE), NaviMumbai, India, 2021, pp. 1-6, doi: 10.1109/ICNTE51185.2021.9487579.

[10] N. Watthanawisuth, T. Lomas and A. Tuantranont, "Wireless black box using MEMS accelerometer and GPS tracking for accidental monitoring of vehicles," Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics, Hong Kong, China, 2012, pp. 847-850, doi: 10.1109/BHI.2012.6211718.

[11] M. R. Wahid, E. Joelianto and N. A. Azis, "System Identification of Switched Reluctance Motor (SRM) Using Black Box Method for Electric Vehicle Speed Control System," 2019 6th International Conference on Electric Vehicular Technology (ICEVT), Bali, Indonesia, 2019, pp. 208-212, doi: 10.1109/ICEVT48285.2019.8994020.

[12] P. Yellamma, N. S. N. S. P. Chandra, P. Sukhesh, P. Shrunith and S. S. Teja, "Arduino Based Vehicle Accident Alert System Using GPS, GSM and MEMS Accelerometer," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 486-491, doi: 10.1109/ICCMC51019.2021.9418317.

[13] R. C. Pandia, (2016). Black Box: A Project Report Submitted in Partial Fulfillment of the Requirements for the Award of Degree of Bachelor of Technology in Electronics and Telecommunication. Orissa Engineering College.

[14] K. N. Kumar, C. Vishnu, R. Mitra and C. K. Mohan, "Black-box Adversarial Attacks in Autonomous Vehicle Technology," 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington DC, DC, USA, 2020, pp. 1-7, doi: 10.1109/AIPR50011.2020.9425267.

[15] M. M. Rahman, A. Z. M. T. Kabir, S. Z. Khan, N. Akhtar, A. Al Mamun and S. M. M. Hossain, "Smart Vehicle Management System for Accident Reduction by Using Sensors and An IoT Based Black Box," 2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Semarang, Indonesia, 2021, pp. 277-282, doi: 10.23919/EECSI53397.2021.9624240.

[16] C. PRABHA, R. SUNITHA, and R. ANITHA, 'Automatic Vehicle Accident Detection and Messaging System Using GSM and GPS Modem,' Int. J. Adv. Res. Electr. Electron.Instrum. Eng.," Int. J. Adv. Res. Electr. Electron. Instrum. Eng., vol. 3, no. 7, pp. 10723–10727, 2014, doi: 10.15662/ijareeie.2014.0307062.

[17] K. A. Hakim, M. M. Hasan, and S. Akter, "Automatic Vehicle Accident Detection and Messaging System Using GSM and GPS Module," B.Sc. Project, City University, Dhaka, Bangladesh, Spring 2019.

[18] M. Krishna Kanth, Moinuddin, and A. Kumar Singh, "Accident Detection and Vehicle Tracking Using Gps, Gsm And Mems," 2013. https://www.slideshare.net/krishnamoparthi/accident-detection-and-vehicletracking-using-gpsgsm-and-mems-16077971 (accessed Jun. 14, 2021).

[19] J. Mounika, N. Charanjit, B. Saitharun and B. Vashista, Accident Alert and Vehicle Tracking System using GPS and GSM (April 17, 2021). Asian Journal of Applied Science and Technology (AJAST) Volume 5, Issue 2, Pages 81-89, April-June 2021, Available at SSRN: https://ssrn.com/abstract=3869132.

[20] A. Kumari, "'Vehicle Accident Detection With GPS and Gsm Modem' Session 2016 Institute of Technology , Korba (C. G.) Department of Electrical and Electronics Engineering."

[21] R. Rishi, S. Yede, K. Kunal and N. V. Bansode, "Automatic Messaging System for Vehicle Tracking and Accident Detection," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 831-834, doi: 10.1109/ICESC48915.2020.9155836.

[22] N. Ahmmed, N. J. Jenny, M. Fowziya Akther Houya, A. I. Binte Alam and M. Adnan Arefeen, "VADet: An Arduino based automated vehicle accident detection and messaging system," 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Dhaka, Bangladesh, 2019, pp. 1-5, doi: 10.1109/ICASERT.2019.8934649.

[23] M. Anil Kumar, M. Venkata Suman, Y. Misra, and M. Geetha Pratyusha, "Intelligent vehicle black box using IoT," Int. J. Eng. Technol., vol. 7, no. 2, pp. 215–218, 2018, doi: 10.14419/ijet. v7i2.7.10296.

[24] V. Anumola, C. Pavan Raja Nadakuduru and K. Vadde, "Implementation of Adaptive Cruise Control and Cloud based Black Box Technology for Modern Automotive Vehicles," 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballar, India, 2023, pp. 1-6, doi: 10.1109/ICDCECE57866.2023.10150905.

[25] S. L. Ranjitha, A. S. Ristha, M. P. Shilpashree, and R. Aravind, "A Black Box with SMS Alert for Road Vehicles," vol. 6, no. 13, pp. 1–6, 2018.

[26] V. Anumola, C. Pavan Raja Nadakuduru and K. Vadde, "Implementation of Adaptive Cruise Control and Cloud based Black Box Technology for Modern Automotive Vehicles," 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballar, India, 2023, pp. 1-6, doi: 10.1109/ICDCECE57866.2023.10150905.

[27] K. K. M. Rahman, M. M. Subashini, M. Nasor and A. Tawfik, "Development of bio-shields for Arduino Uno," 2018 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, Sharjah, Abu Dhabi, United Arab Emirates, 2018, pp. 1-5, doi: 10.1109/ICASET.2018.8376901.

[28] Y. Kyrylenko, I. Kameneva, O. Popov, A. Iatsyshyn, V. Artemchuk, V. Kovach, (2020). Source Term Modelling for Event with Liquid Radioactive Materials Spill. In: Babak V., Isaienko V., Zaporozhets A. (eds) Systems, Decision and Control in Energy I. Studies in Systems, Decision and Control, vol 298, pp 261-279. https://doi.org/10.1007/978-3-030-48583-2_17.

[29] N. I. M. Amir, R. A. Dziyauddin, N. Mohamed, L. A. Latiff and N. S. N. Ismail, "Development of Fall Detection Device Using Accelerometer Sensor," 2021 IEEE 7th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), Bandung, Indonesia, 2021, pp. 93-98, doi: 10.1109/ICSIMA50015.2021.9526289.

[30] O. Popov, V. Iatsyshyn Andrii, Kovach, V. Artemchuk, I. Kameneva, O. Radchenko, K. Nikolaiev, V. Stanytsina, A. Iatsyshyn, Y. Romanenko. Effect of Power Plant Ash and Slag Disposal on the Environment and Population Health in Ukraine. Journal of Health and Pollution 11(31), 210910 (2021). https://doi.org/10.5696/2156-9614-11.31.210910.

[31] S. Mukherjee, A. Ghosh and S. K. Sarkar, "Arduino based Wireless Heart-rate Monitoring system with Automatic SOS Message and/or Call facility using SIM900A GSM Module," 2019 International Conference

on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), Vellore, India, 2019, pp. 1-5, doi: 10.1109/ViTECoN.2019.8899504.

[32] O.O. Popov, A.V. Iatsyshyn, A.V. Iatsyshyn, V.O. Kovach, V.O. Artemchuk, V.O. Gurieiev, Y.G. Kutsan, I.S. Zinovieva, O.V. Alieksieieva, V.V. Kovalenko, A.E. Kiv, 2021. Immersive Technology for Training and Professional Development of Nuclear Power Plants Personnel. In: Lytvynova S.H., Semerikov S.O. (ed.) Proceedings of the 4th International Workshop on Augmented Reality in Education (AREdu 2021), Kryvyi Rih, Ukraine, May 11, 2021. CEUR Workshop Proceedings 2898, 230–254. http://ceur-ws.org/Vol-2898/paper13.pdf.

[33] S. Mukherjee, A. Ghosh and S. K. Sarkar, "Arduino based Wireless Heart-rate Monitoring system with Automatic SOS Message and/or Call facility using SIM900A GSM Module," 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), Vellore, India, 2019, pp. 1-5, doi: 10.1109/ViTECoN.2019.8899504.

[34] I.S. Zinovieva, V.O. Artemchuk, A.V. Iatsyshyn, Y.O. Romanenko, O.O. Popov, V.O. Kovach, D.V. Taraduda, A.V. Iatsyshyn. The Use of MOOCs as Additional Tools for Teaching NoSQL in Blended and Distance Learning Mode. Journal of Physics: Conference Series 1946, 012011 (2021). https://doi.org/10.1088/1742-6596/1946/1/012011.

[35] D. K. Halim, T. C. Ming, N. M. Song and D. Hartono, "Arduino-based IDE for Embedded Multi-processor System-on-Chip," 2019 5th International Conference on New Media Studies (CONMEDIA), Bali, Indonesia, 2019, pp. 135-138, doi: 10.1109/CONMEDIA46929.2019.8981862.

[36] R. Tkachenko, P. Tkachenko, I. Izonin, P. Vitynskyi, N. Kryvinska, and Y. Tsymbal, 'Committee of the Combined RBF-SGTM Neural-Like Structures for Prediction Tasks', in Mobile Web and Intelligent Information Systems, vol. 11673, I. Awan, M. Younas, P. Ünal, and M. Aleksy, Eds., in Lecture Notes in Computer Science, vol. 11673. , Cham: Springer International Publishing, 2019, pp. 267–277. doi: 10.1007/978-3-030-27192-3_21.

[37] Y. Sokolovskyy, O. Sinkevych and R. Voliansky, "Software for Studying Wood Drying Chambers Based on SolidWorks Flow Simulation Experiment," 2019 9th International Conference on Advanced Computer Information Technologies (ACIT), Ceske Budejovice, Czech Republic, 2019, pp. 281-284, doi: 10.1109/ACITT.2019.8780040.

[38] I. Fedorchenko, A. Oliinyk, Jamil Abedalrahim Jamil Alsayaydeh*, A. Kharchenko, A. Stepanenko and V. Shkarupylo, 2020. Modified Genetic Algorithm to Determine the Location of the Distribution Power Supply Networks In The City. ARPN Journal of Engineering and Applied Sciences. (VOL. 15 NO. 23) (pp 2850-2867).

[39] Adam Wong Yoon Khang, Shamsul J. Elias, Nadiatulhuda Zulkifli, Win Adiyansyah Indra, Jamil Abedalrahim Jamil Alsayaydeh, Zahariah Manap, Johar Akbar Mohamat Gani, 2020. Qualitative Based QoS Performance Study Using Hybrid ACO and PSO Algorithm Routing in MANET. Journal of Physics, Conference Series 1502 (2020) 012004, doi:10.1088/1742-6596/1502/1/012004.

Design of University Archives Business Data Push System Based on Big Data Mining Technology

Zhongke Wang, Jun Li
Chengdu Technological University
Chengdu, Sichuan Province, 611730, China

***Abstract*—Aiming at the problems of low accuracy, recall, coverage and push efficiency of university archives business data, a university archives business data push system based on big data mining technology is designed. Firstly, the overall architecture and topological structure of the university archives business data push system are designed, and then the functional modules of the system are designed. Using big data mining technology to mine user behavior, modeling according to user behavior sequence, and designing a model to predict user behavior sequence based on hidden Markov model theory. Finally, the user behavior sequence is analyzed, and the factors such as user collaboration, similarity of user behavior sequence and data timeliness are comprehensively considered to push university archives business data for users. The experimental results show that the proposed method has high data push accuracy, recall, coverage and push efficiency, and can effectively push the required business data for users.**

Keywords*—Big data mining technology; system design; business data pus; hidden markov model; similarity

I. INTRODUCTION

With the rapid development and popularization of information technology, the student file management information system in colleges and universities mainly manages the information related to student files. Relying on the basic platform of campus network, it plays an increasingly important role in teaching and management applications. As a powerful tool, big data mining technology can help university archives departments to better manage and analyze massive data, and discover hidden laws and knowledge from it. The traditional student file management information system generally adopts the client/server architecture mode or browser/server architecture mode, and adopts the centralized management mode. All the information is stored in a database, and the scale is getting larger and larger in colleges and universities. Many colleges and universities have multiple campuses. This mode obviously cannot meet the requirements of improving the efficiency of college student file management. With the rapid development of Internet technology, the data produced and faced by Internet users are increasing, which makes people face the dilemma of "information ocean". Therefore, the recommendation system came into being and became the first choice to help users filter effective information from massive information [1], [2]. When the recommendation system is applied to the data push process of archives business in colleges and universities, users' experience is not good because of the sparse data of archives business, one-sided similarity calculation of users or items and poor real-time recommendation results [3]. Therefore, it is of great significance to study and design an effective data push system for university archives business. The design goal of the push system is to provide an efficient, flexible and intelligent data push platform to help university archives departments manage and utilize data better. The system analyzes and mines the archives business data of colleges and universities through big data mining technology, thus providing valuable information and insight. We will also customize the push content for each user according to individual needs to ensure that users can get the most relevant information in time.

Pan, H et al. [4] proposed a five-layer push system framework, which is divided from bottom to top: the perception layer, the filter layer, the sorting layer, the rule layer and the application layer. At the same time, context awareness technology is applied to the data push process to achieve the push of business data. This method has the problem of low accuracy of data push. Zheng, Y et al. [5] reduced the initial data set by using the characteristics of static Sky-line points. Then, according to the characteristics of the searcher moving in the obstacle space, a distance intersection model is constructed, and a pruning strategy is proposed by using the model and the attributes of the data object. According to the pruning strategy, the data objects that have no influence on the query results when the inquirer moves are filtered, so as to reduce redundant data and get the filtered candidate data set. Finally, according to the non-spatial attributes of data objects and the characteristics of mutual dominance, the events that affect the candidate data set are determined, and these events are used to refine the candidate data set, further reducing redundant calculations and obtaining the result set at the current moment. Jelodar, H et al. [6] established an automatic data recommendation system based on user interest classification. The system includes offline module and online module. The offline module processes and cleans the historical storage data, extracts the user's interest characteristics, and forms the user's interest database submodule; The online module uses the recommendation engine to cluster and correlate the information in the user interest database sub-module, and finds the book information similar to the user interest data, forming a preliminary recommendation set sub-module. This method takes a long time to push data. Moreover, the pushed data is too single, which leads to the problems of low push efficiency and low coverage. Panda et al. [7] proposed to realize effective access control of medical data with forward and backward confidentiality, and medical service providers stored patients' electronic medical records in the cloud to provide high-quality

medical services. Attribute-based encryption is a promising encryption technology, which can realize fine-grained access control of outsourced encrypted data. In this paper, an attribute-based encryption scheme is proposed to support the update of access policy, and it also provides forward security, backward security and user revocation. Performance analysis shows that the scheme has high communication and computing ability, and is suitable for devices with limited resources, but the push accuracy of this method is low. Nour et al. [8] proposed the access control mechanism in the named data network, and the information center network was recently proposed as an important candidate for the future Internet architecture to solve the problems existing in the current Internet host-centric communication model based on TCP/IP. This paper provides a detailed and comprehensive investigation of access control mechanism in named data network. The access control in named data network is studied by comprehensive method. Firstly, the paradigm of information center network is summarized, the change from channel-based security to content-based security is described, and different encryption algorithms and security protocols in named data network are introduced. Then, we divide the existing access control mechanisms into two categories: access control based on encryption and access control independent of encryption. Each category is classified according to the working principle of access control. Finally, the experience and lessons of the existing access control mechanism are summarized, and the challenges of access control based on named data network are pointed out, and the future research direction is emphasized, but the push recall rate of this method is low. Zhang et al. [9] put forward the monitoring of industrial sewage outlet water pollution based on web crawler and remote sensing interpretation technology, and took Luanhe River basin as the experimental point, combined with web crawler and remote sensing interpretation technology, put forward a feasible and efficient method to obtain sewage outlet data. Grab industrial information and spatial location data on the Internet, and get the location of industrial sewage outlet through remote sensing image interpretation. The distribution of main industrial sewage flowing into the tributaries of Luanhe River basin is simulated. By comparing the results with the actual data collected during the field investigation, the accuracy and reliability of the developed method are verified, but the push coverage rate of this method is low.

In order to solve the problems in the above methods, a design method of university archives business data push system based on big data mining technology is proposed. This method designs the overall architecture and topological structure of the university archives business data push system, and analyzes the functional modules of the system. Using big data mining technology to mine user behavior, modeling according to user behavior sequence, and designing a model to predict user behavior sequence based on hidden Markov model theory. On this basis, the user behavior sequence is analyzed, and the factors such as user collaboration, similarity of user behavior sequence and data timeliness are comprehensively considered to push university archives business data for users. The research shows that the proposed method has high data push accuracy, recall, coverage and push efficiency, and can effectively push the required business data for users with good push effect.

II. System Design

A. Overall System Architecture

The client side of the university archives business data push system adopts the client server mode, namely C/S, and the system adopts the B/S (Browser/Server) mode on the browser side. In general, it follows the classic three-tier architecture, which includes the presentation layer, application layer, and data layer. The system architecture of the university archives business data push system is shown in Fig. 1.

As it can be seen from Fig. 1, that the presentation layer is mainly the university archives business data push mobile terminal and PC browser. The application layer, situated in the middle of the three-tier architecture, consists of the control layer, the business logic layer, and the basic service layer. At the bottom is the data layer, including the data access layer and the data storage layer.

The display layer is mainly Android mobile terminal and PC terminal browser, which respectively displays user login interface and administrator login interface. The user end views the file business by logging into the Android mobile APP. The implementation method mainly uses the Android Activity and Fragment drawing interface, and the implementation architecture mainly relies on the Android operating system. The PC terminal mainly manages the system by logging in to the WEB page, which is realized by JavaScript+HTML. The execution architecture consists of a browser and an operating system.

The application layer is the main part of the data push system, which includes three parts to complete different operations. The control layer includes the control of terminal access, which is mainly composed of parameter analysis and session management. The control layer analyzes the file business type, and hands over specific operations to the business logic layer for processing; The business logic layer includes the specific services of user management, data publisher management, data management, intelligent push management and other functional modules. The basic service layer includes resource caching, resource access control, web crawler and vectorization processing, as well as data processing services and data analysis services. The functional modules of the business logic layer are implemented based on the construction in the basic service layer. The basic services offered provide functional support for the aforementioned business operations. The mobile terminal is realized through MVP mode, and the PC terminal is realized through Spring+Spring MVC+MyBatis. The execution architecture is composed of browser and operating system.

The data layer is composed of two components: the data access layer and the data storage layer. It mainly stores user data, archive business publisher data, business data, comment data, etc. of the system. The operation of structured data in the system is completed by MySQL database. Unstructured data is stored in files by data preprocessing.

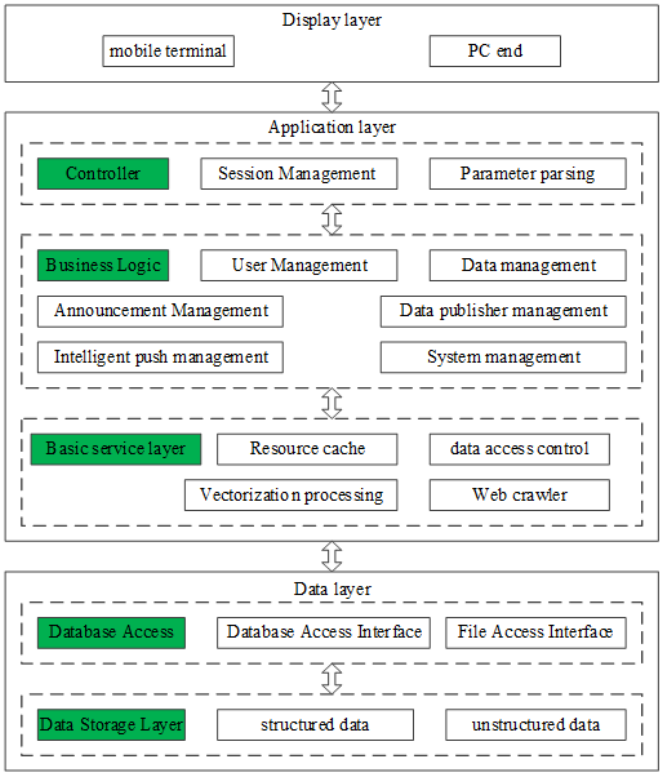


Fig. 1. Overall architecture of university archives business data push system.

In order to maintain the complete extensibility of the system, the system is divided into Controller layer, Business layer, Service layer and DAO layer when implementing the PC side in combination with the specific SSM framework. The business publisher logs in to the system through the browser, accesses the JSP page, and calls the service of the Service layer through the Controller layer.

The request sent by the business publisher to the server first enters the Controller layer. Through the controller layer's control of different businesses, dif-ferent businesses can be distributed to the Business layer to call specific business logic code. The Busi-ness layer encapsulates the database operation ser-vice and the interface provided by the Service layer to the outside. The database operation service can directly call the DAO encapsulated by Mapper. The specific SQL execution statement is implemented in XML. The results are returned to the browser for dis-play through the database operation service.

B. System Topology

Since the university archives business data push system is aimed at users in different geographical locations, the business publisher on the PC side also publishes business in different work areas [10], [11]. The work nodes are distributed in different locations, so the university archives business data push system designed by the proposed method adopts a star to-pology structure, as shown in Fig. 2.

The topological structure of this system is mainly divided into three parts: the network segment where the university archives business data push system is located, the network segment where the third-party organization is located, and the part that communi-cates directly with the Internet. The system is set up inside a communication, with separate WEB server and database server. The WEB server completes the request and response of PC end business publishers and mobile terminal users to the system functions. The database server stores the data information used in the system, including relational data and non-relational file data. At the same time, to ensure concurrent access to the system, the proposed method uses Alibaba Cloud's nginx load balancing to connect to multiple WEB servers. In this network segment, the three roles of system administrator, business administrator and communicator are divided to manage the business information and user information of the system. The roles in the system are connected to the system through the firewall. External users do not need to go through the firewall to access the system, which ensures the speed of external users' access and the security of the internal network.

Through this topology design, different access modes are set for various roles in the system.Whether tourists or registered users, authenticate users [12], [13], use mobile terminals to register and log in to the sys-tem through wireless settings, and complete corresponding queries and other operations. There are three types of file business publishers: correspondent, organization publisher, and individual publisher. The communicator is in the system network segment, the organization publisher is in the third-party organization Ethernet, and the individual publisher is free and can be directly connected to the Internet. The system administrator and news administrator are in the LAN of the system layout to manage user information and file business information.

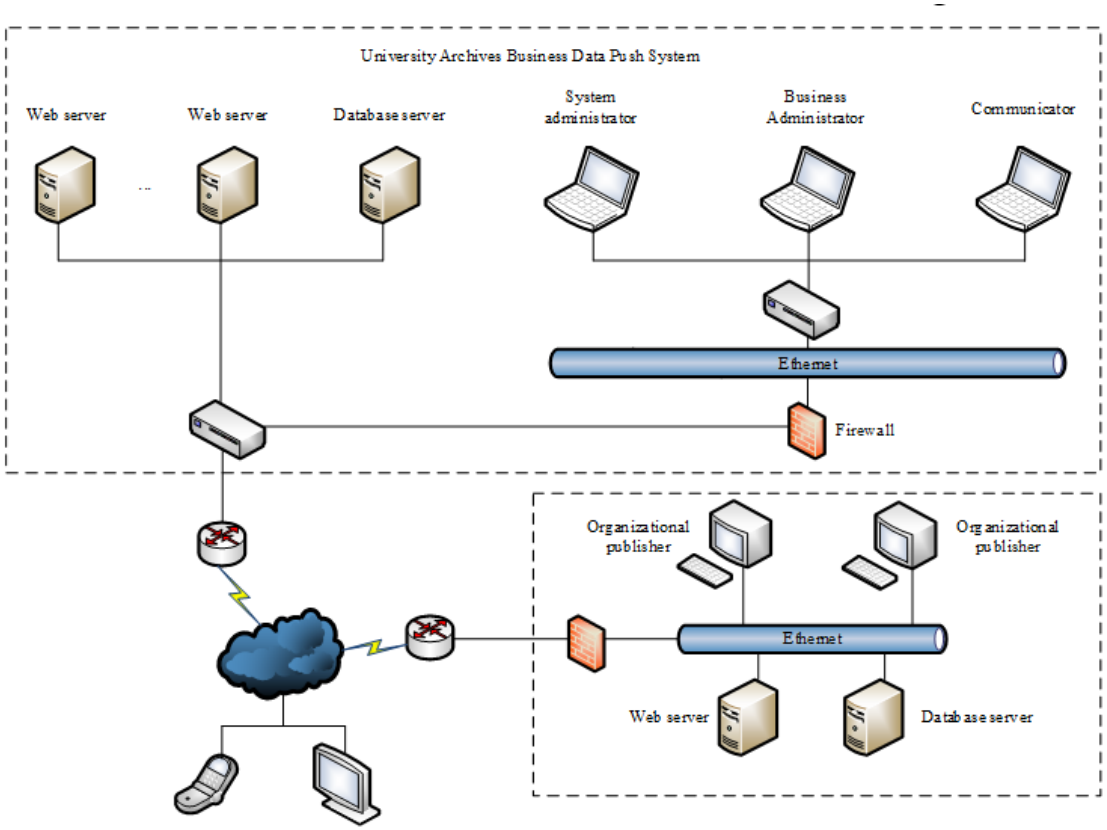


Fig. 2. System network topology.

C. System Function Module

According to the requirements of archives business in colleges and universities, the functional modules of archives business data push system in colleges and universities are divided, and the specific steps are as follows:

Step 1: Demand analysis: It deeply understand the specific needs of university archives business, including interviews with university archives managers, questionnaires or research on existing business processes, with the goal of clarifying the specific requirements and objectives that the system needs to meet.

Step 2: Function identification: Based on the demand analysis, identify the core functions required by the data push system of university archives business to meet the daily operation and management needs of university archives business.

Step 3: Module division: It group and modularize the identified functions. The division of modules should follow the principle of high cohesion and low coupling, so as to ensure that the functions of each module are relatively independent but can work together. The function of the data push system of university archives business is shown in Fig. 3.

1) Push management: You can query the push task records of push users and the records of accessing push pages; Users can also manually add push accounts to the policy server; At the same time, the user can also query the push task record of the active push customer and the record of accessing the push page [14].

2) White list management: This module is mainly responsible for adding users to the white list if they do not want to receive page push reminders, and then adding and deleting the white list if they do not want to be pushed in the future.

3) Customized push management: This module is responsible for adding and modifying the customized push task of the recommendation file business, and is responsible for performing association rule function analysis [15], [16] in the background, and applying the results to the foreground to batch add or delete users of customized recommendation products, and formulating push rules.

4) Push statistics: This module is mainly used to make statistics and analysis on the business data records of each push file in the push system, and present the results to the administrator in the form of a summary table.

5) System management: The system management function module primarily encompasses various functions such as system department management, user management, authority management, password modification, role management, and log query.

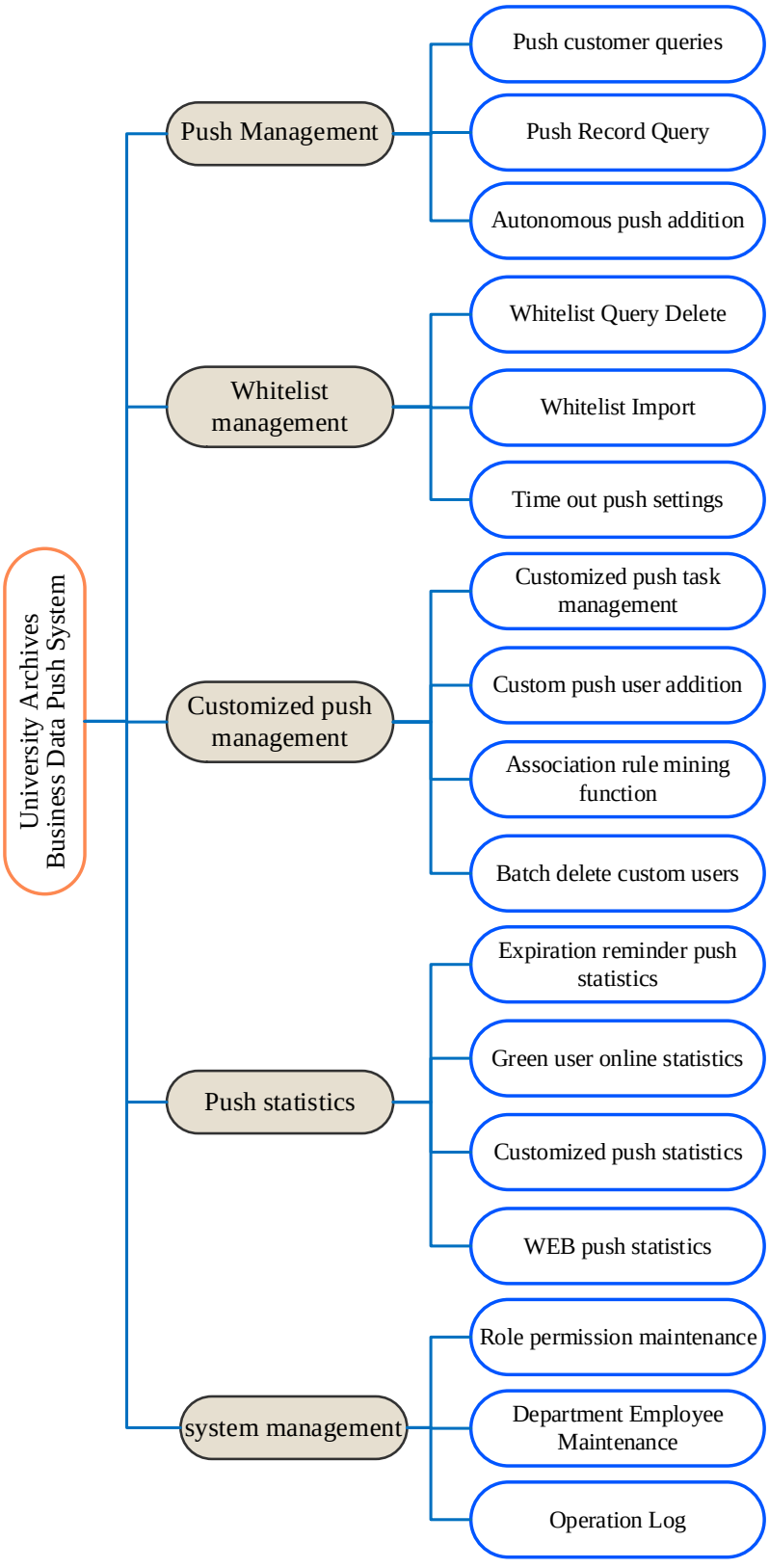


Fig. 3. Function division of college archives business data push system.

III. The Method and Implementation of College Archives Business Data Push

A. User Behavior Analysis

Considering that each user behavior sequence may represent a certain "interest" of the user, and the "entity" of the "interest" is the university file business page pointed to by the behavior sequence. Therefore, user behavior templates are used to record various "interests" of users and their corresponding behavior sequences. The generation rule is to take the user identifier as the name of the file. Each line in the file records a behavior sequence and the identifier of the file business that the behavior sequence points to. This file can be generated synchronously in the process of user behavior sequence extraction.

B. Similarity Calculation

The university archives business data push system designed by the proposed method mainly adopts the idea of user collaborative filtering [17], [18]. The recommendation system based on collaborative filtering needs to find the collection of items or users' nearest neighbor points through similarity calculation, and then carry out the next recommendation work according to the relevant information of these nearest neighbor points. The search of recommended file business data is divided into the following two steps:

- Find the nearest neighbor user of the current user through user feature similarity calculation;
- From the behavior sequences of these neighboring users, we can calculate the similarity of user behavior sequences to find the file business data that the current user may be "interested" in.

1) User feature representation and user feature file: The main data input that the system relies on is the various "filtering and provocation" selected by the user when searching, so that users and data, users and users, and data and data in the system can be associated through user behavior sequences, which can reflect the characteristics of users or data. Therefore, the proposed method takes the user behavior sequence as the original data, uses the vector space model to represent the user, and converts the feature vector β expressed as:

$$\beta = \{(g_1, w_1), (g_2, w_2), \cdots, (g_n, w_n)\} \quad (1)$$

Among them, w_i is the i feature items weight of g_i, indicating that the current user's behavior sequence or behavior template the appears times of g_i ; Characteristic item $g_i$indicates a behavior included in the user behavior sequence; n indicates the type of user behavior in the system.

In order to improve the efficiency of user collaborative search, the proposed method records the feature vectors of all historical users in the form of feature files. This file is a text file, each line of which is composed of the historical user feature vector exported from the user behavior template and the number of this behavior template.

2) User similarity calculation: The common vector-based similarity calculation methods in recommendation systems are as follows:

a) Cosine similarity: Cosine similarity is a commonly used method in information retrieval for measuring the similarity between two documents [19], [20]. It involves treating the feature vectors of the documents as vectors. The cosine similarity is determined by the angle between these two vectors, with a smaller angle indicating a higher degree of similarity, the more parallel they tend to be, and parallelism means they are completely similar. The calculation formula is as follows:

$$s_{im}(u_a, u_b) = \frac{\beta \times \sum_{i=1}^{n} w_{ai} \times w_{bi}}{\sqrt{\sum_{i=1}^{n} w_{ai}^2 \times \sum_{i=1}^{n} w_{bi}^2}} \quad (2)$$

Among them, u_a, u_b represent users eigenvector of a, b; w_{ai},w_{bi} respectively represent eigenvectors u_a,u_b of elements i.

The advantage of cosine similarity is that it is not affected by the size of vector module and the calculation is simple. However, when encountering high-dimensional sparse vectors, the accuracy of similarity calculation will decline.

b) Pearson correlation coefficient similarity: In cosine similarity calculation, two user feature vectors are regarded as two independent variables, but in reality, there may be some connection between them. Especially in the scoring vector based on "user item", there is a certain relationship between different users' ratings of items. Therefore, Pearson correlation coefficient is proposed to measure the approximation of two users [21]. The calculation formula is as follows:

$$s_{im}(u_a, u_b)' = \frac{\beta \times \sum_{s \in S_{ab}} (w_{as} - m_a) \times (w_{bs} - m_b)}{\sqrt{\sum_{s \in S_{ab}} (w_{as} - m_a)^2 \times \sum_{s \in S_{ab}} (w_{bs} - m_b)^2}} \quad (3)$$

Among them, S_{ab} indicates the user a, b subscript set of common features; s express an element in S_{ab} ; w_{as} , w_{bs} respectively represent eigenvectors u_a, u_b of elements s; m_a, m_b respectively represent the average weight of u_a, u_b.

Considering that in the process of college archives business data push, the feature vector of online users is extracted from a behavior sequence, and a behavior sequence contains fewer repeated behaviors, the obtained feature vector of online users is basically a binary vector. Therefore, the similarity calculated by Pearson correlation coefficient is basically equivalent to cosine similarity in effect, but its calculation process is obviously more complex than cosine similarity, so the proposed method uses the cosine value of the vector as the similarity of user characteristics.

3) Similarity calculation of behavior sequence: In general, the number of items found through the current user's nearest neighbor set is still quite considerable, and through previous analysis, it can be seen that each user's behavior sequence is associated with at least one university file, and different behavior sequences of the same user may reflect the user's more subtle interests and preferences. Therefore, the current

user's interest preferences can be more accurately predicted under the direction of the adjacent user's behavior sequence.

An action sequence mainly contains two aspects of information: the relative order in which the user actions in the sequence occur and the value of the actions. Therefore, the similarity of two behavior sequences can be measured from two perspectives, namely, behavior similarity and behavior value similarity.

Define S a behavior sequence the behavior state string of B, which only contains the name of the behavior in B, without the value of the corresponding behavior. Meanwhile in S, the string representing a behavior name is regarded as a state, which is an atomic weight and is indivisible. The calculation formula of the behavioral sequence B_1, B_2 of the behavior similarity of $X_{simseq}(s_1, s_2)$ is as follows:

$$X_{simseq}(B_1, B_2) = \frac{max[L_{len}(B_1), len(B_2)]}{max[L_{len}(B_1, B_2)]} \times sim(u_a, u_b)' \quad (4)$$

Among them, B_1, B_2 represents a common subsequence; $L_{len}(\cdot)$indicates the number of states contained in the behavior state string; $max[\, L_{len}(B_1, B_2)]$ express the number of states s_1, s_2 contained in the maximum common subsequence of.

Values of the same behavior are comparable, and the relative order factors between states should also be included in the scope of value similarity. So, the behavior sequence B_1, B_2 value similarity of $S_{simvalue}(B_1, B_2)$ is based on the maximum common subsequence of s_1, s_2, the calculation formula is as follows:

$$S_{simvalue}(B_1, B_2) = \frac{max[(X_{simseq}(B_1, B_2)]}{max[S_{subseqcount}(B_1, B_2)]} \quad (5)$$

Among them, $S_{subseqcount}(B_1, B_2)$ express the number of the same public status values of B_1, B_2.

Finally, the behavior sequence B_1, B_2 the similarity of $S_{simb}(B_1, B_2)$ given by the linear combination of behavior similarity and value similarity, set parameters $z = \frac{\sigma \times L_{len}(s_1)}{max[L_{len}(s_1), L_{len}(s_2)]}$, the calculation formula of $S_{simb}(B_1, B_2)$ is as follows:

$$S_{simb}(B_1, B_2) = (1 - z)S_{simseq}(s_1, s_2) + S_{simvalue}(B_1, B_2) \quad (6)$$

Among them, σ is a constant between 0 and 1, used to adjust z. Because according to practical experience, the importance of value similarity and behavior similarity is related to the ratio of the number of two behavior states. Generally, the closer the number of behavior states is, the greater the component of value similarity.

C. User Behavior Prediction

The user behavior prediction of the proposed method is mainly based on a set of hidden Markov models [22], [23]. First, relevant machine learning methods are used offline to train the prediction model parameters of various behaviors from historical data; Then, according to the partial behaviors of current online users, a finite step Markov process prediction is performed to obtain a relatively complete sequence of user behaviors; Finally, this behavior sequence is used to guide the push of university archives business data.

1) Prediction model of user behavior sequence: The user behavior sequence prediction model of the proposed method is a model systemcomposed of multiple "single two-layer" mixed state hidden Markov models designed on the basis of hidden Markov model theory [24], in which each sub model is independent of each other and deals with the prediction task of different types of user behavior sequences.

The difference between the "single two-layer" mixed state hidden Markov model and the traditional hidden Markov model is that its state set$S$the contained state elements can be divided into two categories: ordinary state elements and double-layer state elements. The double-layer state is mainly introduced to simulate the double-layer condition in the "filter condition". The detailed description of the model is as follows:

a) Status and observations: In the model, the name of a behavior in the user behavior sequence is regarded as a state, and the behavior value is regarded as the output value of the state. Corresponding to the front search function page of the college file business data push system, the type of the "filter condition" tag is considered as the status, and the "filter condition" represented by the "filter condition" tag is the output value of the corresponding status.

b) Transition probability and initial state vector: In order to reduce the complexity of model calculation, when designing the transition probability matrix, the two-layer state is reduced to a common state, and the root state is used as its representative [25], [26]. From any state$i$transfer to a two-layer state$j$transition probability ofa_{ij}. It can be calculated by the following formula:

$$a_{ij} = \sum_{s \in S_j} p_{is} \times S_{simb}(B_1, B_2) \quad (7)$$

Among them, s is j substatus collection for an element in S_j ; p_{is} is status $i$transfer to status$j$sub state of probability of s. Similar to the transition probability matrix, in the initial state vector, the two-level state is reduced to a common state for processing.

c) Probability vector of observation value: For the double-layer state, the observed values are also divided into two layers. For the root state, all its sub states are regarded as its observed values [27], [28]. Therefore, the double-layer state i of k output probability of observation values $b_i(k)$ is:

$$b_i(k) = \sum_{i=1,k=1}^{m} v_{ik}(j) \times a_{ij} \quad (8)$$

Among them, $v_{ik}(j)$ indicates a sub state k of j the probability of output values.

2) Prediction of user behavior sequence: The prediction of user behavior sequence is to start from a given state, predict the subsequent states and the output values of the states in a limited step, and generate a more complete sequence of user behavior [29]. According to the finite stage optimal decision theory of Markov process, the prediction process is to find the starting point from the specified state in the transfer matrix n step state transfer to obtain the maximum utility of the

transfer path. Define the utility that can be obtained in each step of decision-making, and the predicted result of user behavior sequence is:

$$\gamma(i,j) = a_{ij} \times max[b_i(k)] \quad (9)$$

D. Historical Project Recommendation

Historical project recommendation is to push the university file business data related to historical users to current users. The recommended method is top-N. First, the recommended item set is found by combining user collaborative filtering and user behavior sequence search[30]; Then use the relevant sorting strategy to sort the data in the file business data set according to its sorting weight; Finally select N business data is pushed to users.

1) Historical item search: The search of historical items is to use the similarity between the characteristics of the current user and the historical user, and the similarity between the behavior sequence of the historical user and the behavior sequence of the current user as a clue to filter out the item set that may meet the interests of the current user from the historical items. The specific process is divided into the following steps:

a) User behavior prediction: Use the prediction model to predict backwards n step according to the behavior sequence entered by the current user, to get a more complete sequence of user behaviors s. The size of n is related to the number of behavior states entered by the current user, the more states there are, the small the n is. The maximum number of states contained in a user behavior sequence in the system is ϑ, then the calculation formula of n is as follows:

$$n = \begin{cases} \vartheta - \gamma(i,j) & \vartheta > \gamma(i,j) \\ 0 & \vartheta \le \gamma(i,j) \end{cases} \quad (10)$$

b) Finding user behavior template based on user feature similarity: According to the current user behavior sequence s, generate the user's feature vector, calculate the cosine similarity between the historical user and the current user's feature vector, and take out the similarity greater than the threshold σ_1 user behavior template file. σ_1 can be adjusted by the experimental feedback data.

2) Ranking of recommended items: The ranking of recommended items comprehensively considers factors such as feature similarity between users, behavior sequence similarity, popularity of business data and timeliness of business data, calculates the ranking weight of business data, sorts it, and finally generates *N* list of recommended items of file business data. The specific process is divided into the following steps:

a) Calculate the total feature similarity of users associated with data: One file may be related to multiple users, so the overall feature similarity of file business data $simusers(c,S)$ is the user behavior sequence s similarity between the generated characteristics of the current user and those of all neighboring users containing the data $A_{sim}(c,u_i)$ which is calculated as follows:

$$S_{simusers}(c,S) = n \times S_{simusers}(c,S) \times A_{sim}(c,u_i) \quad (11)$$

b) Calculate the total similarity of the behavior sequence associated with the data: One file business may be related to multiple user behavior sequences, so the similarity of one file business behavior sequence $S_{simstrings}(s,Q)$ is the behavior sequence and prediction sequence associated with s the similarity $S_{simb}(s,q_i)$ which is calculated as follows:

$$S_{simstrings}(s,Q) = \sum_{q_i \in S} S_{simb}(s,q_i) \quad (12)$$

c) Calculate the popularity of data: Count the number of different users associated with each file business, and calculate the popularity of each data. The calculation method is as follows:

$$Z(i) = exp\{ m / \sum_{j=1}^{m} \delta(i,j)\}^{-1} \quad (13)$$

where, $Z(i)$ indicates file business data i popularity, m is the total number of related users.

The values of function $\delta(i,j)$ is as follows:

$$\delta(i,j) = \begin{cases} 1 & m_i \in I_j \\ 0 & m_i \notin I_j \end{cases} \quad (14)$$

where, I_j represents a collection of items contained in a user behavior template.

d) Calculation data timeliness: The timeliness of data is determined by the online time and current time of archive business data. The greater the timeliness, the more innovative the data. The calculation method is as follows:

$$T(i) = exp[\,(t_c - t_i)/\phi] \times Z(i) \times \delta(i,j) \quad (15)$$

Among them, $T(i)$ represent data i timeliness; t_c represents the current time; t_i represent data i online time; ϕ represents a constant greater than 0.

e) Calculate the sorting weight of data: Based on the above factors, the ranking weight of a university's archive business data is:

$$W(i) = \xi_1 \times S_{simusers} + \xi_2 \times S_{simstrings} + \xi_3 \times Z(i) + \xi_4 \times T(i) \quad (16)$$

Among them, ξ_i is the influence factor of each factor, which is set through experimental effect feedback or artificial experience.

f)Generate recommendation list: According to the sorting weight of file business data, the data is sorted from the largest to the smallest, and is selected the first *N* to generate recommendation lists to return to the user.

IV. Experiment and Discussion

In order to verify the overall effectiveness of the design method of university archives business data push system based on big data mining technology, it is necessary to test it. Before the test, prepare three virtual machines and modify the corresponding host names to mini0l, min02, and mini03. Next, set their network mode to NAT and modify their IP addresses [31]. Then modify the hosts file of each machine to configure the mapping relationship between the host name and IP address. Close the firewall of each machine and restart the machine. Finally, configure ssh password free login between virtual machines. Now the Linux environment is ready.

Set five datasets, labeled as Dataset 1, Dataset 2, Dataset 3, Dataset 4, and Dataset 5, specifically, it includes the following contents:

Data set 1: Student file data: This data set contains students' personal information, such as name, gender, date of birth, home address, contact information, etc. It also contains academic information, such as enrollment date, major, course results, rewards and punishments, etc.

Data set 2: Staff file data: This data set contains personal information of staff, such as name, gender, date of birth, contact information, etc. It also contains professional information, such as position, employment date, education, work experience, etc.

Data set 3: School business data: This data set contains various business data of the school, such as school curriculum arrangement, examination arrangement, activity arrangement, etc.

Data set 4: User behavior data: This data set contains the behavior data of users (students and faculty) in the system, such as login times, pages visited, time spent on pages, links clicked, etc.

Data set 5: System log data: This data set contains the running logs of the system, such as error logs and operation logs.

Use the proposed method, reference [4] method, reference [5] method, and reference [6] method to conduct business data push test in the Linux environment, and use the accuracy rate $Precision$, recall rate Z_{Recall} , coverage rate C_{ov} as an evaluation indicator of the method. Accuracy can measure the accuracy of pushed data, recall can measure the integrity of pushed data, and coverage can measure the satisfaction of pushed data to users' needs.

Accuracy P_{re} indicates the ratio of the number of test sets contained in the recommended data list to the number of all recommended items. The calculation formula is as follows:

$$P_{re} = \frac{A}{B} \times 100\% \quad (17)$$

where, A indicates the number of correctly predicted samples, B indicates the number of recommended samples for all users obtained from the test set.

It can be seen from the analysis of Fig. 4 that when the data push test is carried out under the same test environment, the data push accuracy of the proposed method is the highest and relatively stable. The data push accuracy of the reference [4] method and reference [5] method fluctuates greatly, and the data push accuracy of reference [6] method is stable but relatively low overall.

Recall rate Z_{Recall} indicates the ratio between the number of user recommendations and the number of items that users have acted in the test set:

$$Z_{Recall} = \frac{A}{C} \times 100\% \quad (18)$$

where, C indicates the number of samples that all users in the test set have had historical behaviors.

According to the test in Fig. 5, compared with the test results of the method in reference [4], the method in reference [5] and the method in reference [6], the recall rate of the proposed method is higher, indicating that the proposed method has good recommendation ability in the field of large-scale business data push.

Coverage rate C_{ov} represents the ratio of the number of recommended items to the total number of items, and its expression is as follows:

$$C_{ov}=\text{B/D}\times 100\% \quad (19)$$

Where, Drepresents the number of samples in the entire dataset.

It can be seen from Fig. 6 that under different data sets, the data push coverage of the proposed method is more than 90%, which indicates that the proposed method includes many types of data pushed by users with full coverage, and the coverage of the reference [4] method, reference [5] method and reference [6] method is low, which indicates that when the above methods are used to launch data push, the pushed data is relatively simple and cannot meet the needs of users.

The time required for the proposed method, reference [4] method, reference [5] method and reference [6] method in the process of data push test is shown in Table I.

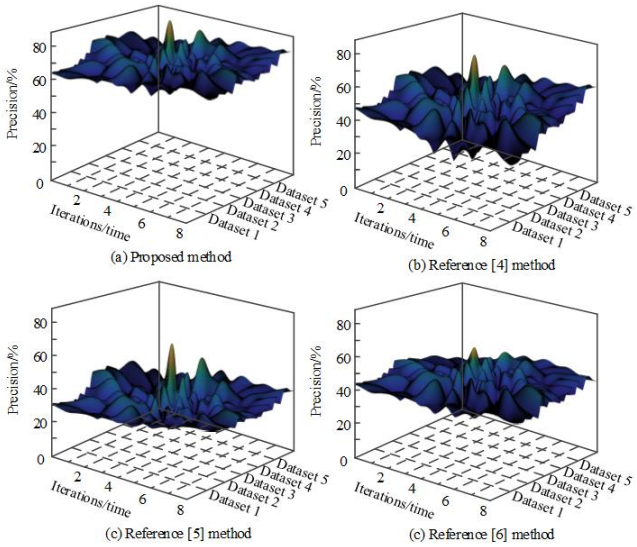


Fig. 4. Accuracy test results.

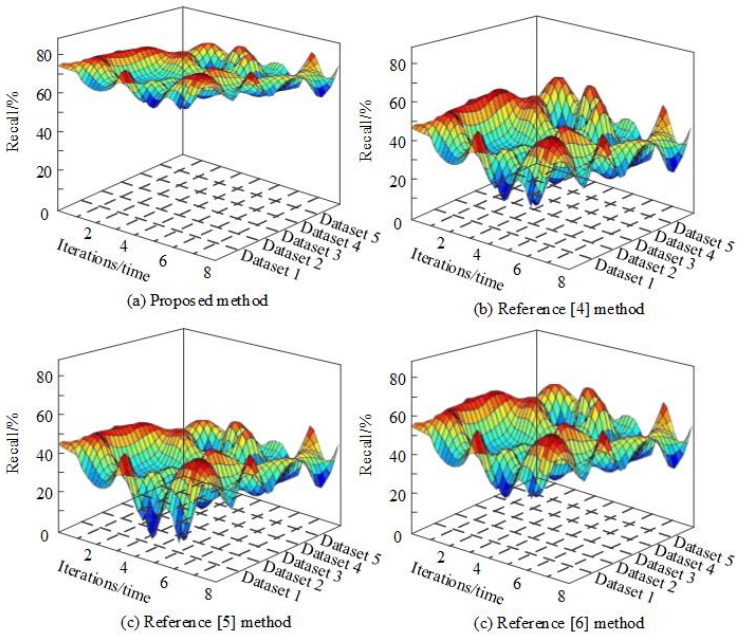


Fig. 5. Recall rate test results.

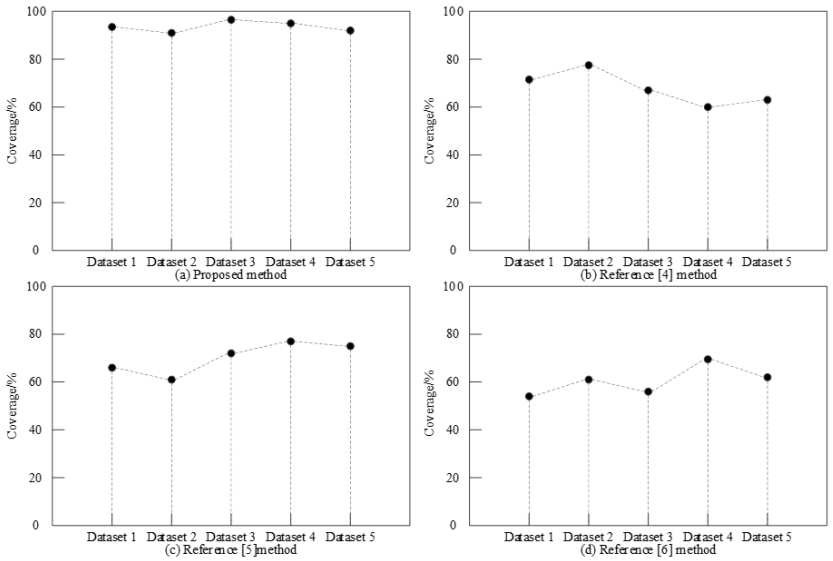


Fig. 6. Coverage test results.

TABLE I. DATA PUSH TIME OF DIFFERENT METHODS

Data samples/piece	Data push time/s			
	Proposed method	***Reference [4] Method***	***Reference [5] Method***	***Reference [6] Method***
100	5.2	9.8	7.6	8.8
200	5.8	10.5	8.3	9.6
300	6.3	11.3	8.9	10.5
400	6.9	12.2	9.7	11.2
500	7.4	13.0	10.4	12.7
600	7.7	14.9	11.7	13.5
700	8.2	15.4	12.6	14.4
800	8.6	16.6	13.7	15.6
900	9.1	17.3	14.9	16.8
1000	9.5	18.1	15.6	17.3

Analysis of the data in Table I shows that the da-ta push time of the proposed method, the reference [4] method, the reference [5] method and the refer-ence [6] method increases with the increase of the number of data samples, but under the same data samples, the push time of the proposed method is far lower than the other three methods, indicating that the proposed method has high data push efficiency.

To sum up, when the data push test is carried out in the same test environment, compared with the methods in [4], [5] and [6], the data push accuracy of the proposed method is the highest and relatively stable; The recall rate is high, the data push coverage rate is above 90%, and the push time required is much lower than the other three methods, which shows that the proposed method has high data push efficiency, and the data pushed for users contains more types and covers all the fields, and it has good push ability in the field of large-scale business data push.

V. CONCLUSION

On the one hand, it takes a lot of time to screen the data, and it is difficult to find the data you need from a large number of data; On the other hand, it will make a lot of redundant information become "hidden information" in the network, which cannot be obtained by ordinary users. In this context, a data push system is proposed. At present, there are some problems in the design method of data push system, such as low push accuracy, low recall, low coverage and low push efficiency. Therefore, a design method of university archives business data push system based on big data mining technology [31] is proposed, which is of great significance and provides decision-making and management support for university management departments. Through the data push function, timely and accurate data transmission and information sharing can be realized, and the efficiency and quality of university archives business can be improved. In this paper, the overall architecture and functional modules of the system are designed, and according to the characteristics of users' behavior, the push of university archives business is realized. It is verified that the data push accuracy of the proposed method is the highest and relatively stable. The recall rate is high, the data push coverage rate is above 90%, and the push time required is much shorter than the other three methods, which can effectively push the required business data for users in a short time. Provide more intelligent and scientific decision-making and management support for university management departments, and improve the efficiency and quality of university archives business.

The prospect of future research work can be carried out from the following aspects:

1) Dig deep into the archives business data of colleges and universities: In the future, we can further expand the data sources, including students, faculty, courses, scientific research projects and other dimensions, in order to obtain more comprehensive data information. Mining hidden rules and trends in data can help university management departments make better decisions and management.

2) Data security and privacy protection: In future research, we need to pay more attention to data security and privacy protection. Explore how to use emerging technologies such as blockchain to ensure the security and credibility of data, and study the methods and technologies of privacy protection to protect sensitive information of individuals and institutions.

COMPETING OF INTERESTS

The authors declare no competing of interests.

AUTHORSHIP CONTRIBUTION STATEMENT

Jun Li: Writing-Original draft preparation

Conceptualization,Supervision, Project administra-tion.

Zhongke Wang: Language review, Methodology, Software.

REFERENCES

[1] J. Yang, H. Chen, and X. Li, “Intelligent products’ recommendation system based on machine learning algorithm combined with visual features extraction,” Int J Biom, vol. 14, no. 2, pp. 125–137, 2022.

[2] B. Peng, “Research and Implementation of Electronic Commerce Intelligent Recommendation System Based on the Fuzzy Rough Set and Improved Cellular Algorithm,” Math Probl Eng, vol. 2021, pp. 1–8, 2021.

[3] C. Liang, R. Fan, W. Lu, and S. Zhao, “Personalized recommendation based on CNN-LFM model,” Computer simulation, vol. 37, no. 03, pp. 399–404, 2020.

[4] H. Pan and Z. Zhang, “Research on context-awareness mobile tourism e-commerce personalized recommendation model,” J Signal Process Syst, vol. 93, pp. 147–154, 2021.

[5] Y. Zheng and D. X. Wang, “A survey of recommender systems with multi-objective optimization,” Neurocomputing, vol. 474, pp. 141–153, 2022.

[6] H. Jelodar et al., “Recommendation system based on semantic scholar mining and topic modeling on conference publications,” Soft comput, vol. 25, pp. 3675–3696, 2021.

[7] S. Panda, S. Mondal, R. Dewri, and A. K. Das, “Towards achieving efficient access control of medical data with both forward and backward secrecy,” Comput Commun, vol. 189, pp. 36–52, 2022.

[8] B. Nour, H. Khelifi, R. Hussain, S. Mastorakis, and H. Moungla, “Access control mechanisms in named data networks: A comprehensive survey,” Acm computing Surveys (cSuR), vol. 54, no. 3, pp. 1–35, 2021.

[9] J. Zhang, T. Zou, and Y. Lai, “Novel method for industrial sewage outfall detection: Water pollution monitoring based on web crawler and remote sensing interpretation techniques,” J Clean Prod, vol. 312, p. 127640, 2021.

[10] S. D. Veeramachaneni, A. K. Pujari, V. Padmanabhan, and V. Kumar, “A hinge-loss based codebook transfer for cross-domain recommendation with non-overlapping data,” Inf Syst, vol. 107, p. 102002, 2022.

[11] S. Soundrya, B. Kumaran, and V. Harini, “An Efficient Two-Layer Framework for Tour Sense Recommendation,” ECS Trans, vol. 107, no. 1, p. 4913, 2022.

[12] A. Jabbari and J. B. Mohasefi, “A secure and LoRaWAN compatible user authentication protocol for critical applications in the IoT environment,” IEEE Trans Industr Inform, vol. 18, no. 1, pp. 56–65, 2021.

[13] C. Hsu, L. Harn, and Z. Xia, “An HSS - based robust and lightweight multiple group authentication for ITS towards 5G,” IET Intelligent Transport Systems, vol. 15, no. 11, pp. 1454‐1460, 2021.

[14] U. Chaisoong, S. Tirakoat, and C. Jareanpon, “Tourist information-seeking behaviours using association rule mining,” ICIC Express Letters, vol. 15, no. 9, pp. 915–923, 2021.

[15] Z. Zhao, Z. Jian, G. S. Gaba, R. Alroobaea, M. Masud, and S. Rubaiee, “An improved association rule mining algorithm for large data,” Journal of Intelligent Systems, vol. 30, no. 1, pp. 750–762, 2021.

[16] Z. Chen, Y. Wang, S. Zhang, H. Zhong, and L. Chen, "Differentially private user-based collaborative filtering recommendation based on k-means clustering," Expert Syst Appl, vol. 168, p. 114366, 2021.

[17] F. Wang, H. Zhu, G. Srivastava, S. Li, M. R. Khosravi, and L. Qi, "Robust collaborative filtering recommendation with user-item-trust records," IEEE Trans Comput Soc Syst, vol. 9, no. 4, pp. 986–996, 2021.

[18] M. Verma and A. Rawal, "An enhanced item-based collaborative filtering approach for book recommender system design," ECS Trans, vol. 107, no. 1, p. 15439, 2022.

[19] E. Türkarslan, J. Ye, M. Ünver, and M. Olgun, "Consistency fuzzy sets and a cosine similarity measure in fuzzy multiset setting and application to medical diagnosis," Math Probl Eng, vol. 2021, pp. 1–9, 2021.

[20] B. Il Kwak, M. L. Han, and H. K. Kim, "Cosine similarity based anomaly detection methodology for the CAN bus," Expert Syst Appl, vol. 166, p. 114066, 2021.

[21] I. Jebli, F.-Z. Belouadha, M. I. Kabbaj, and A. Tilioua, "Prediction of solar energy guided by pearson correlation using machine learning," Energy, vol. 224, p. 120109, 2021.

[22] T. Tamaoka et al., "Denoising electron holograms using the wavelet hidden Markov model for phase retrieval—Applications to the phase-shifting method," AIP Adv, vol. 11, no. 2, 2021.

[23] Y. Aoudni et al., "Cloud security based attack detection using transductive learning integrated with Hidden Markov Model," Pattern Recognit Lett, vol. 157, pp. 16–26, 2022.

[24] C. Raskar and S. Nema, "Metaheuristic enabled modified hidden Markov model for traffic flow prediction," Computer Networks, vol. 206, p. 108780, 2022.

[25] S. Rahimipour, M. Ghatee, S. M. Hashemi, and A. Nickabadi, "A hybrid of neuro-fuzzy inference system and hidden Markov Model for activity-based mobility modeling of cellphone users," Comput Commun, vol. 173, pp. 79–94, 2021.

[26] S. Sefati and N. J. Navimipour, "A qos-aware service composition mechanism in the internet of things using a hidden-markov-model-based optimization algorithm," IEEE Internet Things J, vol. 8, no. 20, pp. 15620–15627, 2021.

[27] Rashid, H. K.,Farkhund, I.,Benjamin, C. M. Fung,J. B. Enabling Secure Trustworthiness Assessment and Privacy Protection in Integrating Data for Trading Person-Specific Information.IEEE Transactions on Engineering Management, vol. 68, no. 1, pp. 149-169, 2021.

[28] Anshu, J.,Vijay, A. Design of Novel Key Generation Technique Based RSA Algorithm for Efficient Data Encryption and Decryption.ECS transactions, vol. 107, no. 1, pp. 2585-2592, 2022.

[29] Vidya, R.,Prema, K.V. DEC-LADE: Dual elliptic curve-based lightweight authentication and data encryption scheme for resource constrained smart devices.IET wireless sensor systems, vol. 11, no. 2, pp. 91-109, 2021.

[30] Abdelhak, E.,Abderrahim, M.,Mohamed, O. Migrating Data Semantics From Relational Database Systems To NoSQL Systems To Improve Data Quality For Big Data Analytics Systems.ECS transactions, vol. 107, no. 1, pp. 19495-19503, 2022.

[31] Gunasekaran, P., Mohamed,S. B.,Ching-Hsien, S. N. K.,Revathi, S.,Priyan, M. K., Bala, A. M. FDM: Fuzzy-Optimized Data Management Technique for Improving Big Data Analytics.IEEE Transactions on Fuzzy Systems: A Publication of the IEEE Neural Networks Council, vol. 29, no. 1, pp. 177-185, 2021.

Augmented Reality SDK Overview for General Application Use

Suzanna[1], Sasmoko[2], Ford Lumban Gaol[3], Tanty Oktavia[4]
Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480[1, 2, 3]
Information Systems Department, Binus Online Learning, Bina Nusantara University, Jakarta, Indonesia, 11480[1]
Information System Management Department, BINUS Graduate Program – Master of Information System Management, Bina Nusantara University, Jakarta, Indonesia 11480[4]

***Abstract*—Augmented Reality Software Development Kits, or as they are commonly called AR SDKs, are useful for developers to build digital objects in AR. This paper presents a comparative study of AR SDKs. This comparison is based on several significant criteria, to select the most suitable SDK. The evaluation used the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) method. Based on a comparative analysis of the features and virtual elements available for application development with the AR SDK, researcher suggests that the main functions of the AR SDK were to be able to offer AR application Editing Platform and facilitate software creation without requiring knowledge of algorithms. Besides that, it is possible to establish some general observations regarding the benefits and limitations of the AR SDK. The result of this research is expected to provide with the clear framework for processing the data that has been collected, summarized, and tested from case study so the researcher will be able to reach useful conclusions. From the literature study has been conducted, it was concluded that among many SDK tools, there are 15 of them which were the most employed by AR developers. These 15 tools were selected based on certain main attributes and support platforms. At the end of this research, it also presents the advantages and limitations of these 15 tools.**

Keywords*—Augmented reality; software development kits; AR SDK; platform; framework; AR technology

I. INTRODUCTION

Augmented Reality Software Development Kits, or AR SDKs, are useful for developers to build digital objects [1]. By utilizing devices like gadgets or tablets, digital objects can be incorporated into the real world through augmented reality (AR). The advantages of the features provided by the AR SDK are the functions for image recognition, 3D object tracking and other multi-tracking, providing simultaneous visual localization and mapping, and many other additional features. This allows developers to create a wide range of digital experiences. AR engineers can leverage this SDK to develop mobile applications, integrate with various CAD platforms, create marketing experiences, develop educational applications, and explore many other possibilities [2]. The AR SDK can generally be used for hardware with certain frameworks [3]. From the results of the literature review, it is also known that several AR SDKs are very flexible to use, so they can be applied to many systems. This makes it easy for AR developers to build AR applications across platforms.

This study presents a comparative study of several AR SDKs. The aim of the comparison that will be carried out is so that AR developers can choose the SDK that best suits their needs so that it can be adapted easily. There are several important criteria that will be explained in the SDK comparison carried out. This research is expected to add value because it has been collected, summarized, and tested from the case study samples that have been carried out so that it can reach useful conclusions for this research.

As we know, AR and virtual reality (VR) have different ways of working. Therefore, AR SDK and VR SDK are also different. The main benefit in the AR SDK category is the functionality of creating AR experiences that can be integrated with various operating systems and various other hardware [1]. Basically, there are three steps in the AR system: recognition, tracking, and blending. At the recognition stage, visual identification occurs in the form of images, whether images of objects, faces, bodies, other body parts, or space will be recognized as virtual objects [4]. The tracking stage is carried out to get real-time visual results in the form of images, objects, faces, bodies, or places such as space so that videos, 3D images, 2D images, text, and others can be added to the AR application [5]. In marker-based AR systems, symbols are used as reference points for overlaying computer graphics. The camera used in this system will continue to search for and identify the target object and then process the image data, starting from the position, orientation, and movement of the visual display that appears. The marking system has a problem, namely that if the lighting is not good, the focus will decrease, and resulting in the image disappearing from the screen, so AR services will not be optimal when using this system [6].

It is different from the AR system without markers [7]. This system uses a combination of electronic devices, such as accelerometers, location data (GPS), compasses, and others, that can determine position in the physical world. With this marker-less AR system, objects can be identified from the direction where the camera is pointing or on which axis the device used operates to obtain the location [6]. This location data determines what the device sees by comparing it to a database, allowing data or computer graphics to appear on the screen. This technological approach sparked the idea of the emergence of 'mobile 'augmented reality' on smartphones, tablets, gadgets, and other mobile technology devices [8].

II. LITERATURE REVIEWS

AR technology is applied in many fields, and AR Software Development Kits ("SDK") have been entered into various applications such as gaming, media and entertainment, automotive, retail, health, education, manufacturing, and others. In the gaming industry, the success of AR technology is that it can offer immersive and interactive experiences that will increase the adoption of AR in the gaming industry. With the success of Pokémon Go, the game company has extensively expanded the technology for games and other applications. For example, the game Ingress Prime adapts the idea of Pokémon Go, where players can go out of the house, travel around, and meet face-to-face with other players. Ingress Prime maximizes the rapid development of AR technology, as seen from the presence of a feature that can make players place a 3D map of a location on the dining table or living room table to set a strategy. Previously, the available portal locations were only numbered in the hundreds of thousands, but now Ingress Prime offers millions of locations spread across the globe [9].

For the media and entertainment industries, the advantage of AR technology is that it can provide real-time experiences. For example, in November 2019, Samsung Electronics launched the Samsung TV True Fit. Using smartphones, both Android and iOS, consumers can see exactly what Samsung's widescreen TV will look like on the wall before buying it. This AR application produced by Samsung can combine futuristic technology with the future of retail shopping and virtual product display. Its main goal is to give consumers the opportunity to choose the right TV at the right size and have a pleasant shopping experience [10].

In the retail industry, AR also offers a promising role. Its ability to provide product and service details tailored to customer needs via smartphones with 3D effects has an impact on market growth. For example, Walmart has used AR so that its customers get shopping experience by scouring supermarket aisles for Waffles + Mochi characters to unlock AR content. After scanning a QR code with a smartphone, image recognition technology allows consumers to search for nine different characters hiding in grocery aisles. As shoppers discover each character in the store, they will earn badges and gain access to unlock more games, recipes, stories, and clips from the Walmart app [11].

In the healthcare sector, there is also an increasing demand for AR because it helps with simulations for surgery, training, and patient care. For example, the Mayo Clinic is the best hospital in the world to develop treatments using AR technology. The past decade has seen tremendous growth and expansion in innovative efforts to address the unmet needs of patients, providers, and care systems. To facilitate this effort, the Mayo Clinic Cardiovascular Medicine established a virtual reality innovation group at the Mayo Clinic in Rochester, Minnesota [12].

In the field of education, the results of a study entitled "Investigating Student Attitudes toward Augmented Reality" found that students have positive attitudes towards AR applications, as seen from the increased interest and motivation in an active and interactive learning environment through AR [13]. This proves that AR technology is able to make applications more interesting to use and increase user interest in learning something [14]. In manufacturing, AR is applied to product design and development, quality control (QC) processes, logistics, employee training, equipment maintenance, and more. With the adoption of AR applications, manufacturers can increase productivity, lower costs, and work faster. For example, AR technology and the Internet of Things (IoT) will allow manufacturing systems to self-assess deficiencies or errors that occur in the system [24].

According to a market analysis perspective by International Data Corporation (IDC), the latest projections estimate the AR market will reach $60.55 billion by 2023 [15]. AR is not just an IT technology; it is a link between the digital and physical worlds and becomes a new interface between humans and machines. The world's leading IT companies, such as Amazon, Facebook, Mayo Clinic, US Navy, and others, have implemented AR, and they are seeing the huge impact AR has on quality and productivity for their companies. The main advantage that AR offers to various fields that use it is the power of AR in processing information so that humans can experience unforgettable immersive experiences. With this immersive power, humans can use their five senses to access information at different speeds. From the research results, it is known that vision provides 90% of the information humans obtain through sight. However, accessing information is not easy because it requires mental capacity that is able to absorb and process information [17]. Therefore, cognitive load, as a demand on human capacity, requires mental effort. Like reading instructions from a computer screen, carrying out calculations, and thinking about the information obtained, all of this provides a greater cognitive load compared to listening to the same instructions because the letters must be translated into words. Words must then be represented so that there is a distance or gap between the presentation of information and its understanding in the context of applying the information [18]. For example, when a driver refers to a smartphone and looks for directions while driving, this requires its own focus. The steps that must be taken are that the driver must read the information from the screen, then memorize the information in his memory, look at the screen, and translate the directions from the screen to the physical environment in front of him. Then follow the instructions when operating the vehicle. There is a lot to do, and there is a cognitive gap between the digital information on the screen and the physical reality on the ground where the information must be applied. The combination of the speed of conveying and absorbing information and the cognitive distance involved in its application is like the term "a picture is worth a thousand words"[16]. When humans view the physical world, they absorb vast amounts of information. The information received is almost immediately absorbed. In the same way, an object or image overlaid with information about the physical world places it in context for human knowledge. This can reduce cognitive distance and minimize cognitive load. This reason can explain why AR can become a trend and is in an important position. There is no better graphical user interface than the physical world we see around us when enhanced with a digital overlay of relevant data and guidance on where and when it is needed. AR eliminates reliance on 2D

information that is out of context and difficult to process on pages and screens, while enhancing humans' ability to understand and apply information in the real world [19].

III. Research Methods

The papers used in this study were collected from a bibliographic database of Scopus and academic publications, such as ERICS, IEEE, Science Direct, Taylor and Francis Journal, and Google Scholar. The required articles were those published between January 2019 and October 2023. The following search keywords resulted in 15,900 papers being collected:

"Augmented" AND "Reality" AND "SDK"

"AR" AND "SDK"AND "Technology"

"AR" AND "Framework" AND "Development"

After selection, there were 235 suitable articles.

The evaluation used the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) method with five stages used to conduct a literature review, namely defining eligibility criteria, defining information sources, selecting literature, collecting data, and selecting data items as shown in Fig. 1. Meanwhile, for the presentation of the results of the literature review that has been carried out, it can be seen in Table I below:

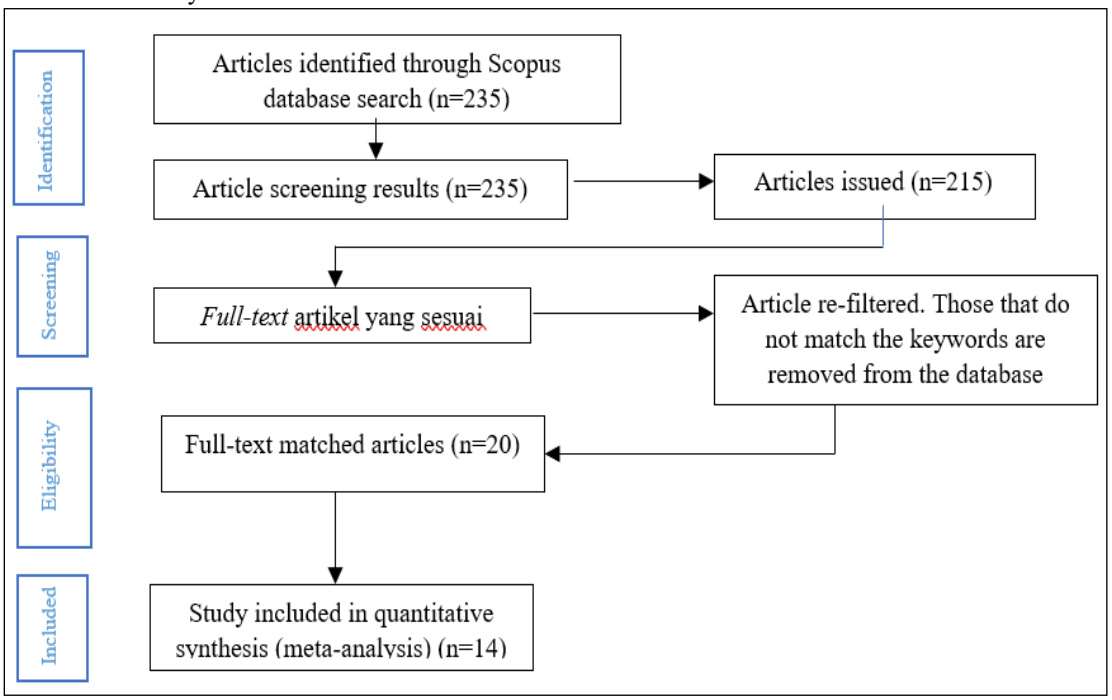


Fig. 1. PRISMA diagram (adapted from Moher et al (2009)).

IV. Analysis of Frameworks and Platforms

From the results of a literature study on AR development tools to explore the functionality and applications used by AR, 15 tools were found that can be used to perform analysis and testing. The data obtained for the development of AR resources is shown in the table below:

1) ARToolkit: Currently for developing AR applications. ARToolkit has the most libraries and the most users. The main advantage of ARToolkit is that it is open source, offers other supporting features, and allows multi-platform use. Its open-source nature allows ARToolkit to continue to develop, and its advantages in tracking planar targets, geolocation, and several other targets can develop along with the needs of AR developers [20][21].

2) Realitykit: The main advantage of Realitykit is rendering realistic photos, creating animations, physics, providing camera effects, and more that are created specifically for AR frameworks. Realitykit makes AR development easier than other tools due to its Native Swift API, ARKit integration, spatial audio, highly realistic physics-based rendering, animation frameworks and transformations, and spatial audio and physics rigid objects, all of which support AR development [2].

3) Aurasma: Aurasma offers an AR platform for education and personal use for free. With support for creating its own mobile applications and web-based platforms, it allows AR developers to transform objects, images, and places. This advantage provides new interactive opportunities using graphic, animation, video, audio, text, and 3D content, enabling planar target tracking and geolocation, as well as tracking via cloud storage in the form of 3D AR [1].

4) BlippAR: The main advantage of BlippAR is that users can register their own bookmarks and connect with many dynamic and interactive visual assets for smartphones. BlippAR supports deep learning algorithms with artificial intelligence, allows users to use cloud technology to track

planar targets, and supports learning of various things that AR developers may need [1][2].

5) CraftAR: The main advantage of CraftAR is that it has a multi-platform SDK, offers mobile applications and a web-based platform, tracks planar images through the cloud, and registers their own bookmarks in various forms such as 3D models, images, audio, video, and more [3].

6) EasyAR: The main advantage of EasyAR is that it has simple and efficient features, is easy to use, and has advanced functions that AR developers have long awaited, such as dynamic target recognition loading, screen recording functions, hard decoding, and more than 1000 types of local target recognition. Apart from that, the web-based platform is also an advantage because it can make it easy for users to register their projects and obtain the necessary licenses. This license is useful for testing and releasing user applications and offers some planar target tracking functionality [1].

7) Kudan: This framework offers an SDK that can be exported to various platforms. Additionally, Kudan has extensive documentation and practical examples and has a support forum for fellow communities. SLAM is also available, allowing the tracking of fewer markers [22] [23].

8) LayAR: Developers can use many features on LayAR, such as adding buttons for social media, web pages, phone calls, email, downloading, voting, shopping, and adding contacts. The programming language used is Java, with back-end XML, PHP, MySQL, and others that support the JSON format. For use on iOS, the iPhone is LayAR's mainstay for integration with specific SDK hardware requirements [2].

9) PixLive: Using PixLive allows the use of media such as images, 3D models, 360-degree images, audio, and video. Developers can also add buttons to social media, web pages, and PDF files and use images and text as buttons. To create scenes and buttons or timers to navigate between scenes, PixLive allows the use of resources for drawing applications and the use of geolocated resources [3].

10)Vuforia: Vuforia is one of the most popular platforms for developing AR. Vuforia provides a web-based environment where users can create and manage their bookmarks and obtain the necessary licenses. Licenses are required to test and publish their applications. Vuforia can also be used to track planar targets, geolocation, multiple targets, text, and 3D objects. Cloud technology can also be used on Vuforia to store data locally on the user's device. Vuforia even enables marker less tracking through two technologies, namely extended tracking and smart terrain.

11)Wikitude: The main advantage of Wikitude is the ability to track planar markers, 3D objects, geolocation markers, multiple targets, and the use of marker less SLAM technology. This platform is paid and offers a web-based 3D AR display and management platform to mobile users. Bookmarks can be created by users and linked to 3D models and other virtual elements.

12)Metaio: A library for creating AR mobile applications. Metaio can perform pattern tracking which will then be compared with existing references. In Metaio, the AR application design has embedded four marker tracking methods to carry out the marker recognition process. These methods have a threshold in the process of recognizing markers[23].

13)D'Fusion: A development platform for building Augmented Reality applications that provides tools for creating all kinds of immersive augmented reality experiences. The platform can also manipulate 3D visuals, combine marker less and gesture recognition, and easily set up and deploy complex AR. More than 300 parameters in advanced functionality are available in this graphical configuration for object recognition and beyond [2].

14)ARMedia: A platform that can help develop Augmented Reality applications effectively and efficiently. This SDK can be used to track unique 3D models from simple AR projects to very complex ones. The SDK provides recognition tools and includes 3D Object, Planar, Location, and Motion Tracking. ARmedia is not only able to recognize planar images and locations, but also 3D objects regardless of size and geometry. The ARMia SDK supports building advanced applications and systems that serve across a wide range of application domains. ARmedia SDK with any 3D Engine and Writing Environment like Unity 3D and Open Scene Graph[1].

15)ARCore: ARCore is a platform for building AR that focuses on mobile devices. The advantage of ARCore is that it can use different APIs; the user's device can observe and receive information about its environment and interact with that information [20][1].

Apart from these 15 tools, there are other tools such as ARUco, ZapWorks, Augment, HP Reveal, PTAM, HandyAR, ARGON, Amazon Sumerian, Mixare and others. It's just that based on the many tools, these 15 tools are the most widely used in research that has been done previously. For tracking target the main attributes in this classification follow a set of criteria that will be used to compare in each SDK. The analysed attributes are:

1) Text recognition: It is used to recognize words and/or groups of words for analysis.

2) Image recognition: It is used to benchmark its planar image recognition.

3) 3D object recognition: It is used to track cylindrical, conical, and general 3D real objects as investigated landmarks.

4) Multi-target recognition: It is used to track multiple targets simultaneously for analysis.

5) Geolocation recognition: It is used to test and verify which tools support geolocation targets.

6) Marker-free recognition: It is used to identify marker-free AR platforms.

From the results of previous research, it is known that there are several support platforms that support the AR SDK including Windows, Linux, OSX, Android, IOS, MAC OS or OSX, Smart Glasses and the Web. From the results of the study conducted by the researcher, the data shown in Table I below was taken:

TABLE I. ANALYSIS OF FRAMEWORKS AND PLATFORMS

Tools	Extension/ Platform	Text Recognition	Image Recognition	3D Object	Multi Targets	Geo-location	Marker less
ARToolKit	Windows, Mac OS, Linux, iOS, Android, Unity Package, Smart Glasses	x	✓	x	✓	✓	x
RealityKit	App (Android, iOS, Windows, Mac)	x	✓	x	x	x	x
Auras-ma	App (Android, iOS)	x	✓	x	x	✓	x
Blipp-AR	App (Android, iOS)	x	✓	x	x	x	x
CraftAR	App (Android, iOS), Unity Package, Apache Cordova	x	✓	x	x	x	x
EasyAR	Windows, Mac OS, Android, Unity Package	x	✓	x	✓	x	x
Kudan	Android, iOS, Unity Package	x	✓	x	✓	x	x
LayAR	App (Android, iOS e BlackBerry)	x	✓	x	x	x	x
PixLive	App (Android e iOS), Apache Cordova e Google Glass	x	✓	x	x	✓	x
Vuforia	Android, IOS	✓	✓	✓	✓	✓	✓
Wiki-tude	App (Android e iOS), Unity Package, Cordova, Titanium, Xamarin e Smart Glasses.	x	✓	x	✓	✓	✓
Metaio	Android, iOS, Unity Package	x	✓	✓	✓	✓	✓
D'Fu-sion	Windows, iOS, Android, Smart Glasses, Unity Package	x	✓	✓	✓	✓	✓
AR-Media	Android, iOS, Unity Package	x	✓	✓	✓	✓	✓
ARCore	Android, iOS, Unity Package	x	✓	✓	✓	✓	✓

In general, AR SDK has similar features, but if we look deeper, each AR SDK has main differences and makes each SDK have specializations that differentiate it from other SDKs. For example, there are some SDKs that have advantages in tracking, but other SDKs excel in recognizing objects. The main feature advantage of all the SDKs studied in this research is their tracking system. Tracking can be done in various ways and each solution offered has its own advantages and disadvantages in the AR experience to be created. There are four main geographic characteristics as seen in Fig. 2 below.

The explanation of Fig. 2 above is as follows:

1) Plane tracking is the localization of one or both ny of the translational offset (X, Y, Z) with rotational orientation (roll, pitch, yaw) in an object with respect to the origin. In AR, the point of origin is usually a mobile device with an embedded camera and almost all AR SDKs have plane tracking although individual results vary.

2) SLAM typically uses Apple ARKIT and Google ARCore as AR SDKs that perform marker less tracking.

3) A common example of face tracking technology is facing tracking with Snapchat AR lenses which uses algorithms to accurately track a user's face. Apart from tracking faces, this technology can also add digital makeup such as eye shadow or lipstick. For game creation, players can become avatars. Face tracking AR SDKs include Vuforia, DeepAR, and Aurasma.

4) Object recognition allows the AR system to identify objects in the real world. This process is still in development but can already display digital information about objects into the field of view and penetrate various areas of life such as education, manufacturing, construction, and design as well as health care.

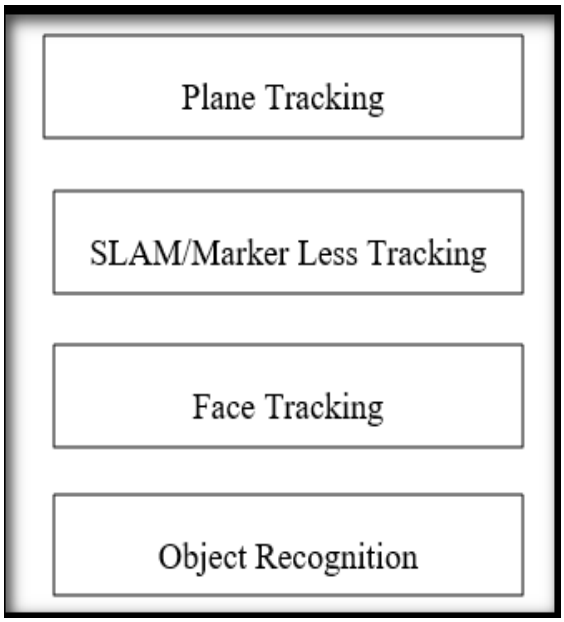


Fig. 2. AR SDKs and Tracking.

V. Result Discussion

Based on a comparative analysis of the features and virtual elements available for application development with the AR SDK, researcher suggests that the main functions of the AR SDK where to be able to offer AR application Editing Platform and facilitate software creation without requiring knowledge of algorithms. Besides that, it is possible to establish some general observations regarding the benefits and limitations of the AR SDK in this study as shown in Table II.

TABLE II. Analysis of Advantages and Limitations of Augmented Reality SDK

	Advantages	Limitations
ARToolKit	- These tools can create AR applications for iPhone and iPad. - ARKit helps developers develop AR applications that can support two devices to share the same virtual item thereby making the AR experience more engaging.	- In less accurate tracking even when the camera and markers are stationary (It does not support location-based). - Free access to AR Library but its development documentation is quite limited.
RealityKit	- High-performance 3D simulation and rendering experiences.	- No Concave Models - No Transparency or Opacity - No Videos - No Shaders
Aurasma	- Allows the association of touching actions on the screen, to start and end the app and after a determined amount of time. - Doesn't required the knowledge to implement algorithms.	- The lack of a control group with which to compare the participants fidelity of videos.
BlippAR	- No coding skills are required. This tool is free for educational use for the next 3 months. - Allows detection of markers and entities. - Allows for a reasonable amount of graphical rendering and animation. - Supports 3D markers for standard shapes.	- Not support marker less AR, though this will be supported soon. - Not support real-time shadows for rendering at present; this will need to be 'faked' by the content creator
CraftAR	- This SDK is most used by iOS apps in the Magazines & Newspapers genre, followed by the Book genre. - Doesn't required the knowledge to implement algorithms.	- Expensive and must be updated to recognize more targets or produce different behaviour.
EasyAR	- EasyAR is an SDK that lives up to its name. The features are simple, easy to use, and efficient. The functions of this SDK are quite advanced according to the needs of AR developers, including hard decoding, dynamic target recognition loading, screen recording functions, and a superior function, namely the recognition of more than 1000 local targets.	- Not free, pricing and payment details are listed on the EasyAR SDK product page. A free trial for EasyAR SDK Pro is provided. Each application will be given a limited time in the trial period.
Kudan	- Kudan is faster than other frameworks. This tool helps mobile AR applications to map multi-polygon models and import 3D models from any of the modelling software packages. - For the number of image recognition is not limited and requires less memory to store files on the device.	- This framework manual is brief and requires additional information. - Limited built-in functionality without direct access to OpenGL.
LayAR	- Offer interactive interactions to users in the context of a brand. Instead of instructing users to download the Screen App, users can bring LayAR interactivity directly into the app without needing to code everything themselves.	- SDK screen is NOT free. A 30-day Trial Period to play with the SDK before proceeding with purchases. But during this trial period users can use the SDK without any restrictions.
PixLive	- Multi-directional recognition and faster synchronization. - Doesn't required the knowledge to implement algorithms.	- Does not provide marker less and offline recognition.
Vuforia	- Enable to maintain tracking even when the target is out of view and view them from greater distance. - Cloud database allows storing thousands of image targets.	- The downside of the Vuforia SDK for Android is that the database is limited; it can only support 100 target images and does not feature any utility functions for loading 3D models, making it difficult for AR developers to create image formats.
Wikitude	- AR Content can be programmed using basic HTML5, JavaScript and CSS. - Enables the use of media such as images, video and 3D models. Allows the insertion and edition of text and buttons for social media. - Doesn't required the knowledge to implement algorithms.	- The SDK component could be a bit heavy, in terms of size. - Tracking of object sometimes was getting reset.
Metaio	- Powerful 3D rendering engine with capability load 3D model of .obj format. - No limit on number of trackable objects depends on device memory.	- Rendering 3D objects is very difficult using this SDK, and it is not easy to carry out developments related to the size of the model to be created.
D'Fusion	- To create high-quality 3D content, this SDK makes it easy to create several different 3D object formats. - Provide encrypted media to prevent privacy or tampering risks. - More than 300 parameters in advanced functionality are available to meet your needs with this graphical configuration tool for object recognition and tracking.	- Video file supported but audio associated with video can't be played.
Armedia	- Depth camera calibration provided which created more immersive experience	- Doesn't support all type of textures for 3D objects.
ARCore	- Its advantages are in motion capture, environmental perception, and light source perception.	- Sometimes it is difficult to scan big virtual object.

VI. CONCLUSION

From the results of the analysis and comparison of features in the AR SDK and virtual elements that have been carried out, the researchers suggest that the main function of the AR SDK is that it can be a platform for editing AR applications and makes it easy to create AR applications without needing to know algorithms or have knowledge of algorithms. The main attributes and supporting platforms were also examined in this study so that they could provide developers especially for the beginners with AR application knowledge. The results of this study showed that the main functions of the AR SDK were to be able to 1) offer an AR application Editing Platform and 2) facilitate software creation without requiring knowledge of algorithms. Developers and especially AR beginners do not need to write any code for the algorithm. These two functions could be in Aurasma, BlippAR, CraftAR, LayAR, PixLive, and Wikitude.

VII. SUGGESTION

Regarding the future perspective, this paper suggests conducting more investigations of the framework, AR functionality and features which required in AR development.

REFERENCES

[1] A. Hanafi, L. Elaachak, and M. Bouhorma, "A comparative study of augmented reality SDKs to develop an educational application in chemical field," ACM Int. Conf. Proceeding Ser., vol. Part F148154, 2019, doi: 10.1145/3320326.3320386.

[2] D. Amin and S. Govilkar, "Comparative Study of Augmented Reality Sdk's," Int. J. Comput. Sci. Appl., vol. 5, no. 1, pp. 11–26, 2015, doi: 10.5121/ijcsa.2015.5102.

[3] J. C. S. Cardoso and A. Belo, "Evaluation of multi-platform mobile ar frameworks for roman mosaic augmentation," GCH 2018 - Eurographics Work. Graph. Cult. Herit., pp. 119–128, 2018, doi: 10.2312/gch.20181348.

[4] J. C. G. Vargas, R. Fabregat, A. Carrillo-Ramos, and T. Jové, "Survey: Using augmented reality to improve learning motivation in cultural heritage studies," Appl. Sci., vol. 10, no. 3, pp. 1–27, 2020, doi: 10.3390/app10030897.

[5] F. De Pace, F. Manuri, and A. Sanna, "Augmented Reality in Industry 4.0," Am. J. Comput. Sci. Inf. Technol., vol. 06, no. 01, 2018, doi: 10.21767/2349-3917.100017.

[6] S. Mystakidis, A. Christopoulos, and N. Pellas, "A systematic mapping review of augmented reality applications to support STEM learning in higher education," Educ. Inf. Technol., no. August, 2021, doi: 10.1007/s10639-021-10682-1.

[7] M. Kochi, "Book Object for Augmented Reality," Comput. Animat. Virtual Worlds, vol. 22, no. 6, pp. 529–541, 2016, doi: 10.1002/cav.431.

[8] C. Llena, S. Folguera, L. Forner, and F. J. Rodríguez-Lozano, "Implementation of augmented reality in operative dentistry learning," Eur. J. Dent. Educ., vol. 22, no. 1, pp. e122–e130, 2018, doi: 10.1111/eje.12269.

[9] "Niantic Luncurkan Game AR Baru, Ingress Prime | Dailysocial." https://dailysocial.id/post/niantic-ingress-prime (accessed Jan. 14, 2022).

[10] "See Your TV in a Way You Never Imagined With the Samsung TV True Fit Mobile Application – Samsung Newsroom Canada." https://news.samsung.com/ca/see-your-tv-in-a-way-you-never-imagined-with-the-samsung-tv-true-fit-mobile-application (accessed Jan. 08, 2022).

[11] "Walmart's In-Store AR Experience Has Shoppers On the Hunt for Waffles + Mochi | RIS News." https://risnews.com/walmarts-store-ar-experience-has-shoppers-hunt-waffles-mochi (accessed Jan. 09, 2022).

[12] "Developing augmented and virtual reality treatments - Mayo Clinic." https://www.mayoclinic.org/medical-professionals/cardiovascular-diseases/news/developing-augmented-and-virtual-reality-treatments/mac-20462595 (accessed Jan. 09, 2022).

[13] M. Sirakaya and E. Kiliç Çakmak, "Investigating Student Attitudes towards Augmented Reality," Malaysia Online J. Educ. Techology, vol. 6, no. 1, pp. 30–44, 2018.

[14] M. Afif and M. Yusof, "Preserving Malaysian Folktales Through Mobile AR Application : Sumpahan Ikan Tapah," J. Comput. Technol. Creat. Content, vol. 6, no. June, pp. 6–9, 2021.

[15] "Market Analysis Perspective: Worldwide Augmented Reality and Virtual Reality, 2021." https://www.idc.com/getdoc.jsp?containerId=US47266821 (accessed Jan. 08, 2022).

[16] K. Kuswara and Y. Sumayana, "Apresiasi Cerita Rakyat sebagai Upaya Memperkuat Karakter Siswa dalam Menghadapi Revolusi Industri 4.0," J. Basicedu, vol. 5, no. 1, pp. 317–326, 2020, doi: 10.31004/basicedu.v5i1.678.

[17] K. A. Butler, B. E. John, and R. J. K. Jacob, "Human-Computer Interaction : Introduction and Overview," no. January 1997, pp. 1997–1999, 2016, doi: 10.1145/286498.286556.

[18] Z. Bhatti, A. Abro, and M. Karbasi, "Be-Educated: Multimedia Learning through 3D Animation," no. February, 2018.

[19] Y.-H. Wang, "Exploring the effectiveness of integrating augmented reality-based materials to support writing activities," Comput. Vol. 113, issue 1, vol. 113, no. 1, 2017.

[20] F. Pankratz, "Penerapan Teknologi Augmented Reality Pada Katalog Rumah Berbasis Android," vol. 90, no. 1, pp. 88–92, 2016, [Online]. Available: https://books.google.co.uk/books/about/Augmented_Reality_for_Developers.html?id=8xhKDwAAQBAJ&printsec=frontcover&source=kp_readbutton&redir_esc=y#v=onepage&q&f=false%0Ahttps://mediatum.ub.tum.de/1286695

[21] A. Nur Amalina, "Desain Sistem Augmented Reality dengan Marker di dalam Film Pendek Beranimasi yang Mengadopsi Cerita Rakyat Sangkuriang," in Universitas Indonesia Library, 2013, pp. 57–60. Accessed: May 17, 2021. [Online]. Available: http://lib.ui.ac.id

[22] A. Yulianti, B. P. Andika, and A. Labellapansa, "Application of Batu Belah Batu Bertangkup industri in Riau Province with Augmented Reality," in ICSEC 2019 - 23rd International Computer Science and Engineering Conference, Oct. 2019, pp. 60–64. doi: 10.1109/ICSEC47112.2019.8974761.

[23] T. A. Vakaliuk and S. I. Pochtoviuk, "Analysis of tools for the development of augmented reality technologies," CEUR Workshop Proc., vol. 2898, pp. 119–130, 2021.

[24] A. Palancı and Z. Turan, "How Does the Use of the Augmented Reality Technology in Mathematics Education Affect Learning Processes?: A Systematic Review," Int. J. Curric. Instr. Stud., vol. 11, no. 1, pp. 89–110, 2021, doi: 10.31704/ijocis.2021.005.

Automatic Extractive Summarization using GAN Boosted by DistilBERT Word Embedding and Transductive Learning

Dongliang Li*, Youyou Li, Zhigang ZHANG
College of Artificial Intelligence, Jiaozuo University, Jiaozuo, 454000, China

***Abstract*—Text summarization is crucial in diverse fields such as engineering and healthcare, greatly enhancing time and cost efficiency. This study introduces an innovative extractive text summarization approach utilizing a Generative Adversarial Network (GAN), Transductive Long Short-Term Memory (TLSTM), and DistilBERT word embedding. DistilBERT, a streamlined BERT variant, offers significant size reduction (approximately 40%), while maintaining 97% of language comprehension capabilities and achieving a 60% speed increase. These benefits are realized through knowledge distillation during pre-training. Our methodology uses GANs, consisting of the generator and discriminator networks, built primarily using TLSTM - an expert at decoding temporal nuances in timeseries prediction. For more effective model fitting, transductive learning is employed, assigning higher weights to samples nearer to the test point. The generator evaluates the probability of each sentence for inclusion in the summary, and the discriminator critically examines the generated summary. This reciprocal relationship fosters a dynamic iterative process, generating top-tier summaries. To train the discriminator efficiently, a unique loss function is proposed, incorporating multiple factors such as the generator's output, actual document summaries, and artificially created summaries. This strategy motivates the generator to experiment with diverse sentence combinations, generating summaries that meet high-quality and coherence standards. Our model's effectiveness was tested on the widely accepted CNN/Daily Mail dataset, a benchmark for summarization tasks. According to the ROUGE metric, our experiments demonstrate that our model outperforms existing models in terms of summarization quality and efficiency.**

***Keywords*—Extractive text summarization; generative adversarial network; transductive learning; long short-term memory; DistilBERT**

I. Introduction

In the digital era, there is an overwhelming amount of online information. Manually extracting insights from this vast data is challenging. Automatic Text Summarization (ATS) is a solution that extracts essential details efficiently. Summarization involves creating a concise version of text from one or multiple sources, capturing the main information for specific users or purposes [1].

There is a plethora of approaches for extractive summarization. Some lean on machine learning techniques such as support vector machines [2] and clustering [3], optimization algorithms [4-7], while others adopt graph-based strategies [8], where sentences are portrayed as graphs, the nodes represent words, and edges signify the relationship between those words. As deep learning evolves, it is becoming more dominant in natural language processing, overshadowing conventional machine learning techniques. Thanks to its complex architecture, deep learning autonomously discerns word, sentence, or document attributes. However, even with numerous deep learning-driven summarization techniques, many grapple with extractive summarization intricacies. Crucial aspects of summarization, like sentence evaluation and choice, often present hurdles. Many existing methods tend to be overly selective, meaning after picking a valuable sentence, they might overlook its relevance in subsequent selections, reducing the overall efficacy of the summary.

Generative Adversarial Networks (GANs) are advanced machine learning tools that consist of a generator and a discriminator. The generator strives to create lifelike outcomes, such as images or text. In contrast, the discriminator tries to discern between genuine and fabricated content [9]. GANs hold potential for optimizing sentence selection in extractive text summarization. When generating a summary, they evaluate the entirety of the document, giving the generator a holistic grasp. Such understanding helps the generator pick subsequent sentences more judiciously, recalling previously identified important sentences and ensuring better summary quality. Adversarial Training significantly boosts the GANs' capability to address the issue of biased sentence selection. The discriminator offers critical insights into the generator about the cohesiveness and quality of the formed summary, allowing the generator to refine its methods. This results in selecting impactful sentences that also harmonize with earlier pivotal sentences. Thanks to adversarial training, GANs adeptly address the risk of omitting key sentences, delivering more unified summaries.

LSTM has been fundamental in numerous sequence-oriented tasks, such as video categorization, machine interpretation, and text summarization [10, 11]. However, standard LSTMs, rooted in inductive learning, shape a universal model from all training datasets. This can sometimes neglect nuances within the data, potentially hampering the adaptability of the model. TLSTM introduces a solution by incorporating a transductive learning method. This approach emphasizes performance improvement around novel data points, accentuating localized nuances. In the structure of TLSTM, this is accomplished through a tailored weighting system, adjusting weights concerning their closeness to the assessment data. Closest data points to the test get a higher

weightage, ensuring optimal performance in those areas. By combining the strengths of LSTM with the adaptability of transductive learning, TLSTM offers a more tailored time-series prediction method. It captures long-term dependencies while addressing overlooked data nuances, making it suitable for intricate time-series challenges [12].

The BERT model in [13] is a notable NLP tool with many parameters. Larger models, while effective, increase computational and environmental costs. DistilBERT [14], a streamlined transformer model, is a distilled version of BERT. It functions 60% swifter and requires 40% less parameters compared to BERT, as evidenced in the GLUE benchmark. In comparison with predecessors like BERT and RoBERTa [15], DistilBERT is a more streamlined variant.

The paper presents an extractive summarizer, founded on DistilBERT word embedding, GAN, and attention mechanism-based TLSTM. GANs are made up of two generator and discriminator components that compete in a process. In this context, the generator's goal is to rate each sentence of the document, while the goal of the discriminator is to distinguish the real from the fake summary, which enhances the performance of the generator. In a non-greedy way, the generator determines the possibility of the presence of sentences in summary at once. The contributions of this article are as follows:

- Using GAN for summarization, the generator improves based on feedback from the discriminator, incorporating both real and fake summaries.
- We use TLSTM to design the generator and discriminator, enhancing accuracy in text summarization.
- By introducing varied noise levels during training and testing, we produce diverse summaries, with a voting system determining the final summary.
- Our proposed model utilizes DistilBERT word embedding to automatically learn and extract complex and meaningful text representations from the input data.

The remainder of the paper is structured as following. Section II covers some related works, while Section III introduces our proposed text summarization method. Section IV presents experimental results, and Section V concludes the paper.

II. Related Works

Abstractive summarization methods, a notable strategy in NLP, aim to produce summaries that don't merely pick and reorganize existing sentences or phrases [16]. These techniques endeavor to grasp the essence of the text and formulate new, succinct, and cohesive statements that reflect the main ideas of the original content [17]. Abstractive summarization seeks to produce summaries with a human-like touch, capturing the heart of the source material without restricting itself to direct extractions. By discerning the core semantics, connections, and subtleties of the document, abstractive methods can potentially craft summaries that are richer, more concise, and linguistically smooth. To realize this, such techniques frequently utilize advanced tools like neural networks and natural language generation models [18, 19]. These models employ methods such as sequence-to-sequence frameworks, attention systems, and reinforcement learning to craft summaries that hold semantic significance and flow smoothly. By grasping the underlying context and essence of the text, abstractive summarization models can reword and restructure the original material, introducing fresh phrases, reshaping statements, and even creating unique expressions to highlight the primary details. This capability to transcend basic extraction allows abstractive summarization to deliver shorter summaries that still encapsulate the primary intent of the original text. Yet, this approach comes with its set of challenges. The crafted summaries must walk the fine line of being brief yet informative, ensuring logical flow and upholding the truthfulness of the source. Moreover, abstractive techniques often demand vast training data and intricate models to effectively decipher the subtleties and variances in natural language [20].

Numerous extractive summarization methods, spanning graph-based to deep learning techniques, have been explored [21, 22]. LeClair et al. [23] delved into code summarization advancements via Graph Neural Network (GNN) application, enhancing summary insightfulness. Zhong et al. [24] derived word features from documents and then determined sentence scores based on word scores. Yousefi et al. [25] scored sentences using the cosine similarity between them and their topics. Cao et al. [26] employed recurrent neural networks for sentence ranking, treating sentences as trees with words as leaves, and deriving sentence scores from a non-linear procedure. Rosca et al. [27] utilized reinforcement learning for summary creation, where sentence coherence was the reward. The policy was crafted as a multilayer perceptron assigning scores to sentences. Abdi et al. [28] showcased a deep learning methodology for creating opinion-focused multi-document summaries. This involved components like sentiment analysis embedding space (SAS), text summarization embedding spaces (TSS), and an opinion summarizer module (OSM) [29]. The SAS uses an RNN with LSTM to capture sequential data, and the TSS applies linguistic knowledge for improved word embeddings. Fitrianah and Jauhari [30] employed LSTM and GRU models in their approach for ETS summarization, leveraging feature engineering techniques. Hin et al. [31] presented LineVD, a deep learning model that identifies vulnerabilities using a Graph Neural Network (GNN) and notably omits vulnerability status in its analysis. Nallapati et al. [32] introduced Summarunner, an RNN-based model, where two RNN layers embed words and sentences, followed by logistic regression for sentence classification. Kobayashi et al. [33] proposed a method centered on embeddings and document-level similarities, representing words through embeddings, treating sentences as word collections, and documents as sentence collections. Chen et al. [34] introduced a deep reinforcement learning technique and an encoder-extractor framework for single-document summarization, extracting sentences post key feature selection [35]. Mikael et al. [36] leveraged continuous vector representations in RNN and achieved top results on the Opinosis dataset. Yin et al. [37] devised a unique sentence selection strategy ensuring a balance

between sentence significance and diversity after developing an unsupervised CNN for phrase representation learning.

In recent times, there has been a transformative shift in the realm of natural language processing, largely attributed to BERT. BERT, a model based on the transformer architecture, has brought about a significant evolution in the domain of NLP by introducing contextual comprehension of words and sentences. In contrast to prior models that processed sentences in a linear manner, BERT takes into account both antecedent and subsequent words, thereby capturing a deeper insight into contextual interdependencies present within the text. This bidirectional approach empowers BERT to construct word representations that carry more profound significance, accurately reflecting their contextual applications. The integration of BERT has yielded notable enhancements across diverse NLP tasks, encompassing aspects such as text classification, identification of named entities, evaluation of sentiment, and particularly, condensing texts. By incorporating BERT into frameworks for text summarization, researchers have achieved summaries that are not only more precise but also informed by the context. BERT's proficiency in grasping linguistic intricacies and generating comprehensive portrayals has led to a paradigm shift in the way we handle and comprehend natural languages. This has, in turn, paved the way for the development of more intricate and efficient NLP applications. As influence of the BERT model continues to stimulate progress in language modeling and comprehension, it carries immense potential for further reshaping the landscape of natural language processing. Koto et al. [38] introduced techniques for probing discourse at the document level, which were utilized to detect connections between documents and appraise the performance of pre-trained language models. They employed BERT, BART, and RoBERTa as model choices to assess the outcomes derived from their assessment. In a separate study, Abdel-Salam and Rafea [39] conducted an evaluation of diverse variations of BERT-based models intended for text summarization. They introduced an unsupervised strategy for creating summaries from multiple documents, leveraging the transfer learning capabilities of the BERT sentence embedding model. The researchers adjusted the BERT model through supervised intermediate tasks extracted from GLUE benchmark datasets. This adjustment included the use of both single-task and multi-task fine-tuning methodologies to enhance the learning of sentence representations. In a different context, Srikanth et al. [40] harnessed the potential of the BERT model to produce extractive summaries through the clustering of sentence embeddings using K-means clustering. Alongside this, they introduced a dynamic approach to ascertain the suitable quantity of sentences to be chosen from the clusters.

A considerable portion of prior deep learning methods face a constraint during sentence selection. They often exhibit an inclination to excessively prioritize the selection of the sentence with the highest score, neglecting its pertinence within the context of subsequent sentence selection. As a result, this methodology contributes to a reduction in the overall excellence of the produced summary.

III. The Proposed Method

To tackle our research challenge, we have combined the strengths of DistilBERT for word embeddings and TLSTM for analyzing temporal data.

DistilBERT, a streamlined variant of BERT, excels in converting words into pertinent vectors. It mirrors BERT's bidirectional transformer architecture but is more efficient due to fewer layers, resulting in quicker computations. Notably, its training employs dynamic rather than static masking.

TLSTM excels at handling sequential data by grasping both short and long-term patterns. Its unique gating mechanism modulates data retention and recall over time.

We have also harnessed GANs for text summarization to boost extractive techniques and combat issues like greediness. Our GAN setup includes a generator, which creates synthetic data, and a discriminator that distinguishes between authentic and fabricated content. We've further enhanced our model by conditioning it on sentence features, allowing more accurate and relevant outputs. Integrating noise into the document representation lets our generator craft varied but consistently high-quality summaries.

In the following, the details of each component are explained.

A. DistilBERT-based word Embedding

The objective of word embedding is to transform words into vectors for utilization in deep learning algorithms. Word embedding has demonstrated its credibility in generating reliable vectors for words, drawing from the surrounding context. Diverse methodologies for word embedding have been introduced, each designed to produce substantial representations suitable for deep models. These approaches encompass Skip-gram [41] and matrix factorization techniques such as GloVe [42].

BERT stands as a profound bidirectional language model, capable of furnishing contextual portrayals. It is frequently subjected to fine-tuning through a hefty neural network layer tailored to various classification undertakings. Its training data encompasses extensive datasets such as Wikipedia. The initial, broadly applicable significance acquired during pre-training can be effectively harnessed to capture context- or issue-specific nuances through the process of fine-tuning, also readying it for classification purposes. BERT adopts a bidirectional transformer architecture, wherein representations are concurrently influenced by both preceding and subsequent context, spanning all tiers. This distinctive characteristic distinguishes BERT from models like GloVe and Word2Vec, which offer embeddings in a singular direction that disregards the contextual intricacies.

DistilBERT integrates the concept of information distillation, wherein a condensed system, the student, learns from a more expansive system's patterns, labeled as the teacher. The student system's learning curve is shaped by a specific loss criterion, reflecting the teacher's probability benchmarks:

$$L_{ce} = \sum_i t_i * \log\ (s_i) \quad (1)$$

Here, t_i and s_i represent the probabilities obtained from the teacher and student techniques, respectively. DistilBERT employs a structure akin to that of BERT, yet with fewer layers. DistilBERT exhibits a 40% reduction in width, operates with 60% enhanced speed, while still retaining 97% of the performance capabilities inherent to BERT. The core aim of the distillation process lies in approximating the comprehensive output distributions of BERT by means of a more condensed model, exemplified by DistilBERT. Consequently, the quantity of layers within the BERT architecture has been curtailed from 12 to 6. The pre-trained model encompasses a total of 66 million parameters, a comparison to the 110 million presents in the BERT model. Notably, DistilBERT's training duration amounts to 3.5 GPU days (using 8 × V100), in contrast to BERT's 12 GPU days (also with 8 × V100). DistilBERT, much like BERT, undergoes training using a dataset of 16 GB sourced from English Wikipedia, specifically the Toronto books corpus. During the training process of DistilBERT, a substantial batch size is employed in conjunction with gradient accumulation. This methodology entails the local amalgamation of gradients from multiple mini-batches prior to the modification of trainable parameters in each phase. Additionally, the training regimen of DistilBERT does not incorporate objectives such as next-sentence prediction and segment embedding learning, which are observed in BERT training. Moreover, the dynamic masking technique employed during inference replaces the static masking mechanism used in the BERT model.

B. TLSTM

LSTM has risen in prominence as a widely embraced and potent method applied to sequence data across diverse domains. Its inherent aptitude to apprehend extended temporal relationships and manage sequential information renders it exceptionally fitting for tasks involving time series analysis and prognosis [43]. An eminent virtue of LSTM lies in its capacity to unravel intricate temporal structures and seize the fluid dynamics of systems that undergo time-driven fluctuations. By dissecting the inherent motifs and inclinations embedded within the data, LSTM architectures can unveil nuanced interconnections that may not be readily discerned using established statistical or machine learning methodologies.

LSTM networks possess the capability to grapple with input sequences of varying lengths, thereby endowing them with adaptability for scenarios involving sequence data wherein the historical data's length may fluctuate across instances [44]. This adaptative trait proves invaluable when confronted with diverse systems or equipment beset with differing operational states or maintenance schedules. Furthermore, LSTM's prowess in dealing with both brief and extensive dependencies within sequence data sets it apart. Traditional methods like autoregressive models or moving average approaches might falter in capturing prolonged trends or subtle intricacies concealed within the data fabric [45]. In contrast, LSTM's memory cells empower it to retain information over extended intervals, thereby empowering the model to apprehend dependencies spanning multiple time increments [46, 47].

The groundbreaking concept of LSTM was originally developed by Hochreiter and Schmidhuber [48]. From its genesis, an array of methods aimed at enhancing its performance have been introduced [49]. In this research, we put into practice the well-accepted architecture advanced by Gers et al. [50], a blueprint that has been leveraged in a multitude of academic endeavors, including [51, 52]. The LSTM mechanism revolves around a gating system which regulates how information is retained over time, skillfully overseeing how long it is stored and determining the appropriate moment for its access through the memory cell. This paper places particular emphasis on the scrutiny of the LSTM cell, as expounded in Graves' work [51]. LSTM employs three gates to optimize information processing. Let i_t, f_t, o_t, c_t, and h_t symbolize the input gate, forget gate, output gate, memory cell, and hidden state at the sequence time t, respectively. When x_t represents the system's input at the same time, the architecture of the LSTM cell can be described as follows [51, 53]:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (2)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \quad (3)$$

$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (4)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (5)$$

$$h_t = o_t tanh(c_t) \quad (6)$$

The sigmoid function, σ(.), acts as an activation function. The logistic sigmoid and hyperbolic tangent are applied element-wise. Weight matrices, W_{xk} and W_{ck}, are linked to the input, forget, output gates, and memory cell. The number of neurons in these gates is preset, with Eq. (2) to Eq. (6) affecting each neuron separately. If n represents the neuron count, then $\{i_t, f_t, c_t, o_t, h_t\}$ are in $R^{n\times 1}$. In discussing the LSTM model, we use w_{lstm} and b_{lstm} for weights and biases. The LSTM equations are presented as follows:

$$\begin{cases} c_t = f(c_{t-1}, h_{t-1}, x_t; w_{lstm}, b_{lstm}) \\ h_t = g(h_{t-1}, c_{t-1}, x_t; w_{lstm}, b_{lstm} \end{cases} \quad (7)$$

Considering Eq. (2) through (6), we derive $g(\cdot)$ and $f(\cdot)$. Assuming $z(\eta)$ represents an unseen series, the state space depiction of the T-LSTM is expressed as:

$$\begin{cases} c_{t,\eta} = f(c_{t-1,\eta}, h_{t-1,\eta}, x_t; w_{lstm,\eta}, b_{lstm,\eta}) \\ h_{t,\eta} = g(h_{t-1,\eta}, c_{t-1,\eta}, x_t; w_{lstm,\eta}, b_{lstm,\eta} \end{cases} \quad (8)$$

In Eq. (8), the model structure markedly deviates from what is outlined in Eq. (7). While in Eq. (7) the model parameters remain stable irrespective of the evaluation point, in Eq. (8) they are molded by the feature vector of that specific evaluation point. The subscript η is introduced to underscore the variation in model parameters resulting from the inclusion of data point z(η). It is imperative to underscore that the evaluation label is presumed to be undisclosed. Throughout the training phase, the primary purpose of the evaluation point is to ascertain the relevance of training data points by examining the affinity between their feature vectors and the vector of the evaluation point.

C. Model

In this research, we use adversarial generating networks for extractive summarization. We employ this network to improve the problems of previous methods, including greed. We will first have a description of this network, and then the proposed model is presented.

Generative adversarial networks (GANs) were first proposed by Goodfellow et al. [54]. These networks consist of two separate networks that are similarly trained: the generator and discriminator networks. The purpose of the generator is to produce data such as images, text, etc., which are structurally similar to real data but are fake. On the other hand, the task of the discriminator network is to strengthen the generator.

These two networks play a two-player min-max game with a value function $V(D, G)$ as follows:

$$min_G\, max_D\, V\,(D.G) = E_{x\sim p_{data}(x)}[\log(D(x))]+ E_{z\sim p_z(z)}[log(1 - D(G(z)))] \quad (9)$$

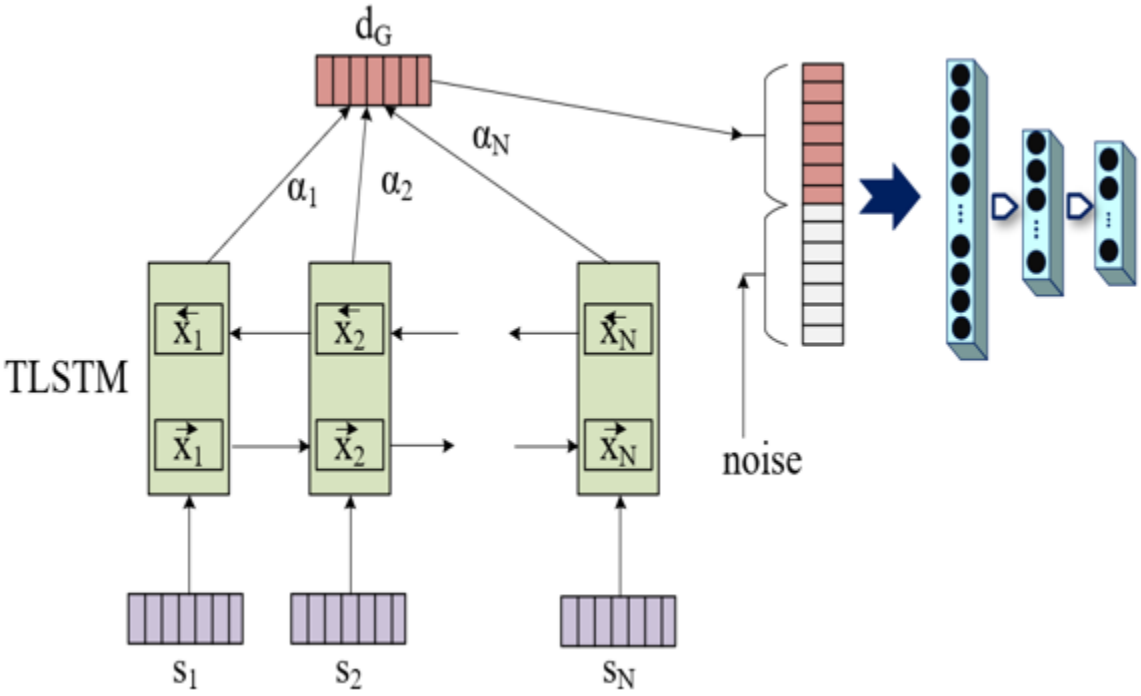


Fig. 1. Generator architecture.

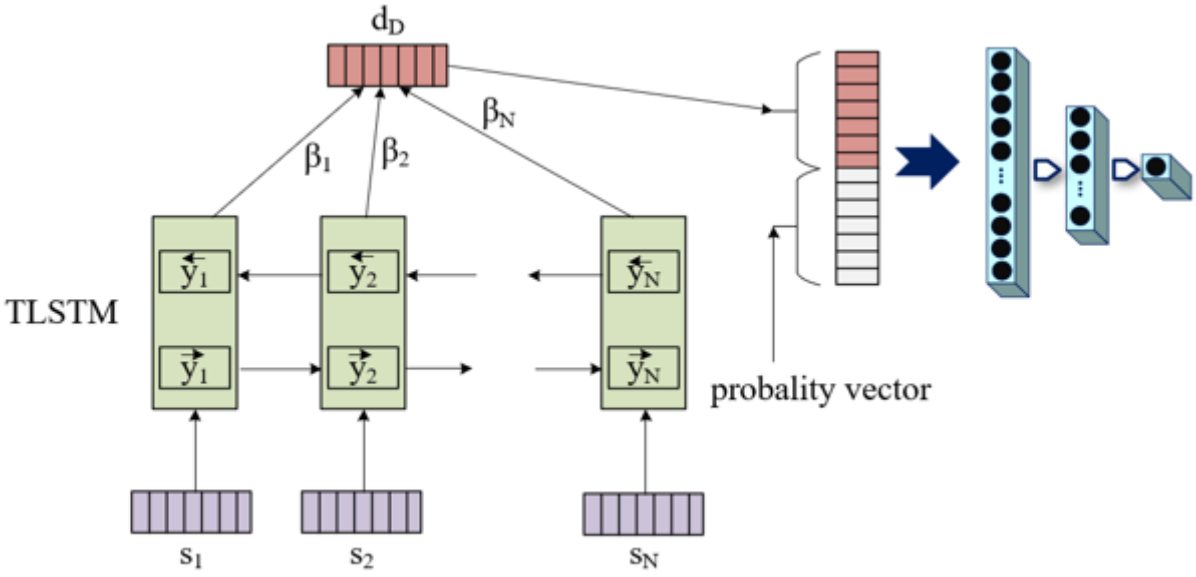


Fig. 2. Discriminator architecture.

Where x and z are input data and noise, respectively. G and D mean the generator and discriminator, respectively. $p_{data}(x)$ and $p_z(z)$ represent the input data distribution and the noise distribution, respectively. E is mathematical expectation.

Generative adversarial networks can be extended to a conditional model If condition y is added to the generator and discriminator input. The value function, in this case, changes as follows:

$$min_G\, max_D\, V\,(D.G) = E_{x\sim p_{data}(x)}[\log(D(x|\boldsymbol{y}))]+ E_{z\sim p_z(z)}[log(1-D(G(z|\boldsymbol{y})))] \quad (10)$$

The proposed generator and discriminator model are shown in Fig. 1 and Fig. 2, respectively. We use sentence features as a condition in the generator and discriminator. Let $D = \{s_1.s_2 \ldots .s_N\}$ represents the document, where the $s_i \in \mathbb{R}^d$ is the extracted features of the i-th sentence. N is the length of document D, which is equal to the number of restricted sentences in each document. The attention mechanism calculates the representation vector of the document in the generator and the discriminator according to the following equations:

$$d_G = \sum_{i=1}^{N} \alpha_i\, [\overleftarrow{x}_i.\vec{x}_i] \quad (11)$$

$$d_D = \sum_{i=1}^{N} \beta_i\, [\overleftarrow{y}_i.\vec{y}_i] \quad (12)$$

where, $\overleftarrow{x}_i \in \mathbb{R}^{d_1}$, $\overleftarrow{x}_i \in \mathbb{R}^{d_1}$, $\overleftarrow{y}_i \in \mathbb{R}^{d_2}$, $\vec{y}_i \in \mathbb{R}^{d_2}$ are the output of step i in BLSTM. α_i and β_i are the coefficients of attention for the i-th sentence in the generator and the discriminator, respectively, which are formulated as follows:

$$\alpha_i = \frac{e^{u_i}}{\sum_{i=1}^{N} e^{u_i}} \quad (13)$$

$$\beta_i = \frac{e^{v_i}}{\sum_{i=1}^{N} e^{v_i}} \quad (14)$$

$$u_i = tanh(W_u[\overleftarrow{x}_i.\vec{x}_i] + b_u) \quad (15)$$

$$v_i = tanh(W_v[\overleftarrow{y}_i.\vec{y}_i] + b_v) \quad (16)$$

where, $W_u \in \mathbb{R}^{2.d_1}.b_v \in \mathbb{R}$, $W_v \in \mathbb{R}^{2.d_2}$, and $b_v \in \mathbb{R}$ are the parameters of the attention mechanism for documents.

In Fig. 1, the document's representation vector is linked with the noise vector, entering a feed-forward neural network. The final layer of this network computes the likelihood of each sentence's presence. The introduction of noise prompts the generator to generate diverse outputs. Each document undergoes multiple iterations of summarization by the generator, with varied noises, leading to distinct outputs. The generator aims to create varied yet similarly high-quality summaries for each document. This process empowers the generator to identify diverse sentence combinations suitable for crafting the summary. Consequently, sentences that might lack individual significance for the summary can contribute to a quality summary when positioned alongside other sentences.

Within the discriminator network, the probability vector of sentences interfaces with the document's representation vector (see Fig. 2). In this context, the probability vector of sentences represents the count of sentences within a document, and each element assumes a value of either zero or one.

1) Real summary: In a typical GAN framework, the output of the generator functions as synthetic data for training the discriminator. Moreover, an authentic target is garnered for each individual sample. In this study, to acquaint the discriminator with quality summaries, more than one summary is drawn from each document, displaying similar levels of quality. Simultaneously, several summaries of inferior quality are generated for each document. Given that actual summaries are text and unsuitable as target data, a method is required to represent the presence or absence of each sentence in a summary as a numerical value. To serve this purpose, a vector with N elements is designated for every document, where N signifies the sentence count. Each element of this vector holds a value of either zero or one, with a value of one indicating the inclusion of the sentence in the summary. This vector is constructed employing a greedy approach as delineated in Fig. 3. Initially, a vector of length N with M ones is generated, where M corresponds to the sentences within the summary. The ones are distributed randomly throughout the vector. Sentences associated with a value of one are then concatenated within this vector to compose a summary, and the quality is assessed using the ROUGE metric.

Subsequently, a one is selected at random and transformed into zero, while a randomly selected zero is converted to one. The ROUGE score is recalculated, and if it surpasses the prior value, the alteration is retained. This sequence is reiterated Itr times, culminating in the selection of the most favorable vector throughout the process, which is then designated as the outcome. It is important to note that for the generation of any genuine target, the algorithm must be restarted from the beginning. The procedure for devising a synthetic target closely mirrors that of a real target, with the exception that if the ROUGE score is lower, the vector supersedes the previous one. The length of all documents is confined to N sentences, with longer documents being truncated to N sentences and shorter ones being padded with zeros.

2) Loss function: The Loss function is calculated based on the discriminator output for the generator as follows:

$$Loss_G = E_{i\sim Dataset}\left[E_{z\sim p_{z(z)}}\left[log\left(1-D\big(G(z|y_i)\big)\right)\right]\right] \quad (17)$$

where, *Dataset* is a set of documents, y_i is features of sentences in document i, and E is the mathematical expectation. The Loss function for the discriminator is computed based on the generator output, real and fake summaries as follows:

$$Loss_D = E_{i\sim Dataset}[\,E_{z\sim p_{z(z)}}[log(1-D(G(z|y_i)|\,y_i))] + E_{k\sim p_{Fake_i}}\big[log\big(1-D(k|y_i)\big)\big] + E_{k\sim p_{Real_i}}[log(D(l|y_i))]] \quad (18)$$

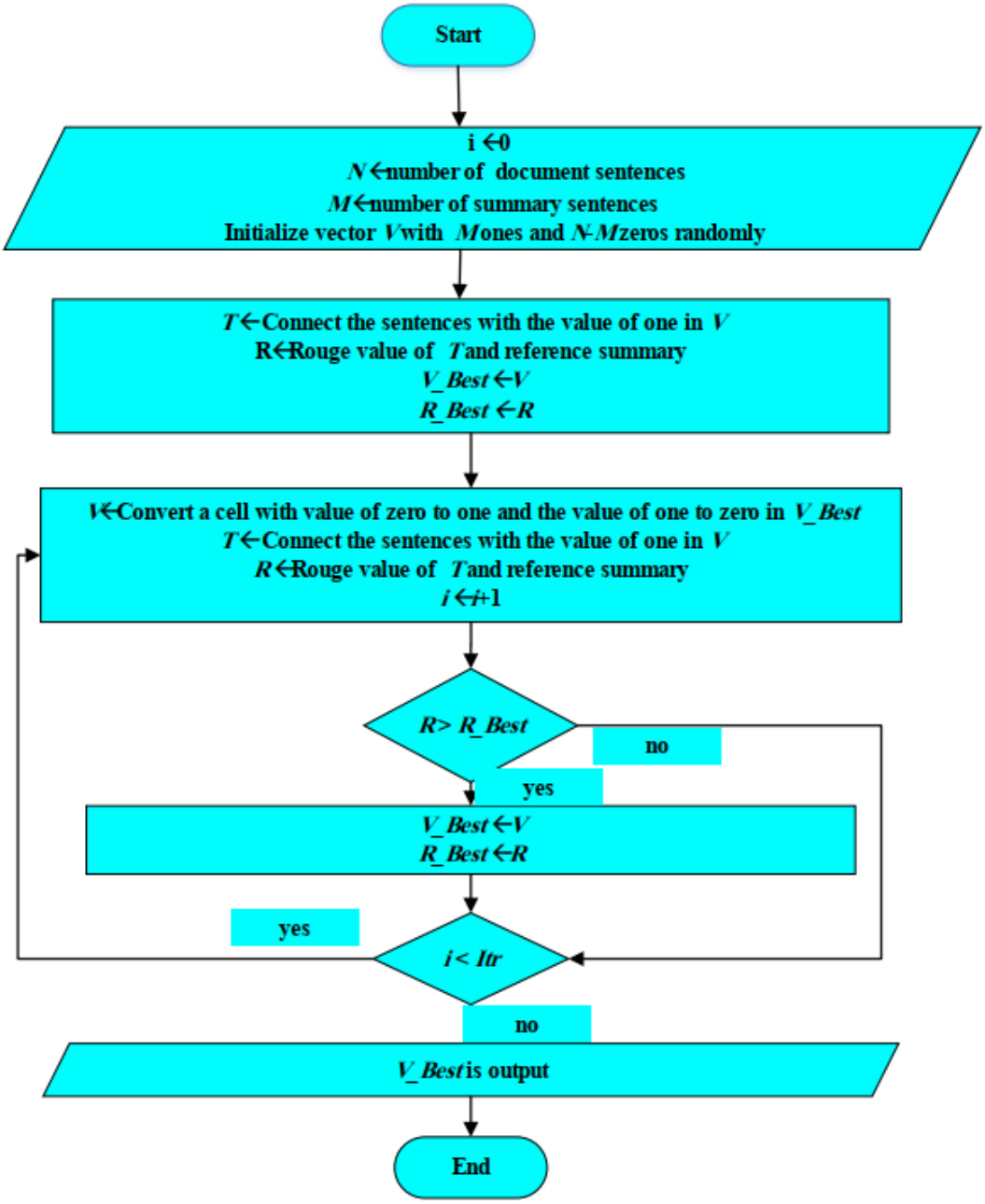


Fig. 3. Generate a real summary for the document.

where, p_{Real_i} and p_{Fake_i} show the distribution of real and fake summaries for the document *i*. Eq. (18) forces the discriminator to learn a set of high-quality and low-quality summaries. On the other hand, be sensitive to the summaries produced by the generator and force the generator to produce a high-quality summary.

IV. Empirical Evaluation

A. Dataset

For our assessments, we employ a familiar dataset known as CNN/Daily Mail. This dataset amalgamates two distinct datasets devised for comprehension, extractive, and abstractive tasks, and it has garnered notable attention from researchers in the domain of automated summarization in recent years. The CNN/Daily Mail dataset is comprised of 287,226 documents earmarked for training, 13,368 for validation, and 11,490 for testing. Within the training data, the average document encompasses approximately 28 sentences. On average, each document's reference summary spans 3 to 4 sentences, and the mean word count per document in the training dataset is approximately 802 words [34]. You can delve into additional particulars outlined in Table I. This dataset exists in two versions: the first version features the replacement of all entities with specific words, while the second version retains the original data. For our model, we choose to adopt the second version of the dataset.

TABLE I. STATISTICS OF THE CNN / DAILY MAIL DATASET

	Train	Validation	Test
Pairs of data	287,113	13,368	11,490
Article length	749	769	778
Summary Length	55	61	58

TABLE II. THE PARAMETERS OF THE MODEL

Parameter	value
batch size	128
embedding dim	60
max sentence length	100
real summary per document	40
fake summary per document	40
activation fun (tlstm & dense)	relu
dense hidden layer	8

B. Detail of Model

For the execution of this study, the Python programming language and the PyTorch library have been harnessed for implementation purposes. The Jupyter environment has been employed as the platform to execute project codes. Additionally, the NLTK library, an instrumental component, has been utilized. This particular library furnishes an assortment of classes and methods dedicated to processing natural language within the Python context. Its capabilities span a broad spectrum of natural language processing tasks.

The model architecture incorporates a dual-layer bidirectional TLSTM structure. Within generative adversarial networks, the discriminator tends to converge at a quicker rate than the generator, often impeding the generator's convergence. In light of this, we have designed a training strategy wherein the discriminator is trained once for every 15 iterations of generator training. Moreover, due to the interconnection of vectors within the two networks, we implement batch normalization prior to data entry into the feed-forward neural network. The parameter values are detailed in Table II.

C. Metrics

We employ the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) package [55] as an evaluation metric in our experiments. This metric calculates the similarity between the generated summary and the reference summary by counting the number of common units. Rouge-n recall between an extracted summary and a reference summary is calculated as follows:

$$\text{Rouge-}n = \frac{\sum_{s \in \{ref\ sum\}} \sum gram_n \in s\ Count_{match}\ (gram_n))}{\sum_{s \in \{ref\ sum\}} \sum gram_n \in s\ Count\ (gram_n))} \quad (19)$$

where, n stands for the length of n-ngram, $Count_{match}\ (gram_n)$ is the maximum number of n- gram co-occurring in the extracted summary and the reference summary. Rouge-1 and Rouge-2 are special cases of Rouge-n in which n = 1 or n = 2. R-L calculates the length of the longest common subsequence between the reference summary and the extracted summary. Based on previous works, Rouge-1(R-1), Rouge-2(R-2) and, Rouge-L(R-L) are most widely used in summarization. For this reason, we use these three metrics in all our experiments.

D. Experimental Results

In the execution of our project, we utilized a Windows operating system that runs on 64-bits, accompanied by 64 GB of RAM and an integrated GPU. For the CNN/Daily Mail dataset, the optimal model emerged after running through 50 epochs. Remarkably, the entirety of our training duration spanned a mere four hours.

Our innovative approach was subjected to a comparative analysis against various methodologies. These included three methodologies rooted in graph algorithms: BGSumm [56], TextRank [57], and EdgeSumm [8]. Additionally, we compared against seven methodologies anchored in deep learning paradigms: SummaRunner [32], RENS with Coherence [58], SHANN [59], HSSAS [60], T5 [61], BART [62], and DeepSumm [63]. Lastly, a foundational model, TLSTM, was part of our comparison. It's noteworthy to mention that the TLSTM model is exclusively reliant on the generator component we devised. To visualize the assessment results for our system using the CNN/Daily Mail dataset, please refer to Table III.

TABLE III. NUMERICAL COMPARISON OF THE PROPOSED METHOD AND OTHER METHODS ON THE CNN / DAILY MAIL DATASET

Model	R-1	R-2	R-L
BGSumm	33.20	12.50	31.74
TextRank	32.16	11.10	29.21
EdgeSumm	34.26	13.10	32.90
DeepSumm	42.91	18.18	37.95
SummaRunner	39.65	16.26	35.39
RENS with Coherence	41.29	18.90	37.79
SHA-NN	35.46	14.74	33.26
HSSAS	42.32	17.81	37.65
T5	42.48	18.08	37.77
BART	36.51	15.14	31.26
TLSTM	26.46	8.48	6.25
Proposed	44.51	18.46	38.90

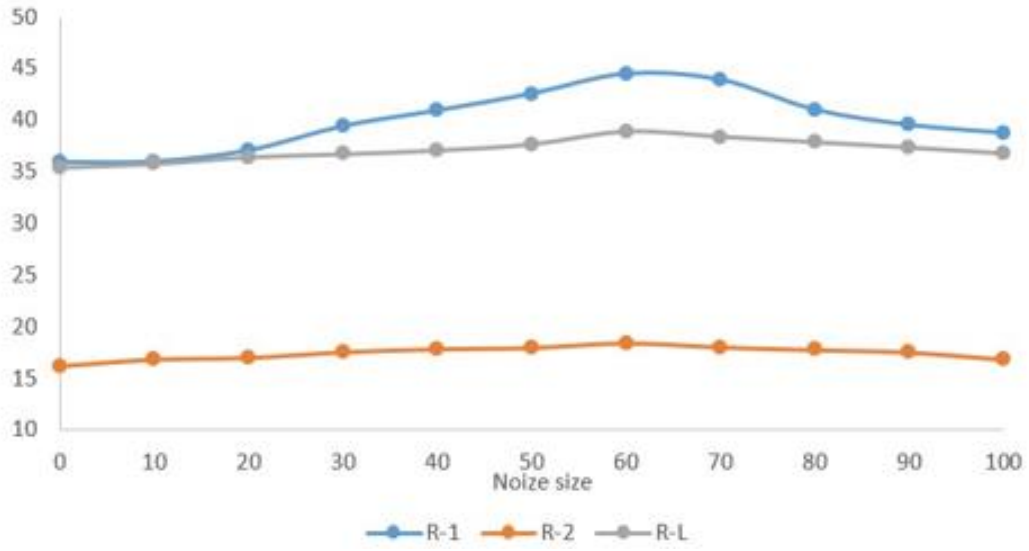


Fig. 4. Results of the proposed model for different noises on the CNN / Daily Mail dataset.

Drawing insights from the graph-centric models, the EdgeSumm model conspicuously outperformed its peers, inclusive of BGSumm, across all evaluated benchmarks. To quantify, EdgeSumm decreased errors by magnitudes of more than 33%, 32%, and 30% for the three primary metrics: R-1, R-2, and R-L, respectively. Intriguingly, even though BGSumm demonstrated its efficiency on a medical dataset, it couldn't replicate its performance for the CNN/Daily Mail dataset. Surpassing even the robust EdgeSumm model, our pioneering model showcased error enhancement rates of approximately 24.29%, 24.30%, and 25.41%. As many would anticipate, models anchored in deep learning exhibited greater efficacy than those rooted in graph algorithms. The RENS with Coherence approach, despite its integration of sentence coherence, didn't match the precision of our model. Among the pantheon of deep learning models, DeepSumm emerged as the frontrunner. However, even DeepSumm lagged behind our proposed model, registering weaker performance metrics of 25.28%, 25.39%, and 26.47%.

1) Explore noise: We undertook additional experimental trials to ascertain the impact of varying noise intensities on the generator's functionality. In these trials, we introduced noise of diverse magnitudes to the generator to observe its effect. The outcomes, specifically for metrics R-1, R-2, and R-L pertaining to the CNN/Daily Mail dataset, are graphically presented in Fig. 4. A noteworthy observation was that elevating the noise level to 60 improved the aforementioned metrics: R-1, R-2, and R-L. However, a declining trend in performance was evident when the noise level ranged between 60 and 100. From our analysis of this dataset, it appears that the optimal noise magnitude stands at 60. It is evident from the findings that the generator's efficiency is enhanced when noise is incorporated, with our proposed model displaying superior results in the presence of noise.

2) Word embedding: Word representations play a pivotal role in the realm of deep learning models. This is primarily because these models interpret input data as vectors; therefore, if there is any discrepancy or error in the embedding process, it could potentially misguide the model. In this research, we have employed the DistilBERT model for word embeddings, which is considered one of the latest advancements in this domain.

TABLE IV. RESULTS OF DIFFERENT WORD EMBEDDINGS ON THE MODEL

Model	R-1	R-2	R-L
One-Hot encoding	23.45	8.11	27.80
CBOW	35.86	12.01	30.80
Skip-gram	36.04	12.14	31.11
GloVe	39.30	15.26	35.39
FastText	40.14	16.98	36.89
BERT	43.40	17.33	37.44
DistilBERT	44.51	18.46	38.90

To rigorously assess the efficacy of various word embedding techniques in tandem with our model, we introduced five different embeddings for our evaluation: One-Hot encoding, CBOW, Skip-gram, GloVe, FastText, and the original BERT model [45]:

- One-Hot Encoding: This basic yet foundational technique is essential for translating categorical variables into a format that can be fed into deep learning models, thereby optimizing prediction and classification outcomes. The essence of this approach lies in representing each unique category with a distinct binary code, ensuring only one bit is "hot" or set to '1' for every class representation.
- CBOW and Skip-gram: These are sophisticated models that employ neural networks to associate words with their respective embedding vectors. Their operational methodologies might differ, but they share a common goal.
- GloVe: This unsupervised learning model taps into the aggregated co-occurrence data of global word pairs from a given corpus, providing a distinctive representation for words.
- FastText: Pioneering an evolution of the Skip-gram model, FastText takes a novel approach by encoding words as letter n-grams rather than representing them as unique vectors.

For a comprehensive understanding of our results, one should consult Table IV. It was anticipated, and the results confirmed, that One-Hot encoding lagged behind other embeddings, registering suboptimal performance. The improvement metrics for our proposed model using this method were approximately 62.70% (R-1), 9% (R-2), and 18% (R-L). Intriguingly, CBOW and Skip-gram, given their analogous architecture, exhibited similar performances, with both overshadowing the GloVe embedding. Among the lot, the BERT model emerged as the most competent word embedding technique. However, its effectiveness diminished slightly when juxtaposed with the DistilBERT model. In comparison to BERT, DistilBERT demonstrated a reduction in errors by 11% (R-1), 10% (R-2), and 19% (R-L).

The real summary is, “A Canadian doctor says she was part of a team examining Harry Burkhart in 2010, Diagnosis: autism, severe anxiety, post-traumatic stress disorder and depression, Burkhart is also suspected in a German arson probe, officials say, Prosecutors believe the German national set a string of fires in Los Angeles”.

3) Examples: Using a practical illustration of the generator's functionality, we've provided three sentences from a document within the CNN/Daily Mail dataset. These sentences, as extracted by the generator, can be viewed alongside their corresponding reference summary in Fig. 5. Upon inspection, it is evident that sentences sharing a greater number of words with the reference summary tend to receive elevated scores. This showcases the generator's ability to prioritize and assign higher scores to sentences that align more closely with the central themes or keywords present in the reference summary.

rank	Sentence	Score
1	Stancheva said she and other doctors including a psychiatrist diagnosed Burkhart with ”autism, severe anxiety, posttraumatic stress disorder and depression.”	0.95
2	Burkhart, a 24-year-old German national, has been charged with 37 counts of arson following a string of 52 fires in Los Angeles	0.86
3	A medical doctor in Vancouver, British Columbia, said Thursday that California arson suspect Harry Burkhart suffered from severe mental illness in 2010, when she examined him as part of a team of doctors.	0.76

Fig. 5. Three sentences extracted by the generator for the CNN / Daily mail dataset.

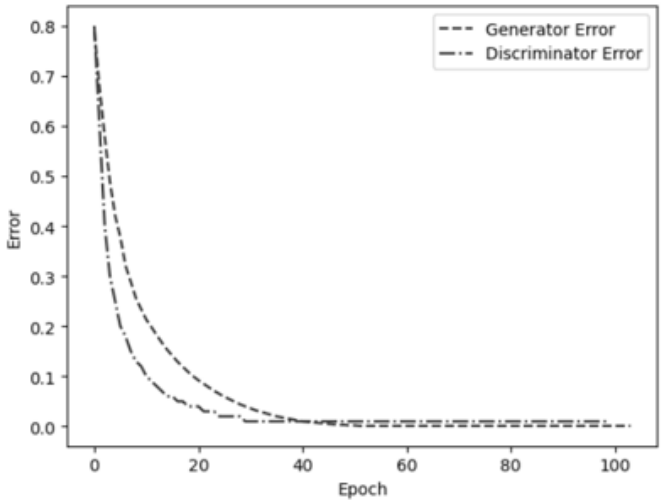


Fig. 6. Comparative diagram of error dynamics.

4) Discussion: This paper unveiled a groundbreaking approach to extractive text summarization, blending a multitude of advanced techniques. Our introduced methodology synergizes the power of a GAN-centric framework, the precision of DistilBERT word embeddings, and the adaptability of TLSTM. Central to efficacious summarization is the interplay between two components within the GAN architecture: the generator and the discriminator. The generator's primary task is to gauge the significance of each sentence in a prospective summary. In contrast, the discriminator's role is to critique and ascertain the caliber of the summaries generated. Such a dynamic within the GAN structure empowers the generator, encouraging it to sift through an array of sentence permutations. As a result, this culminates in the creation of summaries that are both concise and of superior quality. Adding another layer of sophistication is TLSTM, which harnesses the power of transductive learning. Transductive learning is distinctive in its approach, as it assigns augmented weights to data samples that are proximate to a specific test point. This ensures that the most relevant and closely aligned data samples exert greater influence, optimizing the summarization process.

Fig. 6 presents error diagrams for both the generator and discriminator in a GAN across various epochs. Initially, the generator's error is markedly higher at 0.789, revealing its challenges in generating samples mimicking the true data distribution. Yet, as training progresses, the error of the generator exhibits a clear decreasing trend. This suggests the generator is progressively getting better at simulating genuine data, capturing intricate patterns within the dataset. Simultaneously, the discriminator, starting with a slightly lower error of 0.8, undergoes its own evolution. Its role is to differentiate between real and synthetic data. Over the epochs, its error also reduces, though not as sharply as the generator. This indicates that even as the discriminator becomes more skillful, the generator is advancing at a slightly faster pace, producing ever more convincing samples. The interplay between these adversarial elements is crucial to the convergence of the GAN. A diminishing error for both entities across epochs implies a harmonious convergence in the GAN training. The generator refines its outputs, drawing them closer to real samples, while the discriminator sharpens its evaluative abilities. This consistent drop in error highlights the stability and continuous advancement of the GAN in training. The model aptly leverages the adversarial dynamic between its components to enhance performance over epochs. In essence, Fig. 5 underscores the iterative refinement in GAN training, each step bringing the system closer to generating more credible synthetic data distributions.

The paper touts the proposed model's preeminence, substantiating its claims through performance metrics derived from the ROUGE evaluation on the CNN/Daily Mail dataset. Nevertheless, this evaluation is tethered to just one dataset, raising questions about the model's adaptability across a spectrum of diverse datasets. To genuinely encapsulate the model's prowess, it would be prudent to undertake assessments across a myriad of datasets. Relying solely on the CNN/Daily Mail dataset might pave the way for dataset-specific biases. The rationale behind this exclusive dataset choice warrants elucidation, and an exploration into the model's versatility across varied datasets is imperative. Multiple datasets inherently encapsulate nuances in linguistic style, domain specificity, and content diversity, which undeniably bear implications on the outcomes of text summarization. A more holistic evaluation, spanning multiple datasets, would invariably render a more nuanced understanding of the model's capabilities.

The paper touts the proposed model's preeminence, substantiating its claims through performance metrics derived from the ROUGE evaluation on the CNN/Daily Mail dataset. Nevertheless, this evaluation is tethered to just one dataset, raising questions about the model's adaptability across a spectrum of diverse datasets [64]. To genuinely encapsulate the model's prowess, it would be prudent to undertake assessments across more datasets. For this, we can use datasets presented in [65]. Relying solely on the CNN/Daily Mail dataset might pave the way for dataset-specific biases. The rationale behind this exclusive dataset choice warrants elucidation and an exploration into the model's versatility across varied datasets is imperative. Multiple datasets inherently encapsulate nuances in linguistic style, domain specificity, and content diversity, which undeniably bear implications on the outcomes of text summarization. A more holistic evaluation, spanning multiple datasets, would invariably render a more nuanced understanding of the model's capabilities [66].

GANs are built upon a novel framework where two neural networks, the generator and the discriminator, work in tandem. The generator's primary goal is to produce outputs that are indistinguishable from real data, while the discriminator's objective is to differentiate between actual data and the data generated by the generator. This adversarial process, though powerful in theory, presents several practical challenges, particularly during the training phase. One primary concern is the issue of convergence. Given the dynamic nature of the adversarial relationship, ensuring that both networks converge to an optimal solution is not straightforward. If not carefully managed, the training can end up in a loop where each network

constantly tries to outdo the other without reaching a stable equilibrium. This oscillatory behavior can make GANs particularly sensitive to hyperparameters, initialization, and the chosen architecture, often requiring extensive experimentation and fine-tuning. Additionally, the delicate balance between the generator and discriminator can easily be disrupted. If the discriminator becomes exceptionally adept early on in training, it can stifle the generator's ability to learn. The generator, facing constant rejection from the discriminator, may struggle to make any meaningful progress, leading to a stagnation in learning and potentially resulting in mode collapse, where the generator produces limited or repetitive outputs. On the other hand, if the generator dominates the learning process and continually manages to deceive the discriminator, the discriminator may fail to provide meaningful feedback. This can result in generated summaries that, while convincing at first glance, might stray from the original content's essence, compromising the quality and relevance of the output. Given these challenges, there's a growing consensus in the research community about the need for more refined training strategies for GANs [67]. Techniques such as gradient penalty [68], spectral normalization [69], and modified loss functions [70] have been proposed to stabilize GAN training.

Transductive learning is a unique learning paradigm that seeks to make predictions specifically for the given test set without generalizing to the broader population. By concentrating on samples near the test point, it can produce highly optimized results for a specific set of data. However, this precision comes with its own set of challenges, primarily related to model generalization [71]. The inherent nature of transductive learning to prioritize certain instances over others can inadvertently lead the model to capture noise or idiosyncrasies present in the training data. Such a model would be finely tuned to a particular dataset, but might falter when introduced to new, unseen data. This phenomenon, known as overfitting, means that while the model performs exceptionally well on its training data, its performance significantly drops on new, unfamiliar data. In practical scenarios, especially in dynamic environments like news summarization, social media analytics, or customer feedback systems, data distributions can shift rapidly. A model trained with a strong transductive bias might not adapt well to these changing scenarios, thus compromising its effectiveness and reliability. It would continuously require retraining or fine-tuning on new data points, which is resource-intensive and not always feasible. Addressing these challenges necessitates a more balanced approach to learning. One potential avenue is the incorporation of regularization techniques [72]. Regularization, in essence, adds a penalty to the loss function, discouraging the model from fitting too closely to every data point and, in turn, mitigating overfitting. Techniques such as L1 and L2 regularization or dropout [73] can be applied to ensure the model retains a level of generality. Furthermore, blending transductive learning with inductive learning offers another promising solution. While transductive learning focuses on specific test points, inductive learning aims to find a general pattern or hypothesis that can be applied to any input. By combining these two paradigms, one could harness the precision of transductive learning while maintaining the broader applicability provided by inductive learning. Such a hybrid approach would not only cater to specific data instances but also ensure that the model remains versatile and adaptable to a range of data distributions.

V. CONCLUSION

In the research presented, we introduce an innovative approach to extractive text summarization, leveraging a blend of GAN, the DistilBERT word embedding technique, and an attention-centric TLSTM methodology. Utilizing DistilBERT, we crafted a feature vector for each sentence, which was subsequently fed into the generator to deduce the likelihood of that sentence being part of the final summary. In tandem, a discriminator was employed to scrutinize the summaries churned out by the generator, thus honing its capabilities. We further innovated by designing a unique loss function tailored for the training of the discriminator. This function meticulously considers the output of the generator, as well as both authentic and contrived document summaries. An intriguing facet of our methodology is that each document is paired with distinct noise during both training and testing phases. Such a strategy empowers the generator, equipping it to explore a vast array of sentence amalgamations, laying the groundwork for the creation of superior quality summaries. Empirical evaluations conducted on the CNN/Daily Mail dataset lend weight to the efficacy of our model. The results not only underscore the effectiveness of our novel methodology but also highlight its superiority, outpacing other established text summarization techniques in performance metrics.

In forthcoming research endeavors, we intend to focus on enhancing the coherence among sentences within our model. Coherence plays an instrumental role in ensuring that the summarized text is not just a collection of sentences, but a fluid and cohesive narrative that is easy for readers to follow and understand. Addressing this aspect can significantly elevate the quality and readability of the generated summaries. One possible approach to achieve this would be to prioritize coherence during the construction of our target summaries. By doing so, the model would be trained to select sentences that not only contain critical information but also seamlessly connect with one another, ensuring a natural flow of ideas. Additionally, another promising avenue to explore is the incorporation of coherence as a loss function within the generator. By integrating coherence into the loss function, the generator would be incentivized to produce summaries where the sentences logically follow one another, leading to more cohesive and contextually relevant outputs. Introducing such modifications could provide dual benefits: improving the intrinsic quality of the summaries and enhancing the user experience, as coherent and logically structured summaries are more easily comprehensible. This, in turn, would further cement the model's applicability and usefulness in real-world scenarios, catering to a wider range of text summarization needs.

REFERENCES

[1] R. Mitkov, Handbook for Language Engineers edited by Ali Farghaly, Computational Linguistics, 30 (2004) 397-399.

[2] T. Hirao, H. Isozaki, E. Maeda, Y. Matsumoto, Extracting important sentences with support vector machines, COLING 2002: the 19th international conference on computational linguistics, 2002.

[3] S. Akter, A.S. Asa, M.P. Uddin, M.D. Hossain, S.K. Roy, M.I. Afjal, An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm, 2017 ieee international conference on imaging, vision & pattern recognition (icivpr), IEEE, 2017, pp. 1-6.

[4] S. Vakilian, S.V. Moravvej, A. Fanian, Using the artificial bee colony (ABC) algorithm in collaboration with the fog nodes in the Internet of Things three-layer architecture, 2021 29th Iranian Conference on Electrical Engineering (ICEE), IEEE, 2021, pp. 509-513.

[5] A. Kumar, A. Sharma, Systematic literature review of fuzzy logic based text summarization, Iranian journal of fuzzy systems, 16 (2019) 45-59.

[6] S. Vakilian, S.V. Moravvej, A. Fanian, Using the cuckoo algorithm to optimizing the response time and energy consumption cost of fog nodes by considering collaboration in the fog layer, 2021 5th International Conference on Internet of Things and Applications (IoT), IEEE, 2021, pp. 1-5.

[7] S.V. Moravvej, R. Alizadehsani, S. Khanam, Z. Sobhaninia, A. Shoeibi, F. Khozeimeh, Z.A. Sani, R.-S. Tan, A. Khosravi, S. Nahavandi, RLMD-PA: A reinforcement learning-based myocarditis diagnosis combined with a population-based algorithm for pretraining weights, Contrast Media & Molecular Imaging, 2022 (2022).

[8] W.S. El-Kassas, C.R. Salama, A.A. Rafea, H.K. Mohamed, EdgeSumm: Graph-based framework for automatic text summarization, Information Processing & Management, 57 (2020) 102264.

[9] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A.A. Bharath, Generative adversarial networks: An overview, IEEE signal processing magazine, 35 (2018) 53-65.

[10] S. Song, H. Huang, T. Ruan, Abstractive text summarization using LSTM-CNN based deep learning, Multimedia Tools and Applications, 78 (2019) 857-875.

[11] S.V. Moravvej, M. Joodaki, M.J.M. Kahaki, M.S. Sartakhti, A method based on an attention mechanism to measure the similarity of two sentences, 2021 7th International Conference on Web Research (ICWR), IEEE, 2021, pp. 238-242.

[12] S. Hochreiter, J. Schmidhuber, LSTM can solve hard long time lag problems, Advances in neural information processing systems, 9 (1996).

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805, (2018).

[14] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108, (2019).

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692, (2019).

[16] D.d.V. Feijo, V.P. Moreira, Improving abstractive summarization of legal rulings through textual entailment, Artificial intelligence and law, 31 (2023) 91-113.

[17] H. Oh, S. Nam, Y. Zhu, Structured abstract summarization of scientific articles: Summarization using full‐text section information, Journal of the Association for Information Science and Technology, 74 (2023) 234-248.

[18] P.K. Katwe, A. Khamparia, D. Gupta, A.K. Dutta, Methodical Systematic Review of Abstractive Summarization and Natural Language Processing Models for Biomedical Health Informatics: Approaches, Metrics and Challenges, ACM Transactions on Asian and Low-Resource Language Information Processing, (2023).

[19] M.T.R. Laskar, M. Rahman, I. Jahan, E. Hoque, J. Huang, CQSumDP: A ChatGPT-Annotated Resource for Query-Focused Abstractive Summarization Based on Debatepedia, arXiv preprint arXiv:2305.06147, (2023).

[20] Z. Chen, H. Lin, Improving named entity correctness of abstractive summarization by generative negative sampling, Computer Speech & Language, 81 (2023) 101504.

[21] T. Vo, An approach of syntactical text graph representation learning for extractive summarization, International Journal of Intelligent Robotics and Applications, 7 (2023) 190-204.

[22] S. Rai, R.C. Belwal, A. Sharma, Investigating the Application of Multi-lingual Transformer in Graph-Based Extractive Text Summarization for Hindi Text, International Conference on Data Management, Analytics & Innovation, Springer, 2023, pp. 393-403.

[23] L. Alex, H. Sakib, W. Lingfei, M. Collin, Improved code summarization via a graph neural network. In 2020 IEEE, ACM International Conference on Program Comprehension, 2020.

[24] S.-h. Zhong, Y. Liu, B. Li, J. Long, Query-oriented unsupervised multi-document summarization via deep learning model, Expert systems with applications, 42 (2015) 8146-8155.

[25] M. Yousefi-Azar, L. Hamey, Text summarization using unsupervised deep learning, Expert Systems with Applications, 68 (2017) 93-105.

[26] Z. Cao, F. Wei, L. Dong, S. Li, M. Zhou, Ranking with recursive neural networks and its application to multi-document summarization, Proceedings of the AAAI conference on artificial intelligence, 2015.

[27] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, S. Mohamed, Variational approaches for auto-encoding generative adversarial networks, arXiv preprint arXiv:1706.04987, (2017).

[28] A. Abdi, S. Hasan, S.M. Shamsuddin, N. Idris, J. Piran, A hybrid deep learning architecture for opinion-oriented multi-document summarization based on multi-feature fusion, Knowledge-Based Systems, 213 (2021) 106658.

[29] H. Zareiamand, A. Darroudi, I. Mohammadi, S.V. Moravvej, S. Danaei, R. Alizadehsani, Cardiac Magnetic Resonance Imaging (CMRI) Applications in Patients with Chest Pain in the Emergency Department: A Narrative Review, Diagnostics, 13 (2023) 2667.

[30] B.T. Hammad, A.M. Sagheer, I.T. Ahmed, N. Jamil, A comparative review on symmetric and asymmetric DNA-based cryptography, Bulletin of Electrical Engineering and Informatics, 9 (2020) 2484-2491.

[31] D. Hin, A. Kan, H. Chen, M.A. Babar, LineVD: statement-level vulnerability detection using graph neural networks, Proceedings of the 19th International Conference on Mining Software Repositories, 2022, pp. 596-607.

[32] R. Nallapati, F. Zhai, B. Zhou, Summarunner: A recurrent neural network based sequence model for extractive summarization of documents, Proceedings of the AAAI conference on artificial intelligence, 2017.

[33] H. Kobayashi, M. Noguchi, T. Yatsuka, Summarization based on embedding distributions, Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 1984-1989.

[34] L. Chen, M. Le Nguyen, Sentence selective neural extractive summarization with reinforcement learning, 2019 11th International Conference on Knowledge and Systems Engineering (KSE), IEEE, 2019, pp. 1-5.

[35] S. Danaei, A. Bostani, S.V. Moravvej, F. Mohammadi, R. Alizadehsani, A. Shoeibi, H. Alinejad-Rokny, S. Nahavandi, Myocarditis Diagnosis: A Method using Mutual Learning-Based ABC and Reinforcement Learning, 2022 IEEE 22nd International Symposium on Computational Intelligence and Informatics and 8th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics (CINTI-MACRo), IEEE, 2022, pp. 000265-000270.

[36] M. Kågebäck, O. Mogren, N. Tahmasebi, D. Dubhashi, Extractive summarization using continuous vector space models, Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC), 2014, pp. 31-39.

[37] W. Yin, Y. Pei, Optimizing sentence modeling and selection for document summarization, Twenty-fourth international joint conference on artificial intelligence, 2015.

[38] F. Koto, J.H. Lau, T. Baldwin, Discourse probing of pretrained language models, arXiv preprint arXiv:2104.05882, (2021).

[39] S. Abdel-Salam, A. Rafea, Performance study on extractive text summarization using BERT models, Information, 13 (2022) 67.

[40] A. Srikanth, A.S. Umasankar, S. Thanu, S.J. Nirmala, Extractive text summarization using dynamic clustering and co-reference on BERT, 2020 5th International Conference on Computing, Communication and Security (ICCCS), IEEE, 2020, pp. 1-5.

[41] C. Ma, T. Wang, L. Zhang, Z. Cao, Y. Huang, X. Ding, Distributed Representation Learning with Skip-Gram Model for Trained Random Forests, Neurocomputing, (2023) 126434.

[42] S. Aburass, O. Dorgham, J.A. Shaqsi, A Hybrid Machine Learning Model for Classifying Gene Mutations in Cancer using LSTM, BiLSTM, CNN, GRU, and GloVe, arXiv preprint arXiv:2307.14361, (2023).

[43] S.V. Moravvej, S.J. Mousavirad, M.H. Moghadam, M. Saadatmand, An LSTM-based plagiarism detection via attention mechanism and a population-based approach for pre-training parameters with imbalanced classes, Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part III 28, Springer, 2021, pp. 690-701.

[44] S.V. Moravvej, S.J. Mousavirad, D. Oliva, G. Schaefer, Z. Sobhaninia, An improved de algorithm to optimise the learning process of a bert-based plagiarism detection model, 2022 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2022, pp. 1-7.

[45] S.V. Moravvej, S.J. Mousavirad, D. Oliva, F. Mohammadi, A Novel Plagiarism Detection Approach Combining BERT-based Word Embedding, Attention-based LSTMs and an Improved Differential Evolution Algorithm, arXiv preprint arXiv:2305.02374, (2023).

[46] M.S. Sartakhti, M.J.M. Kahaki, S.V. Moravvej, M. javadi Joortani, A. Bagheri, Persian language model based on BiLSTM model on COVID-19 corpus, 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA), IEEE, 2021, pp. 1-5.

[47] L. Hong, M.H. Modirrousta, M. Hossein Nasirpour, M. Mirshekari Chargari, F. Mohammadi, S.V. Moravvej, L. Rezvanishad, M. Rezvanishad, I. Bakhshayeshi, R. Alizadehsani, GAN‐LSTM‐3D: An efficient method for lung tumour 3D reconstruction enhanced by attention‐based LSTM, CAAI Transactions on Intelligence Technology, (2023).

[48] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation, 9 (1997) 1735-1780.

[49] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078, (2014).

[50] F.A. Gers, N.N. Schraudolph, J. Schmidhuber, Learning precise timing with LSTM recurrent networks, Journal of machine learning research, 3 (2002) 115-143.

[51] A. Graves, Generating sequences with recurrent neural networks, arXiv preprint arXiv:1308.0850, (2013).

[52] W. Zaremba, I. Sutskever, O. Vinyals, Recurrent neural network regularization, arXiv preprint arXiv:1409.2329, (2014).

[53] S.V. Moravvej, M.J.M. Kahaki, M.S. Sartakhti, A. Mirzaei, A method based on attention mechanism using bidirectional long-short term memory (BLSTM) for question answering, 2021 29th Iranian Conference on Electrical Engineering (ICEE), IEEE, 2021, pp. 460-464.

[54] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems, 27 (2014).

[55] A.R. Zehan, Web Interface for Rouge Automatic Summary Evaluator, Invited Lectures, 57.

[56] M. Moradi, Frequent itemsets as meaningful events in graphs for summarizing biomedical texts, 2018 8th International Conference on Computer and Knowledge Engineering (ICCKE), IEEE, 2018, pp. 135-140.

[57] R. Mihalcea, P. Tarau, Textrank: Bringing order into text, Proceedings of the 2004 conference on empirical methods in natural language processing, 2004, pp. 404-411.

[58] Y. Wu, B. Hu, Learning to extract coherent summary via deep reinforcement learning, Proceedings of the AAAI conference on artificial intelligence, 2018.

[59] J.-Á. González, E. Segarra, F. García-Granada, E. Sanchis, L.ı.-F. Hurtado, Siamese hierarchical attention networks for extractive summarization, Journal of Intelligent & Fuzzy Systems, 36 (2019) 4599-4607.

[60] K. Al-Sabahi, Z. Zuping, M. Nadher, A hierarchical structured self-attentive model for extractive document summarization (HSSAS), IEEE Access, 6 (2018) 24205-24212.

[61] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research, 21 (2020) 5485-5551.

[62] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461, (2019).

[63] A. Joshi, E. Fidalgo, E. Alegre, L. Fernández-Robles, DeepSumm: Exploiting topic models and sequence to sequence networks for extractive text summarization, Expert Systems with Applications, 211 (2023) 118442.

[64] S. Gong, Z. Zhu, J. Qi, W. Wu, C. Tong, SeburSum: a novel set-based summary ranking strategy for summary-level extractive summarization, The Journal of Supercomputing, (2023) 1-29.

[65] A.P. Widyassari, S. Rustad, G.F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, Review of automatic text summarization techniques & methods, Journal of King Saud University-Computer and Information Sciences, 34 (2022) 1029-1046.

[66] S.V. Moravvej, A. Mirzaei, M. Safayani, Biomedical text summarization using conditional generative adversarial network (CGAN), arXiv preprint arXiv:2110.11870, (2021).

[67] S. Moravvej, M. Maleki Kahaki, M. Salimi Sartakhti, M. Joodaki, Efficient GAN-based method for extractive summarization, Journal of Electrical and Computer Engineering Innovations (JECEI), 10 (2022) 287-298.

[68] S. Bourou, A. El Saer, T.-H. Velivassaki, A. Voulkidis, T. Zahariadis, A review of tabular data synthesis using GANs on an IDS dataset, Information, 12 (2021) 375.

[69] Z. Li, M. Usman, R. Tao, P. Xia, C. Wang, H. Chen, B. Li, A systematic survey of regularization and normalization in GANs, ACM Computing Surveys, 55 (2023) 1-37.

[70] Z. Pan, W. Yu, B. Wang, H. Xie, V.S. Sheng, J. Lei, S. Kwong, Loss functions of generative adversarial networks (GANs): Opportunities and challenges, IEEE Transactions on Emerging Topics in Computational Intelligence, 4 (2020) 500-522.

[71] I. Triguero, S. García, F. Herrera, Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study, Knowledge and Information systems, 42 (2015) 245-284.

[72] A.A. Syed, F.L. Gaol, A. Boediman, T. Matsuo, W. Budiharto, A Survey of Abstractive Text Summarization Utilising Pretrained Language Models, Asian Conference on Intelligent Information and Database Systems, Springer, 2022, pp. 532-544.

[73] M.Q. Pham, B. Oudompheng, J.I. Mars, B. Nicolas, A Noise-Robust Method with Smoothed ℓ1/ℓ2 Regularization for Sparse Moving-Source Mapping, Signal Processing, 135 (2017) 96-106.

Security in Software-Defined Networks Against Denial-of-Service Attacks Based on Increased Load Balancing Efficiency

Ying ZHANG, Hongwei DING*
Hebei Software Institute, Hebei, Baoding 071000, China

***Abstract*—The goal of software-oriented networks (SDNs), which enable centralized control by separating the control layer from the data layer, is to increase manageability and network compatibility. However, this form of network is vulnerable to the control layer going down in the face of a denial-of-service assault because of the centralized control policy. The considerable increase in events brought on by the introduction of fresh currents into the network puts a lot of strain on the control surface when the system is in reaction mode. Additionally, the existence of recurring events that seriously impair the control surface's ability to function, such as the gathering of statistical data from the entire network, might have a negative impact. This article introduces a new approach that uses a control box comprising a coordinating controller, a main controller that establishes the flow rules, and one or more sub-controllers that establish the rules to fend off the attack and avoid network paralysis. It makes use of current (when needed). The controllers who currently set the regulations are relieved of some work by giving the coordinating controller management and supervision responsibilities. Additionally, the coordinator controller distributes the load at the control level by splitting up incoming traffic among the controllers of the flow rules. Thus, a proposed method can avoid performance disruption of the flow rule setter's main controller and withstand denial-of-service attacks by distributing the traffic load brought on by the denial-of-service attack to one or more sub-controllers of the flow rule setter. The results of the experiments conducted indicate that, when compared to the existing solutions, the proposed solution performs better in the face of a denial-of-service assault.**

Keywords*—Security; open balance; denial-of-service attacks; software-oriented networks

I. INTRODUCTION

The next-generation network approach, known as a software-oriented network (SDN), allows for programmable control of the network and makes network management easier by separating network transmission from control operations [1]. The Open_ Flow protocol has been utilized the most out of all the tools that are currently available to actualize the software-oriented network [2]. The flow rules that the controller installs in the flow tables of the switches direct the network traffic in an Open_ Flow network. The controller can generate flow rules using either the proactive method or the reactive way, respectively [3]. In the pre-active method, before launching the software-based network, the controller installs the flow rules in the switches based on the predetermined strategy [4]. No special flow rules are put in the switches beforehand when using the reaction technique [5]. A table loss event occurs for each new flow that does not match the installed flow rules [6]. In response, an Open_ Flow request is delivered to the controller, and the controller chooses a new flow rule based on this request. The reactive technique, which is flexible, is typically employed in software-based networks [7].

Even though Open_ Flow has many benefits for streamlining network management and expanding its adoption, its reactive approach to installing flow rules makes it simple for DoS attacks on the controller to succeed because the controller must deal with all the packets generated by the absence of the table in [8]. The Open_ Flow network may receive a high number of transient phony flows that were created by a hacked host. Many flow requests (packet_in packets) are made to the controller when the Open_ Flow switch receives these malicious flows because they cause events linked to the loss of the table. Consequently, these high-frequency current requests deplete the controller's resources, interfering with regular functioning [9].

A denial-of-service attack on the Open_ Flow network often has the following effects: a) overloading of switches; b) congestion in the data plane and control plane communication channel; c) overloading of the controller; and d) overflowing of switch flow tables [10]. While the anomaly connected to the controller's service might cause disruption and failure of the entire network, overloading the switches and overflowing the switch flow tables only endanger the victim switch. As a result, the majority of threats from a denial-of-service attack in the Open_ Flow network are brought on by the emergence of congestion in the communication channel of the data level with the control level and overloading of the controller [11].

The suggested approach for the control level of the software-based network has been utilized in this article to strengthen the security of the network against denial-of-service attacks and boost its availability [12]. In order to increase the availability of the software-based network against significant changes in the events brought on by the arrival of new flows (due to the temporal and spatial characteristics of the network traffic) and the existence of repeated events, the proposed method consists of a control box [13]. It is intended to gather statistical data from the entire network, which overburdens the controller. A coordinating controller, a primary controller of the current setter, and one or more (as required) sub-controllers of the current setter are all included in this control box [14].

The current installation controllers' job is essentially decreased by giving the coordinating controller responsibility for managing and monitoring the software-based network. It is the responsibility of flow controllers to install flow rules in line with the network applications that are operating on them in order to configure the data level of the software-oriented network [15].

Additionally, the coordinator controller will categorize the incoming traffic in a statistical manner after activating one or more (as necessary) sub-controllers in response to the network's increased traffic load, which can be brought on by a denial-of-service assault [16]. Each of them receives a portion of the incoming traffic that the flow controller divides into known categories. This results in the flow controller's controllers sharing the load of incoming traffic, which lessens the effort of all of the flow controller's controllers. In the suggested method, the redundancy of the controllers and the division of labor among them are used to distribute the traffic caused by the denial-of-service attack to one or more sub-controllers, taking into consideration the multi-controller capability offered in Open_ Flow Specification 1.2. The software-based network controller will experience less of an impact from the attack thanks to the flow installer, and the software-based network will be more resilient to unexpected and severe fluctuations in network traffic. In summary, the article presents a solution to improve the security and resilience of SDNs against DoS attacks. The proposed approach, utilizing a control box with coordinating and sub-controllers, is shown to be effective in managing traffic during attacks and reducing response times. However, the article acknowledges the need to address resource consumption and scalability in future work.

In short, the contribution of the authors in this research is as follows:

- Introducing a new approach: A new approach has been introduced to increase the security and resilience of Software Defined Networks (SDN) against Denial of Service (DoS) attacks. It lies in the creation of a "control box" consisting of coordinating, main and sub-controllers.
- They also contribute to this field by comparing the proposed solution with four other existing methods for mitigating DoS attacks in SDN, including Ryu controllers with different mechanisms. Through this comparison, they show that their approach consistently outperforms these alternatives and provides a more effective way to counter DoS attacks.

The article's structure is described in the paragraphs that follow. The work that has been done to fortify the software-based network's control level against denial-of-service assaults has been mentioned in Section II. The software-based network controller's input load has been looked into in Section III. Section IV describes the suggested technique for enhancing the security of the software-based network control level. In Sections IV and V, it was examined, respectively, how denial-of-service assaults would affect the proposed approach and how a saturation attack on the control surface by the data surface would affect it. The effectiveness of the suggested approach against a denial-of-service attack is contrasted with that of many other solutions in Section VI, and conclusions and recommendations for further work are provided in Section VII and Section VIII respectively.

II. Related Works

Studies on data-level protection and control-level protection that aim to lessen the effects of denial-of-service attacks on software-based networks fall into these two groups. In order to fight against assault, data layer protection focuses on enhancing data layer functionality or adding new features [17]. In order to keep the controller from becoming overloaded with requests, OF-Guard uses a data-level cache. The strategy, however, lacks flexibility because it is uncertain whether such a cache will be established at the data level. The implementation of SYN Proxy is proposed in AVANT-GUARD [18] as a unit that performs the TCP Handshake in the switches before sending the incoming TCP stream to the network. The pressure on the switch buffer is increased using this unit, and there is also a cap on the number of proxy ports. Because LineSwitch can arbitrarily proxy flows from the same IP source that it has already established a TCP handshake with, it is suggested that LineSwitch will enhance AVANT-GUARD [19]. Another AVANT-GUARD-based solution is SDN-Shield, which employs a number of NFV-based attack mitigation units to counter distributed denial-of-service attacks at the software-oriented network data level. However, none of these SYN Proxy-based techniques work with other network protocols.

In order to sustain network policy enforcement, Flood-Guard offers a pre-active flow rule analyzer that can examine the source code of an Open_Flow-based application and generate several pre-active rules via running time monitoring of each application's global state. To produce based on the protocol, Flood-Guard stores the remaining packets caused by table loss in a cache at the data level. The source code analyzer, however, is extremely complicated and unable to be deployed across a network. Additionally, the same attack flow protocol was utilized by additional benign flows, so the data level cache cannot ensure fair behavior against them.

Enhancing the security of the controller policies and implementing detection and filter techniques to lessen denial-of-service attacks are the main goals of control level protection [20]. In Flow Ranger, the controller employs a rating algorithm to recognize typical users and pre-process packets with a higher priority. In order to distinguish malicious communication, SGuard introduces an access control system that leverages six items as feature vectors in the classification unit [21]. Researchers in [22] have suggested a scheduling policy for the proposal controller that divides the flows into various queues. The controller manages each queue based on round-robin scheduling, and each queue corresponds to a switch that is situated on the denial-of-service attack path. MLFQ creates an equitable sharing of control resources between the server and hosts in the network by utilizing a number of expandable and collapsible request queues [23]. Potentially hostile traffic is diverted to the intrusion detection system by SDN-Shield, and the impact of a denial-of-service attack is lessened by appropriately building the flow rules.

In order to lessen channel congestion between the victim switch and the controller, Flood Defender suggests a technique for distributing packets related to table loss from the victim switch to its surrounding switches. Additionally, for the purpose of detecting sporadic attacks, Flood Defender employs a two-stage filter and briefly archives packet_ in packets. To handle new benign flows, the packet_ in packet buffer introduces a significant delay. Additionally, the surrounding switches that are being flooded may suffer damage from the packet distribution caused by the absence of the table. In order to counter the denial-of-service attack, FMD employs a technique based on flow migration at the software-oriented network control level slow. The master controller, to which fraudulent requests with a high volume are delivered, is replaced in this technique by a slave controller. The master controller handles all typical Open_ Flow requests. In order to safeguard the channel of communication between the victim switch and the master controller, FMD sends the threatening Open_ Flow requests to the slave controller after identifying a denial-of-service attack. The migrated requests are briefly kept in the slave controller and transferred to the master controller for additional processing at a limited rate [24] to prevent the master controller from becoming overloaded. SGS is offered to defend the control plane from denial-of-service assaults, and its key component is the clustered deployment of several controllers in the control plane. The abnormal traffic detection and controller dynamic defense units make up the SGS procedures. Anomalous traffic detection uses quadratic feature vectors to separate phony flows from real ones by focusing on switches that already exist at the data level. The impact of the denial-of-service assault at the control level is minimized by the controller's dynamic protection, which involves remapping the controller and sending the access control message to the switch. As it was already indicated, the majority of the techniques for data layer protection include modifying Open_ Flow switches or incorporating unique features into the data layer. However, all techniques in the second category (control level protection) work to reduce the number of resources that a denial-of-service assault uses up. In reality, the controller cannot manage the rate of receiving Open_ Flow packets. Additionally, the main cause of the denial-of-service attack in the control level is the congestion of the communication channel connecting the data level with the control level, which is of the TCP or TLS type. Source in [42] has demonstrated the capacity to enhance networks and withstand DDoS attacks. Thus, the purpose of this survey is to review 65 articles about DDoS attack detection in SDN. As a result, each reviewed article's systematic reviews of the suggested methodology are examined. This work additionally analyzes the performance metrics and their best assault detection accomplishments from each research publication. Furthermore, this work reviews the reduction technique applied in each paper.

In this study, an architecture to handle service source attacks is developed in order to address the restrictions indicated above. A suggestion is to have flow rule setter sub-controllers that are multiples of the main flow rule setter controller and assign incoming traffic from denial-of-service attacks to them regardless of the protocol used, as well as from the channel connecting the data level with the level. Control is comprehensive since it guards against congestion and shields the main flow rule controller from denial-of-service attacks to minimize their negative effects on good new flows. Additionally, the concept is transparent because it does not require any adjustments to the network's infrastructure or application programs to be used.

III. Examining the Software-based Network Controller's Input Load

As depicted in Fig. 1, within a software-oriented network made up of Open_Flow switches, the Open_Flow controller periodically sends multi-part messages [25] to the switches under its control. These messages serve to collect statistical data from the switches, contributing to the construction of the network's information base (NIB). The controller can have a comprehensive view of the entire network under his control based on the information in this network's information base. The initial packet of each flow that reaches the input switch is examined to determine whether it complies with the rules set up in the switch's flow table. Suppose the switch's flow table does not contain a match for the packet. In that case, a packet-in message is provided to the controller containing the lowest priority flow entry that is compatible with any incoming flow (flow entry absent from the table). The request in this message is to begin planning a path for the incoming stream. The Open_Flow controller determines the appropriate route for the incoming flow based on information from the network information base. The route is established by installing the appropriate flow rules in the flow tables of the switches along the route, and as a result, all flow packets are directed to their intended destination. Eq. (1) determines the input load to the controller resulting from the input currents with the input speed of the current per second in a data center with N switches, assuming the average number of switches located in the path is [N/2] and for each flow in the flow table of switches located in the path of two flow rules (to forward in two directions) installed.

$$L_{arrival-flows} = \lambda \times (\lceil m_{pi} \rceil + \lceil m_{po} \rceil + \left\lceil \frac{N}{2} \right\rceil \times 2 \times \lceil m_{fm} \rceil) \quad (1)$$

where, m_{fm} stands for the size of the flow_mod packet, which is the packet delivered from the controller to the Open_Flow switches to install a new flow rule. Additionally, m_{pi} and m_{po} stand for the sizes of the packets in and out, respectively. As a result, there is a chance that the controller will become overloaded by increasing the speed at which fresh currents enter the input switch in the reaction mechanism chosen for each incoming current to be passed into the controller.

The input load brought on by gathering statistical data at the network level is another element that overburdens the software-based network controller. Important statistical data that can be gathered from existing Open_ Flow switches in a data center includes data about the ports connected to servers and data about the flows set up in the switches' flow tables. Eq. (2) can be used to determine the input load to the control level in a data center as a result of gathering statistical data from Open_ Flow switches.

$$L_{statistics} = N \times (([m_{portreq}] + H \times [m_{portrep}] + [m_{flowreq}] + F \times [m_{portrep}])) \times [\frac{1}{T}] \quad (2)$$

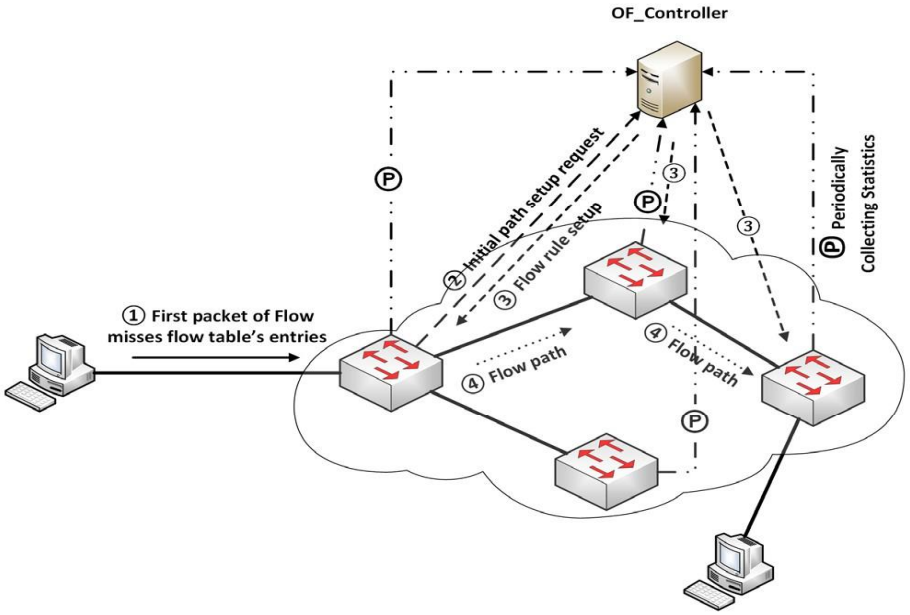


Fig. 1. The software-oriented network's controller operation mechanism creates a path for a new flow.

TABLE I. THE VARIABLES THAT ARE USED TO CREATE THE INCOMING TRAFFIC TO THE DATA CENTER

Name	Model	Parameters
New flow arrivals	Poisson	$\lambda(t) = 192$
Internal rack flow ratio	Bernoulli	$R_{int} = 0.8$
TCP flow ratio	Bernoulli	$R_{tcp} = 0.85$
Flow duration, Dn	Pareto	$a_p = 1.504$ $M_p = 1.0001$
Flow transmission rate, Yn	Gaussian	$E[Y_n] = 4303.69$ $Var[Y_n] = 69936.37$
Flow size, Sn	$Y_n \times D_n$	$E[S_n] = 8517.97$

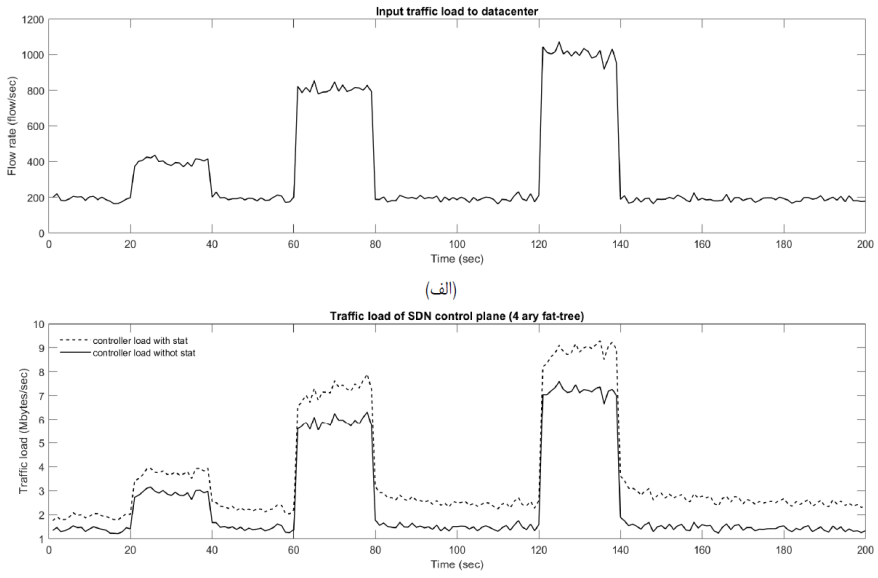


Fig. 2. (a) Incoming traffic to the data center and (b) Incoming load to the controller in a data center with _Ary_Fat_tree4 structure ((-- -) with the load resulting from the collection of statistical information and (-) without it.

where, H is the total number of switch ports linked to servers, and N is the total number of Open_ Flow switches in the data center. F is the number of installed currents in the switches, and T is the controller's periodic request for statistics data. A simulation has been run in MATLAB 2022b to show the impact of the load brought on by the gathering of statistical data (Relation 2) and the admission of fresh flows into the data center (Relation 1) on the software-based network controller.

The data center in this simulation has an Ary Fat_tree 4 structure. The data center's incoming traffic is generated based on the findings of Yoonseon Han and his associates [26], who used the values in Table I. Fig. 2 displays the simulation outcome based on (1) and (2), as well as the relationships offered by Bong-yeol Yu and his associates [27]. As seen in Fig. 2(a), the data center receives 192 flows per second of incoming traffic in the usual state. This amount is in the intervals to observe the impact of speeding up this flow rate on the amount of incoming load to the software-based network controller. There has been an increase in the current per second from 20 to 40 seconds, 60 to 80 seconds, and 120 to 140 seconds, respectively. The input load to the software-based network controller of a data center with an _Ary_Fat_tree4 structure is shown in Fig. 2(b) in two states: (-) with the load brought on by gathering statistical data and (-) without it. As might be predicted, as the pace of incoming traffic to the data center grows, so does the input load on the controller. The important feature in Fig. 2's lower portion is the sizeable input load brought on by the controller's statistical data collection, which has a significant impact on the occurrence of interference with the transmission of basic controller messages (such as current installation), delays the arrival of information to the controller, and ultimately lowers the controller's efficiency. Therefore, packet_in can improve the controller's efficiency and availability by separating the load arising from the gathering of statistical data from the load resulting from the arrival of packages.

IV. Suggestion for Enhancing the Security of SDN Control Levels

In software-based networks, the controller is not just in charge of performing the functions of a straightforward switch to transfer flows throughout the network in order to take advantage of its central management. It is required to gather statistical data from the network level in order to implement efficient programs, such as balancing the load on communication lines and servers in the network or quality control of service provided to flows system by system. For successful and integrated management, software-oriented networks need the controller to be highly available.

There are two ways to provide high availability for Open_ Flow controllers in software-based networks [28]. Reduce the load on controllers as the initial step in the approach. The Open_ Flow controller communicates with the Open_ Flow switches frequently, particularly while running in reactive mode, as was covered in the preceding section. The controller has become overwhelmed. As a result, making it is unable to process incoming messages. The second option is to replace one controller with many controllers to add redundancy.

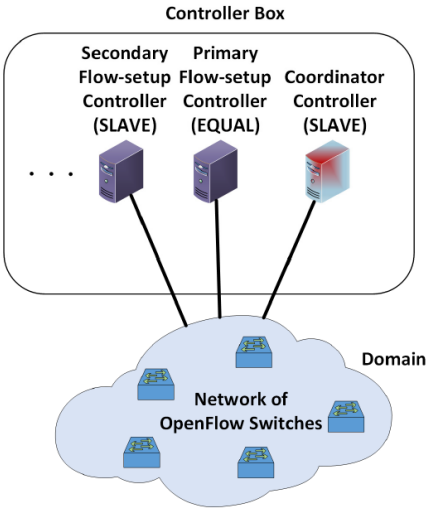


Fig. 3. The recommended approach for software-based network controllers' structure.

Both of the aforementioned methods are taken into account in the proposed control box, which is employed in the proposed method to increase the availability of Open_ Flow controllers in software-based networks (see Fig. 3). One of the controllers in this control box serves as a coordinator, connecting to Open_ Flow switches only to gather statistical data from them. Additionally, there is a main controller and one or more (depending on the demand) sub-controllers in this control box that are in charge of adding flow rules to the flow tables of Open_ Flow switches. So, in addition to lessening the strain on the controller, more controllers have been deployed by splitting the work between the coordinating controller and the controllers that establish the flow rules. It is necessary to employ several controllers to control Open_ Flow switches, which are accessible in Open_ Flow version 1.2 and later, in order to implement the control box concept in software-based networks. Since Open_ Flow version 1.2, it has been possible to have numerous controllers, each of which can control Open_ Flow switches in one of the three modes of SLAVE, MASTER, or EQUAL. The MASTER and EQUAL modes have complete access to the switch and can receive all asynchronous communications (such as packet-in) from the switch, among the other two modes [29].

Each Open_ Flow switch is permitted to communicate with one MASTER controller but numerous EQUAL controllers. In SLAVE mode, the controller can only read from the switches through its access to them; it is, therefore, unable to receive any other asynchronous messages than the answer of the multi-part message containing the switch's status. Each controller can modify its state by sending the switches the OFPT_ROLE_REQUEST message [30]. The switch transmits the OFPT_ROLE_REPLY message to the controller after receiving this message.

The other controllers of that switch will convert to the SLAVE state if the switch gets a message from it instructing it to change the controller state to MASTER. Due to the switch's ability to have several Open_ Flow channels, it is not necessary to re-establish the channel in the event that one of the

controllers connected to it fails. Open_ Flow switches are able to transmit packet-in messages solely to the controller whose ID is in the matching table entry by giving each of the flow rule set controllers in the control box a distinct ID. Their direction of flow is known. Utilizing the Nicira features integrated within the RYU network operating system; this capability is possible [31]. The NXTSetControllerId function is used to accomplish this by giving each flow rule controller a special identification number. The number of flow rule set controllers that can be used in the control box will thus no longer be constrained. As a result, the control box is expandable, allowing the number of active controllers that establish the flow rules to grow as necessary. This boosts the availability of the control level of the software-oriented network by dividing the load associated with packet-in messages among the controllers of the flow rules and reducing their individual loads.

A. Coordinating Controller

All Open_ Flow switches in the domain are connected to the coordinator controller that is operating in SLAVE mode. As seen in Fig. 4, the coordinating controller has a variety of components for regulating the operation of the controllers that establish the flow rules and for keeping an eye on the network. Allocating incoming traffic to the flow controller's sub-controller is the primary responsibility of the coordinating controller. When the coordinating controller notices that the main controller is on the verge of overflowing, this is what happens. By doing this, the existing installation's main controller's traffic load is lessened, and it is once again able to function.

In order to prevent the overload of the controller controllers as the amount of incoming traffic to the network increases, the coordinating controller repeats this process by turning on new flow controller sub-controllers and allocating the overload of incoming traffic to them (see Fig. 5). On the other hand, by lowering the load of traffic entering the network, the coordinating controller transfers all of the traffic load entering the network to the main controller of the flow setter while deactivating the sub-controller of the flow setter. The description of the coordinating controller's role as well as information on each of its component parts, are provided below.

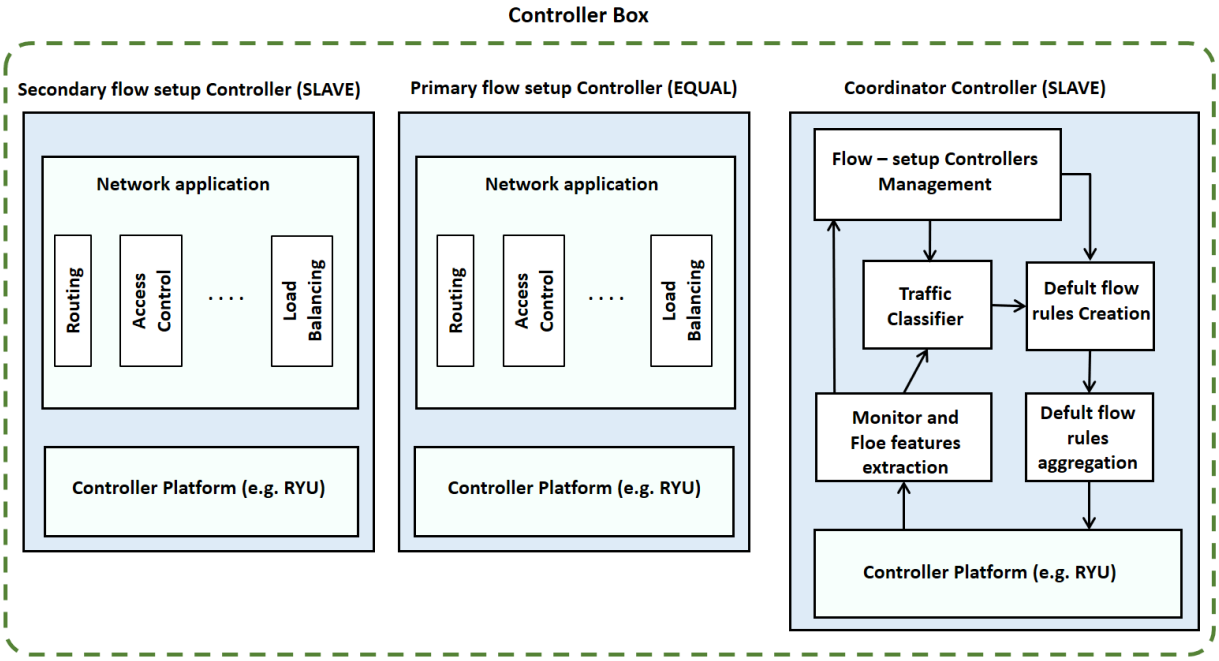


Fig. 4. Block diagram of the components of the proposed method.

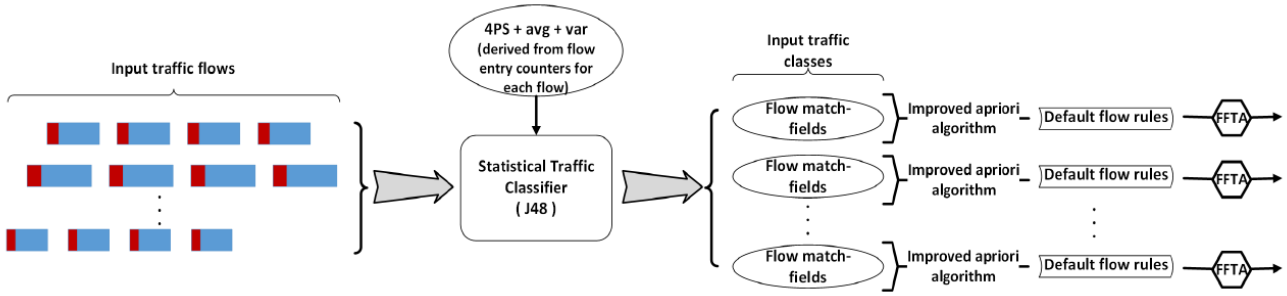


Fig. 5. The process of creating default flow rules and statistically classifying incoming network traffic.

B. How to Switch the Sub-controller On and Off

The main controller of the current installation may become overloaded as a result of changes in the rate at which traffic enters the network. The current setter's sub-controller is activated by the coordinating controller when it notices that the main controller is on the verge of overload as shown in Fig. 6(a). The flow controller's sub-controller receives the be-Active signal from the coordinating controller in order to accomplish this. The sub-controller of the flow installer sends the OFPT_ROLE_REQUEST message to all the Open_ Flow switches attached to it to alert them that its state has changed to EQUAL after receiving the be-Active signal. The sub-controller of the flow installer enters EQUAL mode after receiving the OFPT_ROLE_REPLY signal from the Open_ Flow switches. The coordinating controller changes its state from SLAVE to EQUAL upon receiving the activated signal from the flow controller's sub-controller, and after implementing the default flow rules related to one of the traffic categories entering the network, the output of the controller's classification algorithm. The gate switches activated current controller switches back to SLAVE mode.

The sub-controller of the current setter is deactivated once again by the coordinating controller when it notices that the main controller of the current setter is no longer in overload mode as shown in Fig. 6(b). The coordinating controller delivers the be-passive signal to the current setter sub-controller after switching the mode from SLAVE to EQUAL and removing the default flow rules connected to the active current setter sub-controller from the gateway switches. A remark is sent. The coordinating controller then reverses its state, going back from EQUAL to SLAVE. The flow installer's sub-controller sends Barrier request messages to all of the switches connected to it upon getting the be-passive signal, causing these switches to complete processing all of the messages they had been receiving from the flow installer's sub-controller. The flow controller sub-controller then sends the OFPT_ROLE_REQUEST message to all the switches connected to it, changing their status from EQUAL to SLAVE. The current regulator's sub-controller is rendered inactive as a result.

C. Investigating How the Proposal is Affected by a Control-Level Saturation Attack at the Data Level

A UDP flood attack was applied against three scenarios (the first scenario: using one controller, the second scenario: using a desaturation controller and a number of flow rule set controllers, and the third scenario: using the proposed architecture) in order to demonstrate the impact of redundancy on the availability of the proposed architecture. This experiment was carried out using the structure depicted in Fig. 7 and implemented in the Mininet 2.2.2 emulator environment [32]. According to Table II, this implementation uses OpenVSwitch [33] software switches for the Open_ Flow switches and the RYU platform [34] for the controllers. Additionally, the iPerf [35] tool has been used to produce UDP flood attack traffic in order to test the scenarios. The attacker launches a UDP flood assault on the Open_ Flow switch while a client is corresponding with the load balancer between servers at a consistent pace of 50 flows per second. In this experiment, the round-trip time (RTT) of client requests collected by running the ping command has been used to assess the performance of various scenarios. The round-trip time of client requests has been observed for all three situations, as seen in Fig. 8, despite the fact that the speed of the UDP flood attack rises linearly up to 800 packets per second (PPS). Fig. 8 shows that the proposal approaches saturation when the attack speed hits 800 packets per second, whereas for scenarios 1 and 2, this happened at 450 and 500 packets per second, respectively. In fact, the controllers of the flow rules are out of reach and are unable to install a new route on the Open_ Flow switch in exchange for the arrival of fresh flows once the attack speed surpasses the saturation point of each of the scenarios. The results show that the proposal has a high availability compared to alternative scenarios and is particularly resistant to the assault of saturating the control surface due to the usage of redundancy in the flow rule controllers.

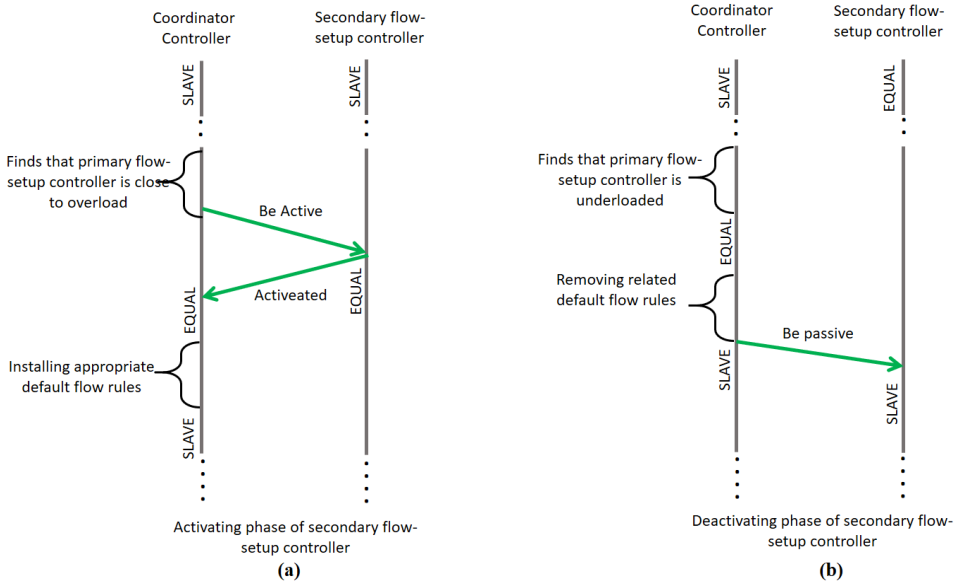


Fig. 6. How to modify the proposal's state of the existing regulator's sub-controller: a) activation and (b) deactivation.

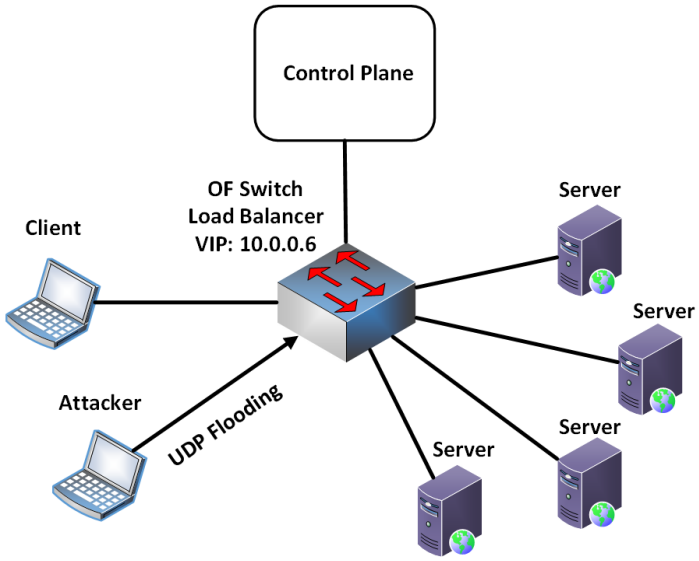


Fig. 7. A structure that was utilized for the UDP flood assault test.

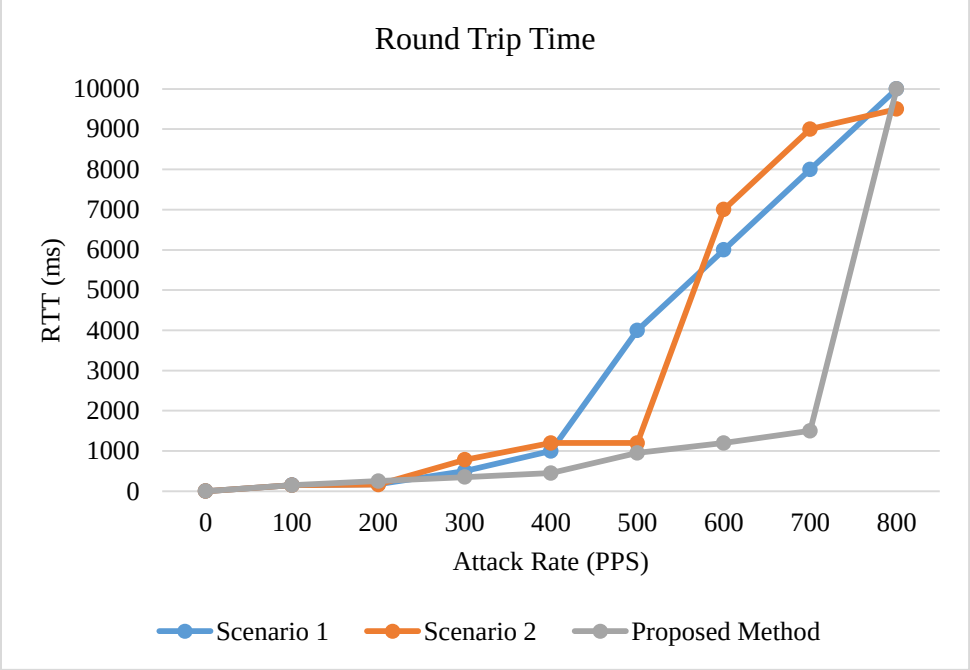


Fig. 8. Determining, during the flood attack, the turnaround time for client requests for proposals and scenarios 1 and 2. a) The first scenario: using one controller, (b) the second scenario: using a coordinating controller and a number of controllers that set the flow rules, and (c) the third scenario: using a proposed method.

TABLE II. LABORATORY PARAMETERS FOR THE PROPOSED EVALUATION

Virtual Machine	Oracle VM VirtualBox (Version 5.1.10 r 112026)
Guest OS	Ubuntu 16.04
CPU	Intel core i7-4720hq cpu@2.60ghz
RAM	16GB
Emulator	Mininet 2.2.2
Network Operating System	RTU 4.15
Open_Flow Switch	OpenVSwitch 2.4.0
Network Traffic Generator	iPerf 2.08b

V. Analyzing How Denial-of-Service Attacks Affect the Proposal

One of the elements affecting the availability of software-based networks is denial-of-service attacks. This section has researched how three different denial-of-service attacks affect the proposal. These attacks include ICMP/Ping flooding, UDP flooding, and TCP.SYN flooding. These attacks, which are a subset of volume denial-of-service attacks, were created using the Hyenae tool version 0.36 [36]. In this section, experiments were carried out utilizing the topology depicted in Fig. 7 that was constructed using the.222 Mininet simulator [37] and [38].

Additionally, the attacker conducts denial-of-service assaults on the Open_ Flow switch while a client is interacting with it (the switch is an Open_ Flow load balancer between servers) at a constant speed of 50 flows per second. The intended parameters in this test are the input load on the controllers in the proposed method, the round-trip time (RTT) of customer requests, and the impact of denial-of-service attacks on the proposed method's availability and performance. The ping command is used in conjunction with the 062bwm-ng v. utility to acquire these parameters. These attacks and the outcomes of their use have been addressed in the paragraphs that follow.

A. Study of the ICMP / Ping Flood Attack's Effects

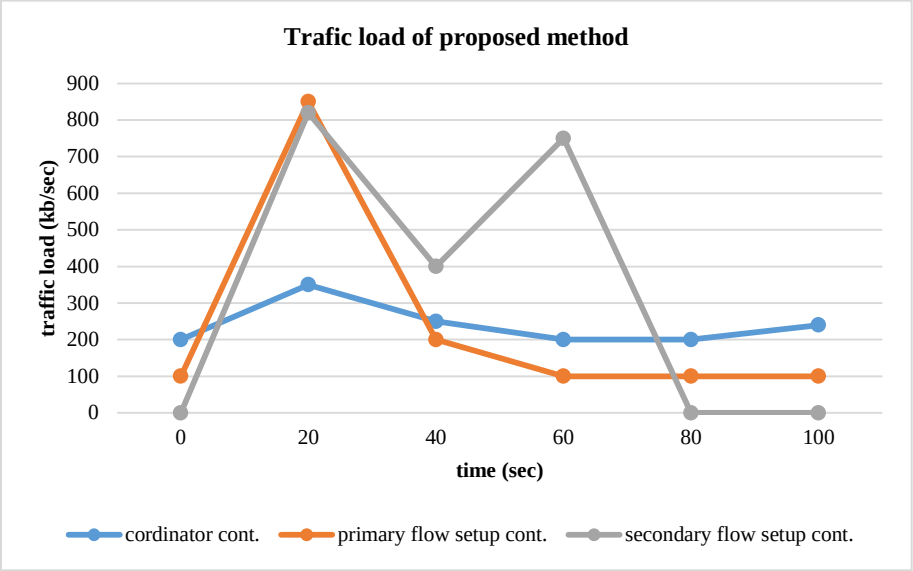
This kind of attack involves the attacker bombarding the target with ICMP echo requests, which interfere with other services that the target's other programs rely on. This kind of attack requires the attacker to rapidly transmit to the victim a large number of echo_request packets that appear to be coming from various sources and have random addresses. By having many flows, the attacker can interfere with the network. Each time a new packet-in is generated when an attacker attacks a victim using many sources with random addresses, the controller will experience a significant increase in traffic. The controller begins to sink when it receives a lot of packet-in packets, and after some time, it begins to disregard fresh incoming packets. The amount of production traffic described in terms of bytes, packets, or streams can be used to identify this kind of assault. Fig. 9 displays the outcomes of an ICMP/Ping flood assault conducted on the software-oriented network using the suggested technique, whose topology is depicted in Fig. 7. The volume of incoming traffic using the suggested method before, during, and after the attack is depicted in Fig. 9(a). Fig. 9(b) also depicts the duration of back-and-forth for client inquiries. The ICMP/Ping flood attack begins at 10 seconds and ends at 40 seconds, as can be shown. As soon as the attack begins, the time it takes for the customer's requests to move back and forth dramatically increases due to the main controller of the flow rules being overloaded and unable to handle requests entered as packet-in packages. The amount of incoming traffic load to the main controller of the flow rules setter is decreased by the activation of the sub-controller of the flow rules setter by the coordinating controller due to the detection of overloading of the main flow rules setter controller and the allocation of additional load resulting from the attack on it. The amount of incoming traffic to the main controller of the flow rules is decreasing, and as a result, this controller is once more able to handle and process requests that are received as packet-in packages. As a result, the amount of time required to go back and forth for customer requests returns to a reasonable value. This test demonstrates that the proposed method's availability is unaffected by an ICMP/Ping flood attack, save for a brief period of time (the time needed for the coordinating controller to notice that the main controller of the flow rules setter is overloaded before activating the sub-controller and adding additional load to it).

B. Investigating the Effect of UDP Flood Attack

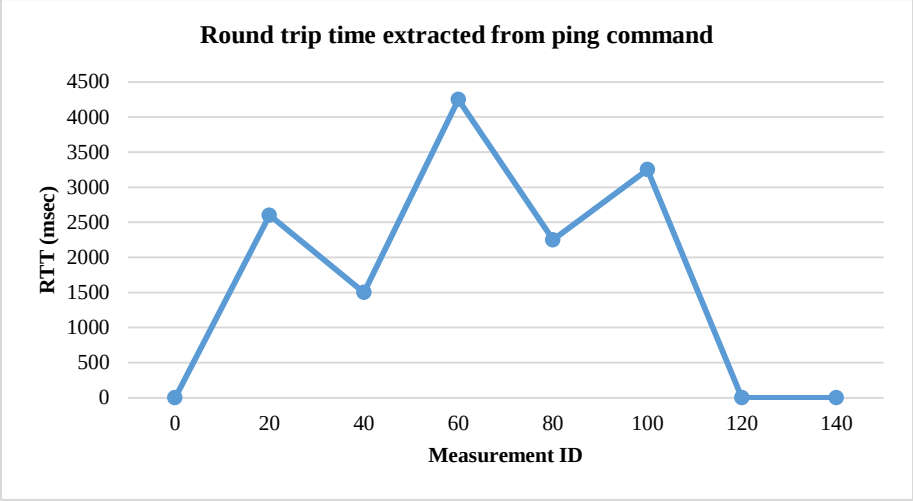
A UDP flood assault aims to bombard the target with a large volume of UDP packets. A host employs a huge number of bogus sources IP addresses when launching a UDP flood attack. The buffer overflow occurs in the victim as a result of a huge quantity of UDP packets arriving at the victim. Since sending a large number of bytes to the victim is the primary characteristic of UDP flood attacks, they produce the most traffic among denial-of-service attacks. In this attack, bogus source IP addresses are used to produce UDP packets that appear to send the target a lot of fresh streams per second. As a result, the software-based network controller starts pouring new flows after a given amount of time. The results of a UDP flood assault on a software-based network using the suggested technique, whose structure is depicted in Fig. 7, are displayed in Fig. 10. The volume of incoming traffic using the suggested method before, during, and after the attack is depicted in Fig. 10(a).

Additionally, Fig. 10(b) displays the amount of time spent responding to client inquiries. The UDP flood attack begins at 10 seconds and stops at 50 seconds, as can be shown. As soon as the attack begins, the time it takes for the customer's requests to move back and forth dramatically increases due to the main controller of the flow rules being overloaded and unable to handle requests entered as packet-in packages. Fig. 10(b) shows that this increase in time is caused by the software-based network controller receiving more incoming traffic than during the ICMP/Ping flood attack.

The amount of incoming traffic load to the main controller of the flow rules setter is decreased by the activation of the sub-controller of the flow rules setter by the coordinating controller due to the detection of overloading of the main flow rules setter controller and the allocation of additional load resulting from the attack on it. The main flow rule controller's ability to handle and process requests received in the form of packet-in packets is restored with the reduction of incoming traffic load, and as a result, the time required for the customer's requests to move back and forth to its reasonable value. Therefore, in this instance, similar to the ICMP/Ping flood attack, the UDP flood attack has no impact on the proposed method's availability other than for a brief time (the time needed for the coordinating controller to detect the main controller of the flow rules setter's load before activating the sub-controller and allocating additional load to it).

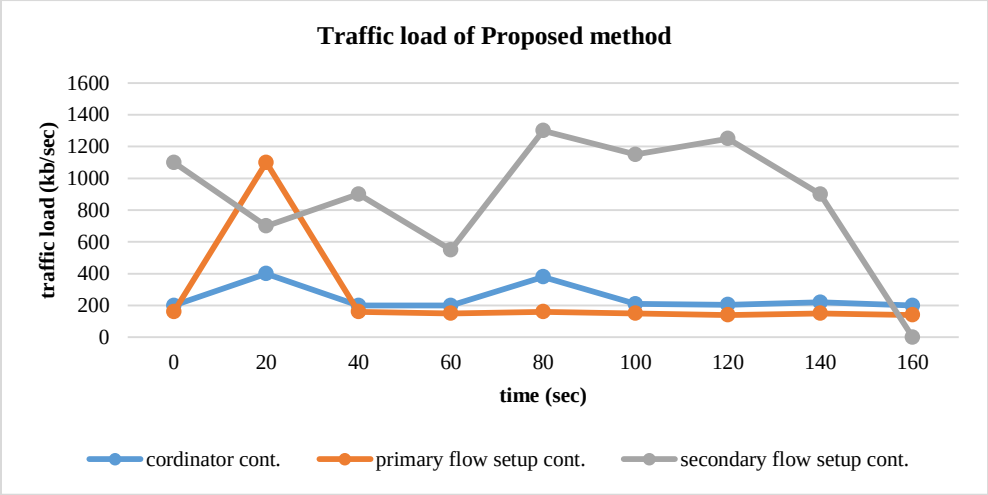


(a)

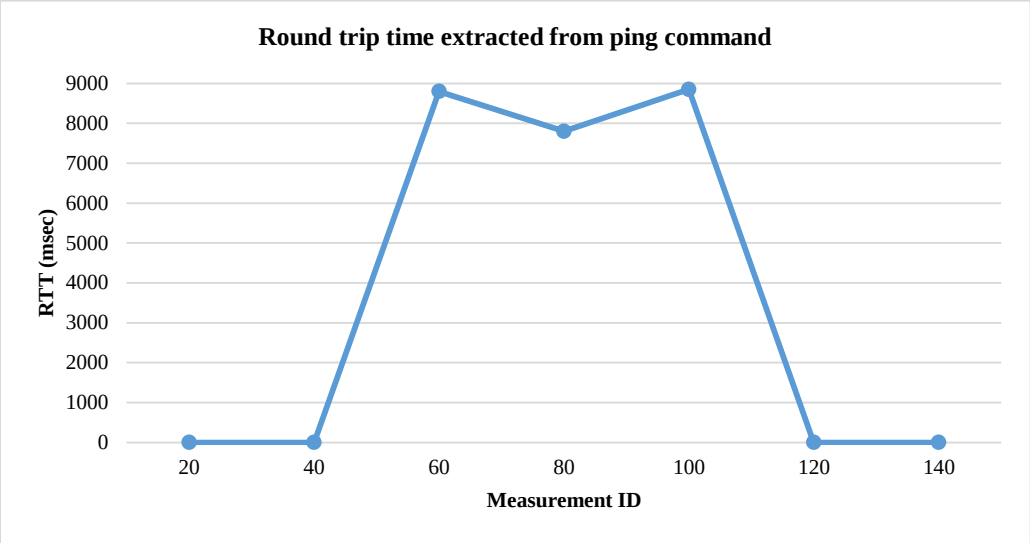


(b)

Fig. 9. (a) The amount of incoming traffic in the proposed method before the ICMP / Ping flood attack, during the attack and after it and (b) the round trip time (RTT) of customer requests.



(a)



(b)

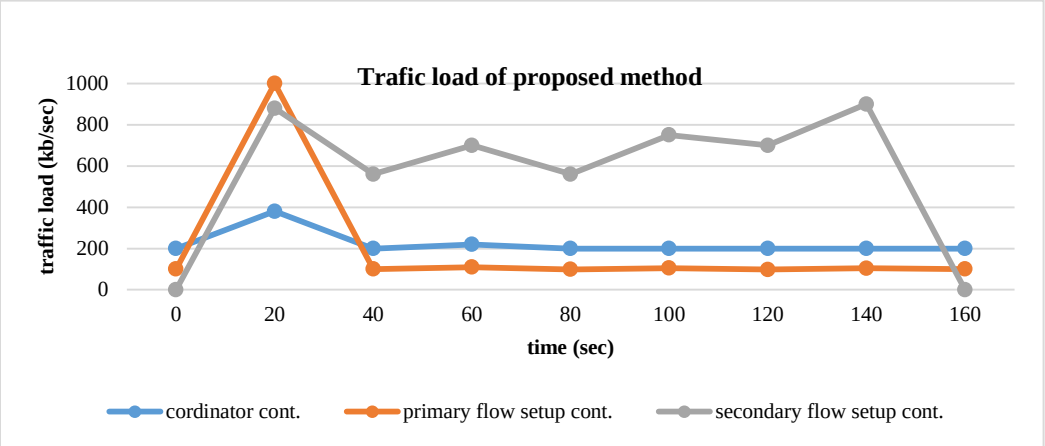
Fig. 10. (a) The amount of incoming traffic in the proposed method before the UDP flood attack, during the attack and after it and (b) the round trip time (RTT) of the client's requests.

C. Investigating the Effect of TCP_SYN Flood Attack

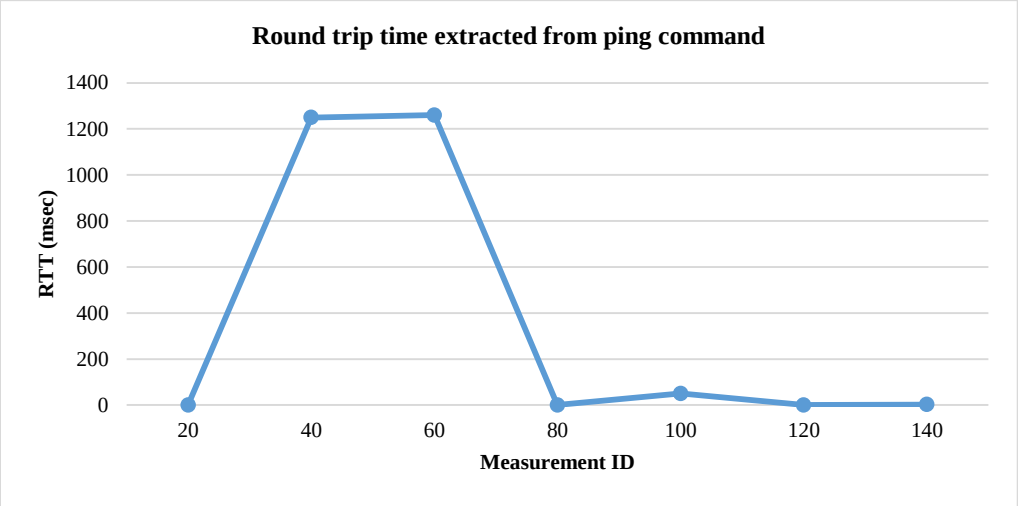
A TCP_SYN flood attack bombards the target with bogus SYN requests that were generated using fictitious source IP addresses. The victim never gets an answer to their SYN/ACK packets because the source IPs are faked. As a result, the attack's port is still open unnecessarily. All of the victim's ports are blocked as a result of several bogus SYN requests, making it impossible for the victim to connect to trustworthy people. It only takes a small amount of bandwidth for this kind of attack to keep the false connections open and render the victim unreachable. Hyenae tool version 0.36 was used [39] to develop a TCP_SYN flood attack, which begins by sending the victim a large number of low-speed streams containing TCP_SYN packets.

Additionally, the Open_Flow load balancing switch only distributes a load of incoming traffic across three servers to analyze the effect of the TCP_SYN flooding attack on the software-based network controller. From the fourth server, the round-trip time is measured. Customer requests are employed in (RTT). The software-based network controller can see a significant number of flows in this assault since random source IPs are being used to produce flow towards the target. This significantly hinders the efficiency of the software-based network controller by adding a lot of traffic to it.

The results of the TCP_SYN flooding assault on the software-based network using the suggested technique, whose structure is depicted in Fig. 7, are displayed in Fig. 11. The volume of incoming traffic using the suggested approach before, during, and after the attack is depicted in Fig. 11(a). Fig. 11(b) also depicts the duration of back-and-forth travel in response to customer demands. The TCP_SYN flood assault begins in three seconds and concludes in 45 seconds, as can be shown. As soon as the attack begins, the time it takes for the customer's requests to move back and forth dramatically increases due to the main controller of the flow rules being overloaded and unable to handle requests entered as packet-in packages. Due to the software-based network controller receiving far less incoming traffic than the UDP flood attack, the time increase seen in Fig. 11(b) is noticeably reduced. The amount of incoming traffic load to the main controller of the flow rules setter is decreased by the activation of the sub-controller of the flow rules setter by the coordinating controller due to the detection of overloading of the main flow rules setter controller and the allocation of additional load resulting from the attack on it. The main controller of the flow rules can once again handle and process requests that are received in the form of packet-in packets by reducing the amount of incoming traffic. As a result, the time it takes for the customer's requests to go back and forth returns to a reasonable value. As a result, in this instance, similar to ICMP/Ping and UDP flood attacks, the TCP_SYN flood attack has no effect on the proposed method's availability other than for a brief period of time (the time needed to detect the coordinator controller loading the main controller of the flow rules and then the activation of the installer) Flow rules and allocation of additional load to it.



(a)



(b)

Fig. 11. (a) The amount of incoming traffic in the proposed method before the TCP_SYN flooding attack, during the attack and after it and (b) the round trip time (RTT) of customer requests.

VI. Comparing the Effectiveness of the Proposed Solution to those Already in Use

This section compares the effectiveness of the proposed strategy with four different ways of preventing denial-of-service attacks. The four options are as follows: Ryu controller [40] without any protection method, Ryu controller [41] with MLFQ mechanism, Ryu controller [18] with Flood Defender mechanism, and Ryu controller [40] with FMD-ARA system. The method used by the researchers in [19] provides the basis for this comparison. Fig. 12 (implemented in the Mininet 2.2.2 emulator) depicts the network structure utilized to conduct this comparison [36], [37].

The control layer network in this experiment was out-of-band (a different network from the data layer network), in contrast to the prior tests where the control layer had an in-band network (using the data layer network to communicate with switches and other controllers).

As seen in Fig. 12, in the scenario taken into consideration for the test, 2h interacts with 3h, and 7h communicates with 8h while attackers 1h and 5h perform a UDP flood attack and steadily speed up this attack. Response Request Time (RRT) is measured at about 2h with 3h and 7h with 8h as a standard to evaluate the effectiveness of each solution. RRT is the average response time from the controller and demonstrates the efficiency of the controller in producing network connections. Yes, that is doable. The response request time in Fig. 13 is in relation to 2h and 3h demonstrates the effectiveness of the approach against the denial-of-service attack. The coordinating controller speeds up the attack until it reaches 1000 packets per second (PPS), at which point it activates the sub-controller of the flow rules setter and sends the traffic produced by the UDP flood attack to it. As demonstrated in Fig. 13, this operation has significantly lowered the average response time to legitimate and safe requests to build the flow by reducing the traffic load on the main controller that establishes the flow rules. Fig. 13's findings, derived from the researchers' article from [20], indicate that while the efficiency of the other four techniques improved as the attack speed increased, so did their average response times. The control level is resistant to denial-of-service attacks and performs better thanks to its unique structure (the presence of sub-controllers that determine the flow rules in the main and coordinating controllers).

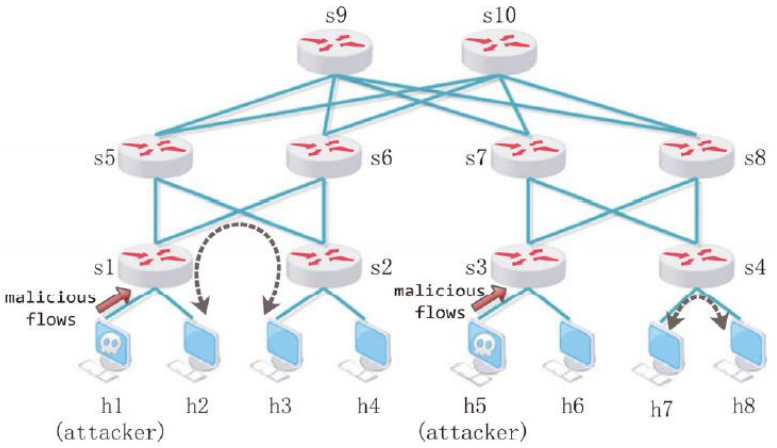


Fig. 12. Network architecture consisting of Open_Flow switches to evaluate the proposed efficiency against denial-of-service attacks.

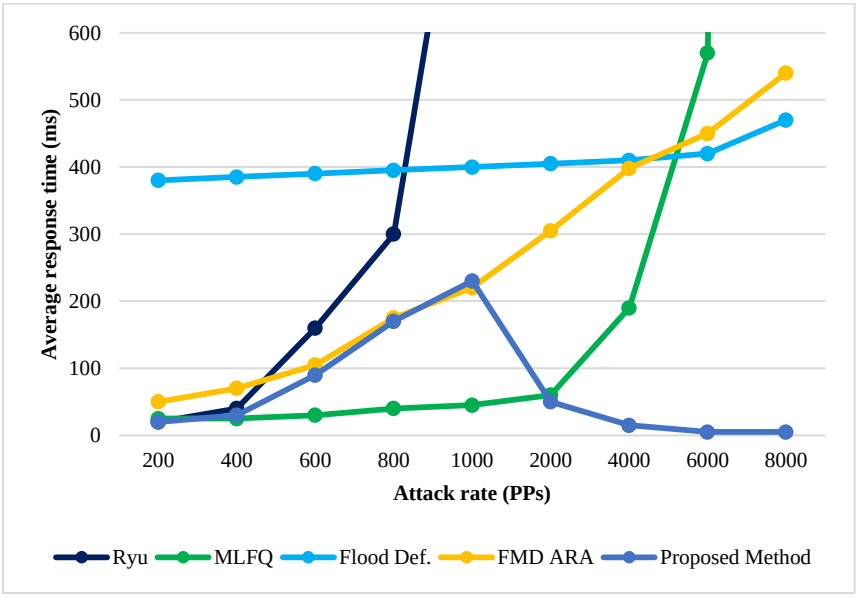


Fig. 13. For current approaches, the typical reaction time is between 2 and 3 hours.

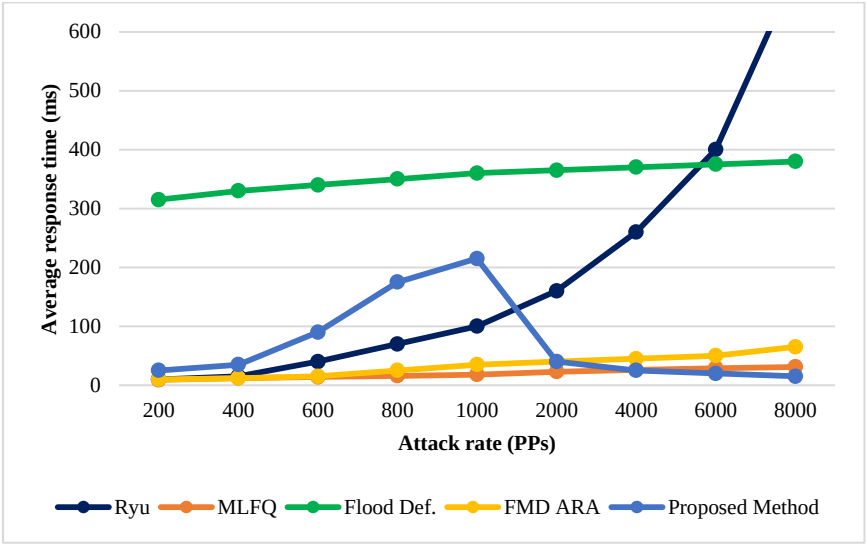


Fig. 14. Based on current techniques, the average reaction time in communication is between 7 and 8.

Fig. 14 illustrates the typical response time for communications between 7 and 8 hours. In contrast to the previous instance, a connection is established between two hosts not directly linked to the switch targeted in the attack. In this instance, the recommended efficiency is not much different from the prior state. Among the four solutions, with the exception of Flood Defender, which relies on a pre-processing system and must temporarily store each new Open_Flow request in a packet buffer before reacting to it, all show better performance than the previous setup. This situation results from the fact that the main factor causing an increase in the reaction time delay in the three solutions, MLFQ, Ryu, and FMD-ARA, is the presence of congestion in the communication channel of the switch with the controller. If, as proposed, the switch's communication channel with the main controller of the flow rules becomes congested and the main controller becomes overloaded, increasing the response time delay, the additional load that caused the congestion is sent through a different communication channel to the main controller. The set of flow rules acts as a guide for a sub-controller. The coordinating controller activates the sub-controller of the flow rules when it notices that the main controller of the flow rules is overloaded. This not only overloads the main controller but also creates a separate communication channel between the sub-controller and the switch. To avoid clogging up the main controller's communication channel with the switch.

The key findings of the study highlight the effectiveness of a novel approach to enhance the security and resilience of software-defined networks (SDNs) against denial-of-service (DoS) attacks. The introduced "control box" solution, consisting of coordinating, main, and sub-controllers, successfully manages and distributes incoming traffic during DoS attacks, reducing response times and maintaining network availability. Experimental validation, using various DoS attack scenarios, demonstrates the method's superiority over existing solutions. Furthermore, comparisons with alternative methods, including Ryu controllers with various mechanisms, consistently show that the proposed approach outperforms them. While the solution offers promising results, the study acknowledges the need for further research to address resource consumption and scalability issues, paving the way for more resource-efficient and scalable network security solutions in the future.

VII. Conclusion

In this article, a new solution for the control layer has been developed with the aim of increasing access to software-based networks by increasing the security of its control layer. The control box is the main part of the proposed method. This control box consists of a main controller that adjusts the flow rules, one or more sub-controllers that adjust the flow rules, and a coordinating controller. The task of the coordinator controller is to monitor the controllers that define the flow rules and monitor the network by collecting statistical data from the Open_Flow switches. Controllers that install flow rules are also responsible for installing flow rules. Based on network information base (NIB) data obtained from the coordinating controller, these flow rules are generated by network programs running on the controllers that install the flow rules. Based on the traffic load entering the network, the coordinating controller also coordinates the operation of the main and sub-controllers of the flow rules. The high tolerance of the proposed method for attacks that lead to the saturation of the control plane by the data plane of SDN networks is due to the redundancy of flow rule set controllers in the method. In summary, the importance of the proposed solution to increase the security and resilience of Software Defined Networks (SDN) against Denial of Service (DoS) attacks is emphasized. They highlight the unique contribution of their "control box" approach, which effectively manages and distributes incoming traffic during DoS attacks, thereby reducing response time and maintaining network availability. The following are the limitations and future works.

The proposed solution, while effective in reducing DoS attacks, requires the deployment of additional resources, which can lead to increased energy consumption and cost. This limitation highlights the need for further optimization to reduce the resource overhead associated with the approach. This article also mentions a limitation on the number of available switch ports for adding sub-controllers to a flow rules installation. This scalability limitation needs to be addressed to ensure that this approach can be applied to larger and more complex SDN environments.

VIII. Future Work

1) Resource efficiency: Future research should focus on optimizing the proposed solution to reduce resource requirements, making it more resource efficient and environmentally sustainable.

2) Scalability solutions: Addressing scalability limitations is a priority. Researchers could explore ways to scale the approach to accommodate larger SDN deployments, possibly through innovations in controller communication and load distribution techniques.

3) Security extensions: Expanding research to cover a wider range of cyber threats and security challenges in SDNs would be valuable. This could include exploring ways to increase security against other types of attacks beyond DoS, ensuring comprehensive network protection.

References

[1] O. E. Tayfour and M. N. Marsono, “Collaborative detection and mitigation of distributed denial-of-service attacks on software-defined network,” Mobile Networks and Applications, vol. 25, pp. 1338–1347, 2020.

[2] H. Farhady, H. Lee, and A. Nakao, “Software-defined networking: A survey,” Computer Networks, vol. 81, pp. 79–95, 2015.

[3] M. A. Aladaileh, M. Anbar, I. H. Hasbullah, Y.-W. Chong, and Y. K. Sanjalawe, “Detection techniques of distributed denial-of-service attacks on software-defined networking controller–a review,” IEEE Access, vol. 8, pp. 143985–143995, 2020.

[4] E. Khezri and E. Zeinali, “A review on highway routing protocols in vehicular ad hoc networks,” SN Comput Sci, vol. 2, pp. 1–22, 2021.

[5] A. F. Abdullah, F. M. Salem, A. Tammam, and M. H. A. Azeem, “Performance analysis and evaluation of software defined networking controllers against denial-of-service attacks,” in Journal of Physics: Conference Series, IOP Publishing, 2020, p. 012007.

[6] W. Li, W. Meng, and L. F. Kwok, “A survey on OpenFlow-based Software Defined Networks: Security challenges and countermeasures,”

Journal of Network and Computer Applications, vol. 68, pp. 126–139, 2016.

[7] M. F. Hyder and T. Fatima, "Towards crossfire distributed denial-of-service attack protection using intent-based moving target defense over software-defined networking," IEEE Access, vol. 9, pp. 112792–112804, 2021.

[8] J. Benabbou, K. Elbaamrani, and N. Idboufker, "Security in OpenFlow-based SDN, opportunities and challenges," Photonic Network Communications, vol. 37, pp. 1–23, 2019.

[9] M. Sakthivel, R. Kamalraj, S. Sivanantham, and V. Krishnamoorthy, "An Analysis of Machine Learning Depend on Q-MIND for Defencing The Distributed Denial of Service Attack on Software Defined Network.," International Journal of Early Childhood Special Education, vol. 14, no. 5, 2022.

[10] Cao, Y., Xu, N., Wang, H., Zhao, X., & Ahmad, A. M. (2023). Neural networks-based adaptive tracking control for full-state constrained switched nonlinear systems with periodic disturbances and actuator saturation. International Journal of Systems Science, 54(14), 2689-2704.

[11] B. Mladenov and G. Iliev, "Optimal software-defined network topology for distributed denial-of-service attack mitigation," Bulletin of Electrical Engineering and Informatics, vol. 9, no. 6, pp. 2588–2594, 2020.

[12] E. Khezri, E. Zeinali, and H. Sargolzaey, "SGHRP: Secure Greedy Highway Routing Protocol with authentication and increased privacy in vehicular ad hoc networks," PLoS One, vol. 18, no. 4, p. e0282031, 2023.

[13] Wang, T., Zhang, L., Xu, N., & Alharbi, K. H. (2023). Adaptive critic learning for approximate optimal event-triggered tracking control of nonlinear systems with prescribed performances. International Journal of Control, 1-15.

[14] H. Wang, L. Xu, and G. Gu, "OF-GUARD: A DoS attack prevention extension in software-defined networks," The Open Network Summit (ONS), no. 2014, 2014.

[15] D. Agnew, S. Boamah, R. Mathieu, A. Cooper, J. McNair, and A. Bretas, "Distributed Software-Defined Network Architecture for Smart Grid Resilience to Denial-of-Service Attacks," arXiv preprint arXiv:2212.09990, 2022.

[16] M. Ambrosin, M. Conti, F. De Gaspari, and R. Poovendran, "Lineswitch: Tackling control plane saturation attacks in software-defined networking," IEEE/ACM Transactions on Networking, vol. 25, no. 2, pp. 1206–1219, 2016.

[17] M. Trik, A. M. N. G. Molk, F. Ghasemi, and P. Pouryeganeh, "A hybrid selection strategy based on traffic analysis for improving performance in networks on chip," J Sens, vol. 2022, 2022.

[18] Yue, S., Niu, B., Wang, H., Zhang, L., & Ahmad, A. M. (2023). Hierarchical sliding mode-based adaptive fuzzy control for uncertain switched under-actuated nonlinear systems with input saturation and dead-zone. Robotic Intelligence and Automation, 43(5), 523-536.

[19] K. Chen, A. R. Junuthula, I. K. Siddhrau, Y. Xu, and H. J. Chao, "SDNShield: Towards more comprehensive defense against DDoS attacks on SDN control plane," in 2016 IEEE conference on communications and network security (CNS), IEEE, 2016, pp. 28–36.

[20] T. Semong et al., "Intelligent load balancing techniques in software defined networks: A survey," Electronics (Basel), vol. 9, no. 7, p. 1091, 2020.

[21] Zhao, H., Wang, H., Xu, N., Zhao, X., & Sharaf, S. (2023). Fuzzy approximation-based optimal consensus control for nonlinear multiagent systems via adaptive dynamic programming. Neurocomputing, 553, 126529.

[22] A. K. A. Al-Mashadani and M. Ilyas, "Distributed Denial of Service Attack Alleviated and Detected by Using Mininet and Software Defined Network," Webology, vol. 19, no. 1, pp. 4129–4144, 2022.

[23] Zhang, H., Zou, Q., Ju, Y., Song, C., & Chen, D. (2022). Distance-based support vector machine to predict DNA N6-methyladenine modification. Current Bioinformatics, 17(5), 473-482.

[24] O. Polat and H. Polat, "An intelligent software defined networking controller component to detect and mitigate denial-of-service attacks," Journal of Information and Communication Technology, vol. 20, no. 1, pp. 57–81, 2021.

[25] Cao, C., Wang, J., Kwok, D., Cui, F., Zhang, Z., Zhao, D., ... & Zou, Q. (2022). webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. Nucleic acids research, 50(D1), D1123-D1130.

[26] S. Sharathkumar and N. Sreenath, "Distributed Clustering based Denial of Service Attack Prevention Mechanism using a Fault Tolerant Self Configured Controller in a Software Defined Network," 2023.

[27] Arefanjazi, H., Ataei, M., Ekramian, M., & Montazeri, A. (2023). A robust distributed observer design for Lipschitz nonlinear systems with time-varying switching topology. Journal of the Franklin Institute, 360(14), 10728-10744.

[28] E. Khezri, E. Zeinali, and H. Sargolzaey, "A novel highway routing protocol in vehicular ad hoc networks using VMaSC-LTE and DBA-MAC protocols," Wirel Commun Mob Comput, vol. 2022, 2022.

[29] Wang, Z., Jin, Z., Yang, Z., Zhao, W., & Trik, M. (2023). Increasing efficiency for routing in internet of things using Binary Gray Wolf Optimization and fuzzy logic. Journal of King Saud University-Computer and Information Sciences, 35(9), 101732.

[30] M. Samiei, A. Hassani, S. Sarspy, I. E. Komari, M. Trik, and F. Hassanpour, "Classification of skin cancer stages using a AHP fuzzy technique within the context of big data healthcare," J Cancer Res Clin Oncol, pp. 1–15, 2023.

[31] J. Sun, Y. Zhang, and M. Trik, "PBPHS: a profile-based predictive handover strategy for 5G networks," Cybern Syst, pp. 1–22, 2022.

[32] M. Trik, H. Akhavan, A. M. Bidgoli, A. M. N. G. Molk, H. Vashani, and S. P. Mozaffari, "A new adaptive selection strategy for reducing latency in networks on chip," Integration, vol. 89, pp. 9–24, 2023.

[33] G. Shang, P. Zhe, X. Bin, H. Aiqun, and R. Kui, "FloodDefender: Protecting data and control plane resources under SDN-aimed DoS attacks," in IEEE INFOCOM 2017-IEEE Conference on Computer Communications, IEEE, 2017, pp. 1–9.

[34] P. Wu, L. Yao, C. Lin, G. Wu, and M. S. Obaidat, "Fmd: A DoS mitigation scheme based on flow migration in software-defined networking," International Journal of Communication Systems, vol. 31, no. 9, p. e3543, 2018.

[35] Y. Wang, T. Hu, G. Tang, J. Xie, and J. Lu, "SGS: Safe-guard scheme for protecting control plane against DDoS attacks in software-defined networking," IEEE Access, vol. 7, pp. 34699–34710, 2019.

[36] Y. Han, J.-H. Yoo, and J. W.-K. Hong, "Poisson shot-noise process based flow-level traffic matrix generation for data center networks," in 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM), IEEE, 2015, pp. 450–457.

[37] D. Mokhlesi Ghanevati, E. Khorami, B. Boukani, and M. Trik, "Improve replica placement in content distribution networks with hybrid technique," Journal of Advances in Computer Research, vol. 11, no. 1, pp. 87–99, 2020.

[38] M. Trik, S. P. Mozaffari, and A. M. Bidgoli, "Providing an adaptive routing along with a hybrid selection strategy to increase efficiency in NoC-based neuromorphic systems," Comput Intell Neurosci, vol. 2021, 2021.

[39] B. Yu, G. Yang, and C. Yoo, "Comprehensive prediction models of control traffic for SDN controllers," in 2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft), IEEE, 2018, pp. 262–266.

[40] Q. Gao, "Recommended System Optimization in Social Networks based on Cooperative Filter with Deep MVR Algorithm," 2022.

[41] Y. Ashgevari and M. Karami, "Study of Atmospheric Discharge Effects in Distribution Networks with a Novel Residential Buildings Protection Approach," Advances in Engineering and Intelligence Systems, vol. 2, no. 02, 2023.

[42] Karthika, P., & Karmel, A. (2023). Review on distributed denial-of-service attack detection in software defined network. International Journal of Wireless and Mobile Computing, 25(2), 128-146.

Optimization of Unsupervised Neural Machine Translation Based on Syntactic Knowledge Improvement

Aiping Zhou
Department of Fundamental Courses, Southeast University Chengxian College, Nanjing, 210088, China

***Abstracts*—Unsupervised Neural Machine Translation is a crucial machine translation method that can translate in the absence of a parallel corpus and opens up new avenues for intercultural dialogue. Existing unsupervised neural machine translation models still struggle to deal with intricate grammatical relationships and linguistic structures, which leads to less-than-ideal translation quality. This study combines the Transformer structure and syntactic knowledge to create a new unsupervised neural machine translation model, which enhances the performance of the existing model. The study creates a neural machine translation model based on the Transformer structure first, and then introduces sentence syntactic structure and various syntactic fusion techniques, also known as the Transformer combines grammatical knowledge. The results show that the Transformer combines grammatical knowledge paired with Bi-Long Short-Term Memory proposed in this research has better performance. The accuracy and F1 value of the combined model in the training dataset are as high as 0.97. In addition, the time of the model in real sentence translation is controlled within 2s, and the translation accuracy is above 0.9. In conclusion, the unsupervised neural machine translation model proposed in this study has better performance, and its application to actual translation can achieve better translation results.**

***Keywords*—Unsupervised; Neural network; Machine translation; Grammatical knowledge; Transformer; LSTM**

I. Introduction

With the progress of globalization and the increasing frequency of information exchange, machine translation, is an important artificial intelligence technology, plays an important role in connecting different languages and cultures [1-2]. Machine Translation (MT) is the process of using computer software to convert text or speech from one natural language to another. MT, as a branch of computer-assisted translation, aims to achieve barrier-free language communication between people. Unsupervised Neural Machine Translation (UNMT) belongs to one kind of MT, and the use of UNMT model to carry out translation tasks can not only improve the translation speed and save the cost, but also be able to deal with multiple languages at the same time, which is an important value of language utilization. As a research direction that has been developing rapidly in recent years, the main purpose of UNMT is to provide the best solution to the problem in the field of unsupervised translation, and to provide the best solution to the problem. Its main purpose is to carry out automatic translation without parallel corpus, so as to improve the speed and accuracy of machine translation [3-4]. Currently, the traditional UNMT model is still facing a series of challenges, for example, in the environment without parallel corpus, UNMT is often difficult to obtain effective linguistic correspondences, thus affecting the quality and accuracy of translation [5]. Against this background, the emergence of the Transformer structure has revolutionized the field of machine translation, especially its demonstrated efficiency in sequence-to-sequence learning [6]. At the same time, how to better integrate grammatical knowledge into the UNMT model has become an urgent challenge. Based on the above problems and challenges, this study aims to deeply explore and propose a novel UNMT model, which not only incorporates the advantages of the Transformer structure, but also the theoretical features of grammatical knowledge. The newly constructed UNMT model aims to achieve higher translation quality and effect, and at the same time solve the technical and theoretical problems of the traditional UNMT model in machine translation, so as to provide certain technical reference value for the field of machine translation.

In order to facilitate readers to better understand the content of the article and the framework of the article, this research divides the article into a total of six sections, which are the introduction, literature review, method design, result analysis, discussion and conclusion chapters. Section I mainly introduces the background of the study, the current status of the study, the research methodology, and the significance of the study. The literature review chapter mainly analyzes and summarizes the related studies of others so as to prove the novelty of this research which is mentioned in Section II. The research methodology in Section III mainly explains how to build the optimized UNMT model and introduces some related machine translation techniques. The result analysis section mainly tests the performance and practical application effect of the UNMT model, so as to prove the effectiveness of the model which is mentioned in Section IV. The discussion in Section V is mainly to further analyze and summarize the reasons for the better performance of the model according to the experimental results. The conclusion in Section VI is a concise summary of the whole paper.

II. Related Work

UNMT is an approach to MT whose main feature is to translate without a parallel corpus. Currently, it has been optimised by a number of experts in combination with deep learning. In order to address the drawback that remotely supervised relational reminding is seriously affected by

mislabelling in practical applications, Xiao et al. proposed a Transformer module for remotely supervised relational reminding with multi-instance learning using a hybrid attention mechanism. On the remote supervised relational reminding task, experimental results demonstrated that the method developed by the research performed better than the state-of-the-art algorithms at the time [7]. The Transformer concept has additionally been used in the machining sector. The a priori knowledge of the target text could not be fully utilised by the conventional automatic speech mistake detection systems, according to Zhang et al. Therefore, Zhang et al. proposed to apply the Transformer model to it, and the results of this study showed that the method could obtain a relative improvement of 8.4% on the F-1 scoring metrics, which is advancing significance for the optimisation of automatic speech error detection methods [8]. Li et al. concluded that the existing anomaly detection methods in the power industry do not fully exploit the potential value of the data. Therefore, an anomaly detection model based on graph attention and Transformer was proposed. Li et al. designed experiments based on power data in a region of China [9]. Li et al. argued that current neural TTS models suffer from robustness problems thus leading to audio anomalies. In order to construct a Neural network (NN) model capable of synthesising both natural and stable audio, thus a Transformer based TTS model called RobuTrans was proposed in [10]. Through experiments, it was found that the model solves the robustness problem that exists in the TSS model. According to Xiao et al., the current entity and relation extraction suffers from noise labelling issues and is unable to recognise the relationship between relations and phrases. This led to the proposal of a hybrid depth NN model based on Transformer and other models. The outcomes of many experiments demonstrated that the model was superior at entity and relation extraction and could filter noisy words [11].

As a posteriori regularisation technique to direct the training efficiency of unsupervised MT models during repeated reverse translation, Ren et al. presented a phrase-based statistical MT model. This study jointly optimises the SMT and NMT models under a unified expectation maximisation framework and gradually improves the performance of both models during the iterative process. The results of the study show that the proposed scheme can achieve two advantages. Filtering the noise in the phrase table by SMT can promptly mitigate the negative impact of errors during iterative back translation. Meanwhile, NMT can make up for the inherent lack of fluency in SMT [12]. Li et al. utilised spatio-temporal maps obtained from videos and the spatial and temporal interactions of objects to facilitate potential spatial alignment and remove translation ambiguity in UNMT. The designed model employs multimodal backtranslation and feature pseudo-visual hubs, and learns a shared multilingual visual-semantic embedding space that fuses visual hub subtitles as additional weak supervision. The proposed model is validated on the VATEX Translation 2020 and HowToWorld datasets for translation in sentences and words with good generalisation performance [13]. Sun et al. empirically investigated the performance of four different languages (French, German, Chinese, and Japanese) on the English UNMT model. In addition, a simple general method is proposed for improving the translation performance of these four language pairs. To address the shortcoming that different language pairs have significant delayed convergence in the denoising process, Sun H et al. proposed a pseudo-data based UNMT [14].

In summary, a number of scholars have carried out a series of studies on Transformer structure and UNMT model. Among them, the research on Transformer structure mainly focuses on the detection of various abnormal signals and data, while the research on UNMT model focuses on the optimisation of model translation effect. Based on the above background, this research innovatively fuses the Transformer structure with GK and uses the fused model in the field of UNMT, aiming at better extraction of English grammatical error features and detection of its incorrect grammar.

III. UNMT Model Construction Based on GK Improvement

The emergence of neural MT models has led to the gradual replacement of end-to-end single MT, and the continuous optimisation of neural MT models has also made the translation effect of various small language translation models closer and closer to that of human translation. In this research, the traditional Long Short-Term Memory (LSTM) and the neural MT model under the Transformer structure are firstly introduced, and then it is optimised by combining with the GK structure, and a new UNMT model is proposed.

A. Research on Neural MT Modelling Based on Transformer Structure

With the continuous combination of deep learning and MT technology, the neural MT model has become the most mainstream intelligent translation model. The biggest advantage of Recurrent Neural Network (RNN) in English translation is its ability to remember the historical information of a sentence, to expand the individual words in a sentence in time steps, and to check its grammar [15]. RNN automatically corrects grammatical errors by transforming the input natural language text into the output of the RNN so that grammatical errors can be corrected automatically. The standard expression form of the RNN is shown in Eq. (1) and Eq. (2).

$$h_t = \varphi_h\left(W_h x_t + U_h h_{t-1} + b_h\right) \tag{1}$$

In Eq. (1), x_t denotes the input of the time step at moment t and $t = [1, 2, \cdots m]$. h_t denotes the implied state of the output of the time step at moment t. W_h, U_h, and b_h denote the relevant parameters of the output implied state, respectively. φ_h, on the other hand, denotes the nonlinear activation function of the output implied state [16].

$$y_t = \varphi_y\left(W_y h_t + b_y\right) \tag{2}$$

In Eq. (2), y_t denotes the output of the network at the moment t. W_y and b_y are the relevant parameters of the output state of the network, respectively. φ_y denotes the nonlinear activation function of the network output. Researchers optimised the RNN and designed new loop units

by adjusting the nonlinear activation function in the network. The common LSTM gradually gets new applications in MT problems. The unit structure of LSTM is shown in Fig. 1.

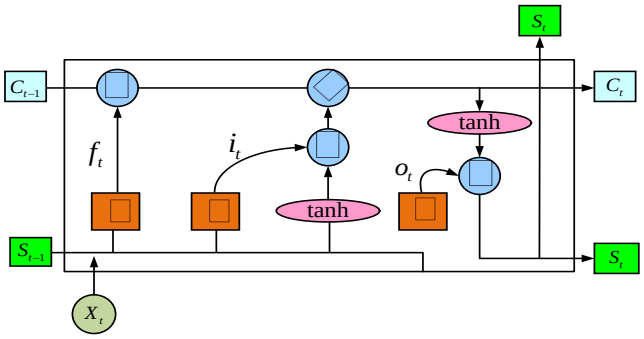


Fig. 1. LSTM cell structure diagram.

In Fig. 1, the LSTM consists of three gate structures, and selectively receives information through memory units. Where X_t denotes the input data of the input layer at the moment t. S_t denotes the neuron state of the hidden layer at the moment t. C_t denotes the memory unit at the moment t. The three σ in Fig. 1 indicate the three gate structures in LSTM from left to right. f_t, i_t, and o_t indicate the parameters of the three gates, respectively [17].

$$f_t = \sigma' W_x^f \cdot X_t + W_s^f \cdot S_{t-1} + b_f \tag{3}$$

In Eq. (3), W_x^f, W_s^f denotes the weight matrix, b_f denotes the input gate bias vector. σ' denotes the input gate. f_t denotes the input gate parameters.

$$i_t = \sigma'' W_x^i \cdot X_t + W_s^i \cdot S_{t-1} + b_i \tag{4}$$

In Eq. (4), W_x^i and W_s^i denote the weight matrix, b_i denotes the forgetting gate bias vector. σ'' denotes the forgetting gate. S_{t-1} denotes the neuron state at the moment of $t-1$. i_t denotes the forgetting gate parameter.

$$o_t = \sigma''' W_x^o \cdot X_t + W_s^o \cdot S_{t-1} + b_o \tag{5}$$

In Eq. (5), W_x^o, W_s^o denotes the weight matrix, b_o denotes the output gate bias vector. σ''' denotes the output gate. o_t is the output gate parameter.

$$\tilde{c}_t = \tanh\left(W_x^c \cdot X_t + W_s^c \cdot S_{t-1} + b_c\right) \tag{6}$$

In Eq. (6), $\tilde{c}_t$ is the output of the memory cell obtained by the forgetting gate parameter i_t after the calculation of the tanh function. W_x^c and W_s^c denote the weight matrix. b_c denotes the bias vector of the forgetting gate after the calculation of the tanh function.

$$c_t = i_t \square \tilde{c}_t + f_t \square c_{t-1} \tag{7}$$

In Eq. (7), $\square$ denotes the Hardman product.

$$S_t = o_t \square \tanh(c_t) \tag{8}$$

Eq. (8) is the equation for the neuron state of the hidden layer at moment t.

Since both RNN and LSTM are prone to the problems of gradient vanishing and gradient explosion when performing parameter updates, the study builds a small language translation model by combining the Transformer structure. Transformer is an NN structure used for sequence-to-sequence learning, which is better able to deal with the problem of long text. The structure of Transformer is shown in Fig. 2.

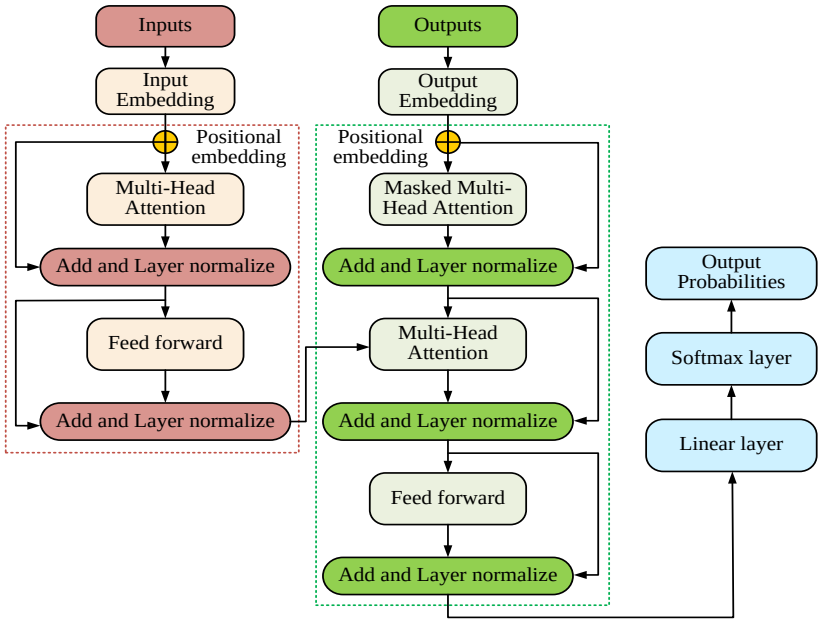


Fig. 2. Structure of the transformer model.

Fig. 2 depicts the Transformer model's structural layout. The operation flow of the encoder and decoder is shown in Eq. (9) and Eq. (10) [18].

$$e_1, e_2, \cdots, e_m = encoder\left(X1, X2, \cdots, Xm\right) \quad (9)$$

In Eq. (9), $e_1, e_2, \cdots, e_m$ denotes a string of input text sequence. $encoder\left(X1, X2, \cdots, Xm\right)$ denotes the encoder for encoding.

$$Y_t = decoder\left(e_1, e_2, \cdots, e_m, Y1, Y2, \cdots Y_{t-1}\right) \quad (10)$$

In Eq. (10), Y_t denotes the probability distribution vector of the decoded data, which is obtained by calculating the SoftMax function. $decoder\left(e_1, e_2, \cdots, e_m, Y1, Y2, \cdots Y_{t-1}\right)$ denotes the decoding operation on the probability distribution vector.

B. Study of UNMT Modelling Incorporating the Transformer Structure and GK

Since the traditional Transformer model tends to ignore the semantic information of the sentence in the translation process, which leads to the translation result deviating from the actual meaning, this study further proposes an optimized UNMT model based on the traditional Transformer structure combined with GK, notated as Transformer combines grammatical knowledge (TCGK). Traditional UNMT is an approach for training in MT tasks without using parallel corpus, i.e., corresponding sentence pairs between source and target languages. Fig. 3 depicts the UNMT model's fundamental structure.

Fig. 3 depicts the UNMT model's overall structure. Two monolingual semantic repositories, a language modelling board, and a reverse translation board are the primary components of the UNMT architecture shown in Fig. 3 [19]. When the words in the two monolingual semantic repositories are input into the model, they first need to be initialised. In the initialisation process, the main purpose is to encode the words, phrases and words so that each word can be recognised by the UNMT model, thus achieving the purpose of training the model. After the initialisation process, language modelling is required. In the modelling process, the encoder-decoder structure is used for denoising and at the same time allowing the encoder to learn the semantic information of the monolingual data. The mathematical expression for language modelling is shown in Eq. (11).

$$L_{\min} = E\left[-\log P_{s\to s}\left(x\middle|C\left(x\right)\right)\right] + E\left[-\log P_{t\to t}\left(x\middle|C\left(x\right)\right)\right] \quad (11)$$

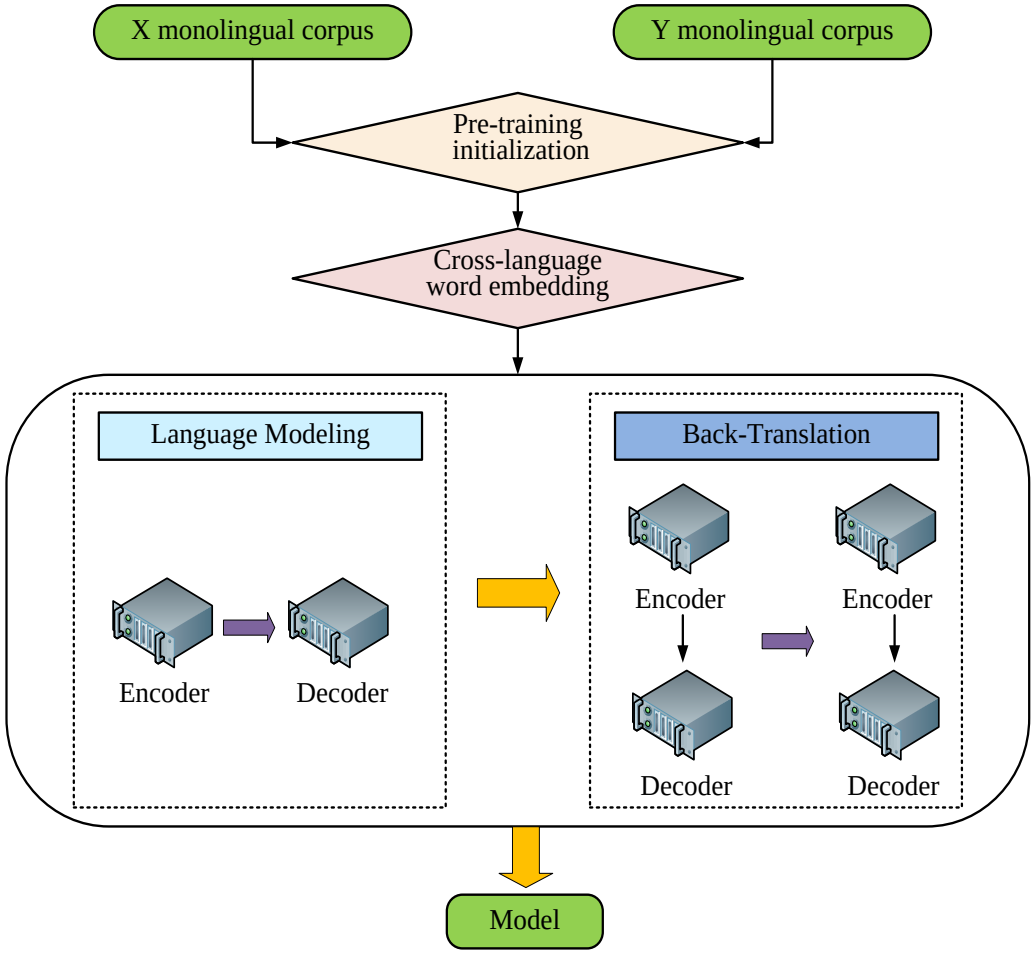


Fig. 3. Unsupervised neural machine translation model structure.

In Eq. (11), $L_{\min}$ denotes the minimum loss in the modelling process. C denotes the noise model. x denotes the sentence in the X monolingual semantic base. $P_{s\to s}$ denotes the source-side encoder-decoder combination. $P_{t\to t}$ denotes the target-side encoder-decoder combination. E denotes the energy consumption in the modelling process. The mathematical expression for reverse translation is shown in Eq. (12).

$$L'_{\min} = E\left[-\log P_{s\to t}\left(y\middle|u*(y)\right)\right] + E\left[-\log P_{t\to s}\left(x\middle|v*(x)\right)\right] \quad (12)$$

In Eq. (12), $L'_{\min}$ denotes the minimum loss in the reverse translation process. y denotes a sentence in the Y monolingual semantic base. $u*(y)$ denotes translating the source language according to the target language. $v*(x)$ denotes translating the target language according to the source language. $P_{s\to t}$ denotes the direction of translation from the target language to the source language. $P_{t\to s}$ denotes the direction of translation from the source language to the target language.

In order to learn the translation relationships between the source and destination languages, neural MT models typically need a sizable parallel corpus for training that comprises corresponding sentence pairs between the two languages. However, UNMT does not rely on parallel corpus, but is trained by using monolingual corpus. The UNMT model is obtained by optimising the encoder-decoder structure, and the basic idea is to learn the correspondence between source and target languages through self-supervised learning of monolingual corpus, so as to achieve the MT task. This study takes GK and syntactic structure into account for the model's optimisation on the basis of the conventional UNMT model in order to give the UNMT model a better translation effect that can accurately translate according to the syntactic structure and be close to the actual semantic environment. Firstly, the sentence syntactic structure is introduced, as shown in Fig. 4.

Fig. 4 shows the syntactic tree structure diagram of the sentence. To translate a complete sentence according to the actual context and grammatical structure, it is necessary to split its sentence syntactic structure first [20]. In Fig. 4, it can be seen that a complete sentence is composed of sentences or phrases. For phrase structure, its extracted syntactic labels contain constituent categories and phrase structure information. The hierarchical output of the phrase structure syntax contains the information of the various categories of words and the attributes of the words.

In order to allow unsupervised MTs to have a better knowledge of syntax, this research decided to use the results of the syntactic analysis to optimise the translation results of the model. After linearising these results and extracting their syntactic labels, and then combining them with the corresponding sentences, the combined data is used to train the source side of the denoising autoencoder. In this way, sentences and syntactic information can be jointly encoded into a new vector, thus creating a language model that incorporates syntactic information. The process of training the model with fused syntactic knowledge is shown in Fig. 5.

A flowchart of the model training process incorporating syntactic knowledge is demonstrated in Fig. 5. In Fig. 5, the optimised model has adapted the inputs of the denoising autoencoder, and multiple encoders are used to process the monolingual corpus, lexicality, phrase structure, and dependency syntax, respectively. During training, the model absorbs lexical and syntactic information and optimises the shared encoder and decoder parameters to better capture the implicit syntactic information in sentences. When decoding, the model utilises semantic, lexical and syntactic information to assist in the predictive generation of target words. By incorporating lexical and syntactic knowledge, the challenge of not being able to explicitly learn syntax can be addressed and the accuracy of the translation can be improved.

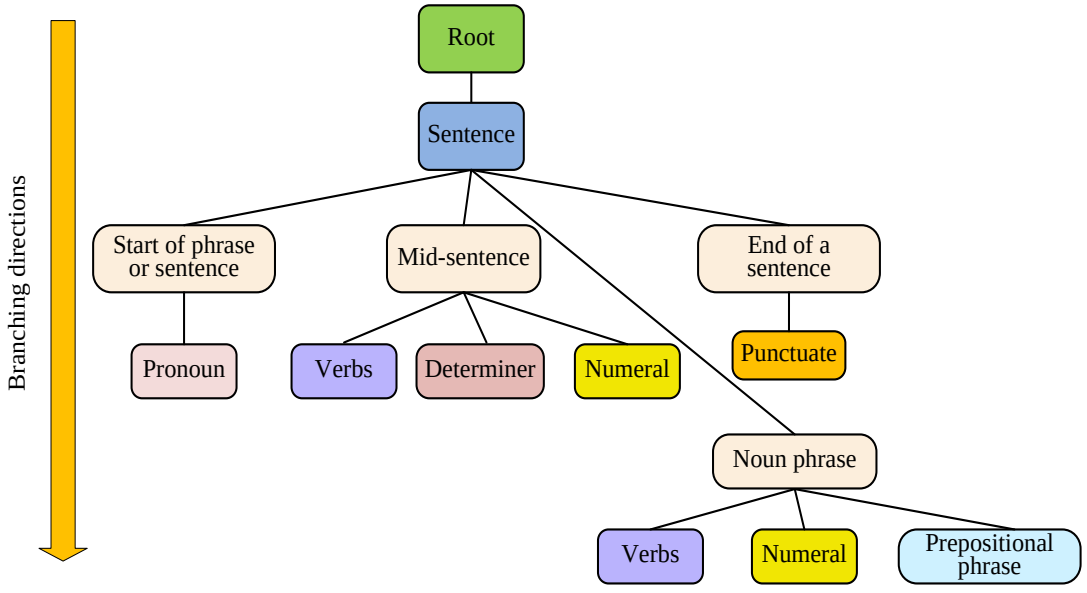


Fig. 4. Sentence syntactic tree structure.

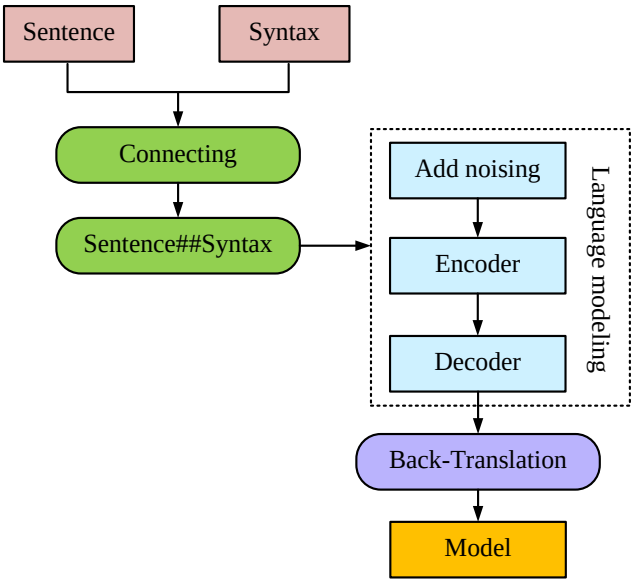


Fig. 5. Flow chart of model training incorporating syntactic knowledge.

In addition to adopting the approach of fusing syntactic knowledge in Fig. 5, the study also uses Bi-Long Short-Term Memory (Bi-LSTM) in NN to extract vectors of features for syntactic analysis sequences and splice them with sentence vectors to obtain a vector containing semantic and syntactic information, and uses this vector to train TCGK, so as to improve the translation quality of the TCGK model. The flow chart of Bi-LSTM extraction of syntactic vectors is shown in Fig. 6. The flowchart of Bi-LSTM for extracting syntactic vectors is shown in Fig. 6.

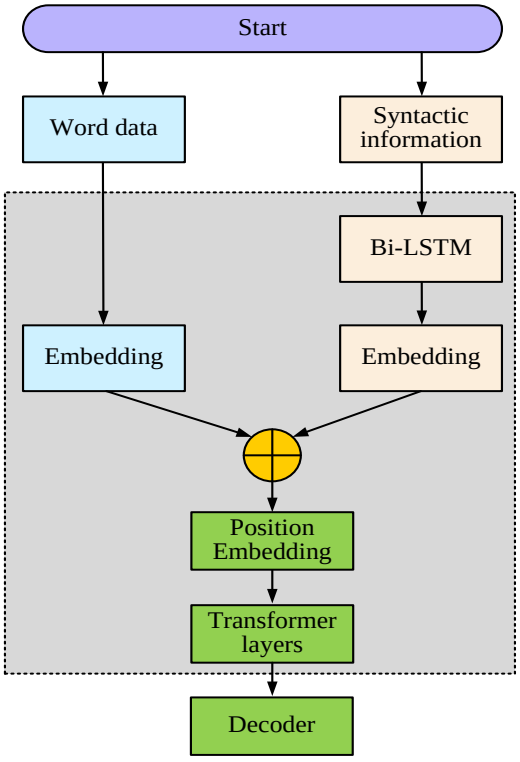


Fig. 6. Flow chart of sentence normal vector extraction under Bi-LSTM.

As shown in Fig. 6, Bi-LSTM is used to extract syntactic vectors. In order to more closely combine monolingual sentences and explicit syntactic information, the syntactic tree sequence is first linearised. At the input of Transformer, this study combines sentence vectors with syntactic vectors processed through Bi-LSTM to form new fusion vectors. The syntactic features are first transformed into high-dimensional vectors through the word embedding layer of the NN and then spliced with the sentence vectors, and this resulting integrated feature vector does not involve modifying the syntactic content. Using this fused vector, the encoder-decoder starts iterative training and stops iterative training until the model has better translation results.

IV. Performance Analysis of UNMT Models using Transformer Structure and GK

The result analysis section tested the performance of various types of translation models before testing the UNMT model created in the aforementioned study. This demonstrated that the translation performance of the model used in this study was superior through the indicators of detection accuracy, change of loss curves, and F1 value. In addition, the study further compares the translation effect of each translation model in practical applications. The results of the study found that the UNMT model combined with Bi-LSTM has higher translation accuracy and teacher-student satisfaction.

A. Performance Analysis of Different Translation Models

The News Crawl dataset was first chosen as the experimental dataset, and Newstest2020 and Newstest2021 were chosen as the experimental training dataset and test dataset, respectively, to test the performance of the model under the Transformer structure. In Newstest2020 and Newstest2021, there were 5000 corpora each. The corpus for Newstest2020 and Newstest2021 is 5000. Table I displays the settings for the experimental model's parameters.

The basic network parameters in the Transformer model are given in Table I, including the number of its encoder-decoder layers, the number of layers of the multi-head attention mechanism, the dimensions of the word embedding and hidden layers, and the learning rate. The detailed composition of the Newstest2020 and Newstest2021 experimental dataset information is shown in Table II.

Table II shows the details of the Newstest2020 and Newstest2021 experimental datasets, describing the source of the datasets, the composition of the language pairs, the number of samples, and the purpose of the dataset usage, respectively. In addition to utilizing the Newstest2020 and Newstest2021 experimental datasets for testing, the study also selected some public language datasets for testing. In order to compare the performance of the two syntactic fusion methods in the model TCGK, this study notated the syntactic fusion approach in Fig. 5 as TCGK+Common coding, and the syntactic fusion approach in Fig. 6 as TCGK+Bi-LSTM, and introduced the traditional Transformer model as well as the LSTM model for the comparison of the model translation performance. The detection accuracy of the four models in different datasets is shown in Table III.

TABLE I. PARAMETER SETTING OF THE EXPERIMENTAL MODEL

Transformer model parameters	Parameter values
Number of encoder and decoder layers	4 layers
Number of layers of multi-head attention mechanism	8 layers
Dimension of word embedding and hidden layers	1024
Learning rate under Adam's optimization method	0.001
Batch Size	32
Model regularization Dropout	0.1

TABLE II. EXPERIMENTAL DATA SET INFORMATION TABLE

Data set information	introduce
Source	It mainly contains news texts, which are sourced from various news websites and agencies.
Language pair	These two datasets cover multiple language pairs. For example, English to German, English to French, English to Chinese, etc.
Sample size	The number of corpus in Newstest2020 and Newstest2021 is 5000.
Aim	Evaluate the translation performance of machine translation models.

In Table III, a total of five public datasets, Newstest2020, Newstest2021, Para Crawl, Europarl, and Common Crawl, are selected for testing. Europarl: dataset is a dataset based on the records of the European Parliament, covering 21 European languages. Common Crawl is a multilingual aligned dataset based on web crawling. Para Crawl is a multilingual parallel corpus for large-scale web crawling. As shown in Table III, the detection accuracies of LSTM in Newstest2020, Newstest2021, ParaCrawl, Europarl, and Common Crawl are 0.78, 0.77, 0.71, 0.66, and 0.69, respectively. Transformer in Newstest2020, Newstest2021, ParaCrawl, Europarl, and Common Crawl with detection accuracies of 0.82, 0.83, 0.75, 0.71, and 0.72, respectively. TCGK+Common coding in Newstest2020, Newstest2021, Para Crawl, Europarl, and Common Crawl were 0.88, 0.89, 0.86, 0.82, and 0.83, respectively. The detection accuracy of TCGK+Bi-LSTM in Newstest2020, Newstest2021, Para Crawl, Europarl, Common Crawl are 0.96, 0.97, 0.91, 0.92, and 0.93, respectively. Among them, the TCGK+Bi-LSTM model is able to achieve the highest detection accuracy in the datasets Newstest2020 and Newstest2021. The translation performance of the four models will be further tested in combination with the datasets Newstest2020 and Newstest2021.

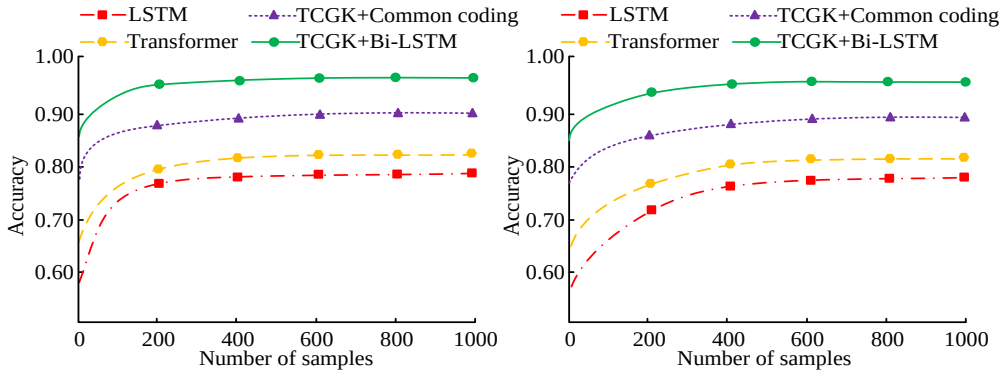
In Fig. 7, the incorrect syntax detection accuracy values for the various translation models in the training dataset and test dataset are displayed. Fig. 7(a) and Fig. 7(b) among them illustrate the detection accuracy of the four translation models for the training dataset and the test dataset, respectively: LSTM, Transformer, TCGK+Common Coding, and TCGK+Bi-LSTM. The four translation models' detection accuracies for faulty grammar samples exhibit an increasing trend as the number of detected samples rises, as shown in Fig. 7(a) and Fig. 7(b). The four translation models, LSTM, Transformer, TCGK+Common coding, and TCGK+Bi-LSTM, each have detection accuracy scores in the training dataset that are 0.78, 0.82, 0.90, and 0.97, respectively. In the testing dataset, the highest detection accuracy values of the four translation models, LSTM, Transformer, TCGK+Common coding, and TCGK+Bi-LSTM, are 0.78, 0.82, 0.90, and 0.97, respectively. coding, and TCGK+Bi-LSTM, the highest detection accuracy values of the four translation models are 0.77, 0.81, 0.88, and 0.96, respectively.

The graphs of the variation of loss values for different translation models are shown in Fig. 8. Among them, all the figures in Fig. 8 show the actual loss curves and the specific changes of the training loss curves of the four translation models, namely, LSTM, Transformer, TCGK+Common coding, and TCGK+Bi-LSTM, in the training process, respectively. Comparing the loss change curves of the four models, it can be found that compared to the other three models, the training loss curve and the actual loss curve of TCGK+Bi-LSTM basically overlap during the training process, so the stability of this model is better during the training process, and there will not be large data fluctuations.

The variance of F1 values for various translation models in the training dataset and test dataset is depicted in Fig. 9. The study introduces F1 values for testing to better represent the detection performance of each model. The F1 values obtained by the four models in the training dataset and test dataset are displayed in Fig. 9(a) and Fig. 9(b), respectively. In Fig. 9(a), the highest F1 values of the four translation models, LSTM, Transformer, TCGK+Common coding, and TCGK+Bi-LSTM, are 0.76, 0.81, 0.90, and 0.97, respectively. In Fig. 9(b), the highest F1 values of the four translation models, LSTM, Transformer, TCGK+Common coding, and TCGK+Bi-LSTM, the highest F1 values of the four translation models are 0.75, 0.80, 0.89, and 0.96, respectively.

TABLE III. DETECTION ACCURACY OF FOUR MODELS IN DIFFERENT DATA SETS

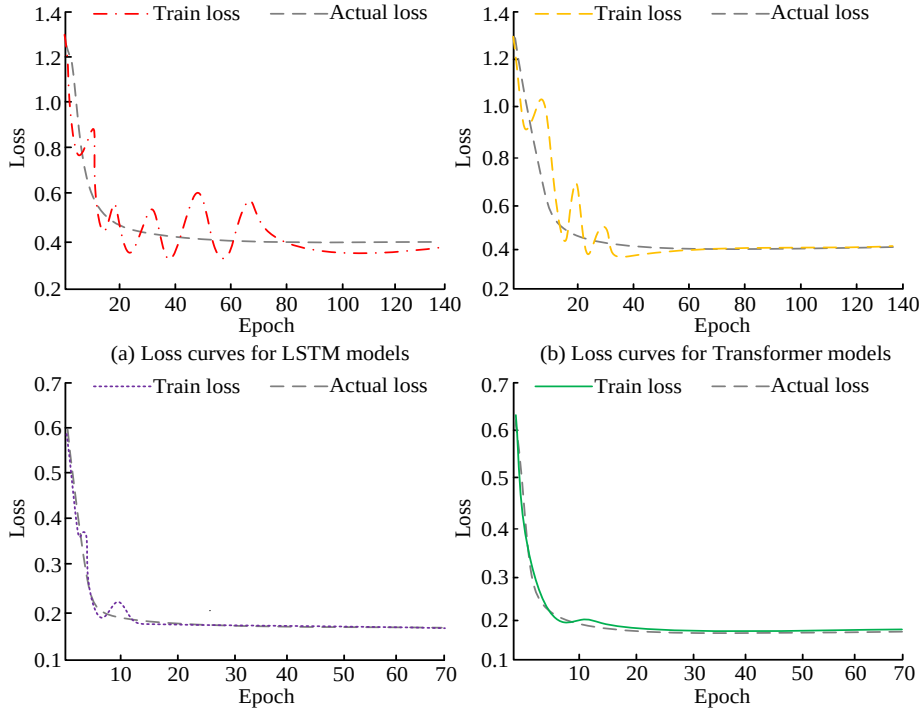
Model	Data set type				
	Newstest2020	Newstest2021	ParaCrawl	Europarl	Common Crawl
LSTM	0.78	0.77	0.71	0.66	0.69
Transformer	0.82	0.83	0.75	0.71	0.72
TCGK+Common coding	0.88	0.89	0.86	0.82	0.83
TCGK+Bi-LSTM	0.96	0.97	0.91	0.92	0.93



(a) Translation accuracy of different models in the training dataset

(b) Translation accuracy of different models in the testing dataset

Fig. 7. Translation accuracy values for the different translation models.



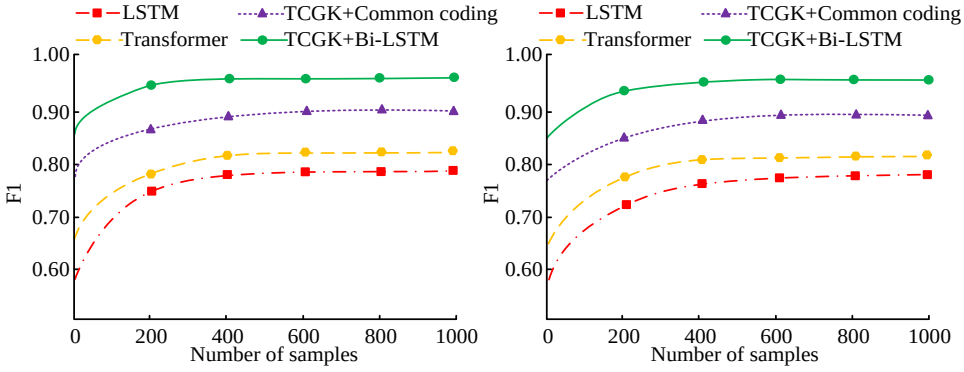
(a) Loss curves for LSTM models

(b) Loss curves for Transformer models

(c) Loss curves for TCGK+Common coding models

(d) Loss curves for TCGK+Bi-LSTM models

Fig. 8. Loss values for the different translation models.



(a) F1 score of different models in the training dataset

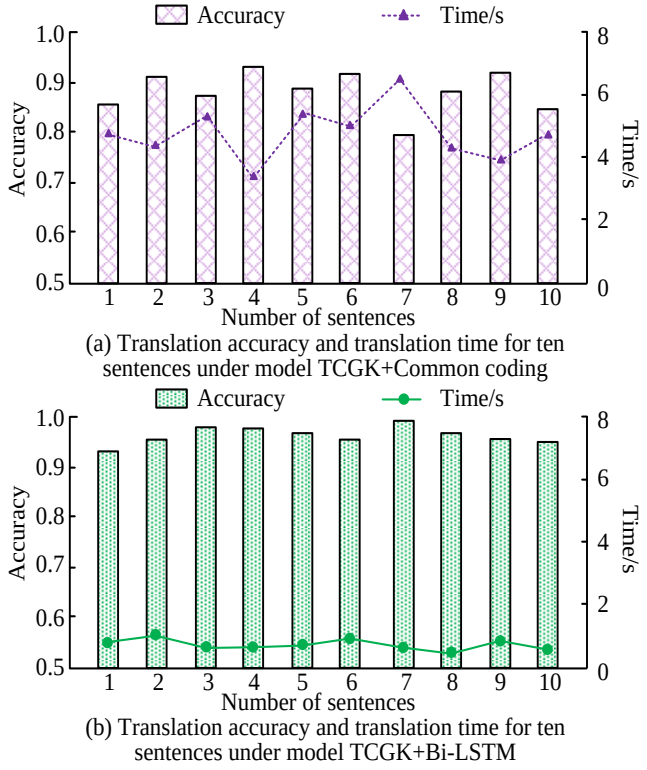
(b) F1 score of different models in the testing dataset

Fig. 9. Translation F1 values for the different translation models.

B. Analysis of the Effectiveness of the Application of Different Translation Models

The results of the analysis of the above performance indicators show that TCGK+Bi-LSTM has better performance compared with the other three translation models. TCGK+Bi-LSTM not only has better error grammar recognition accuracy values and F1 values, the change of the loss curve of this network during training is also basically the same as the actual change. To test the effectiveness of the four models in real English sentence translation, the study randomly selected 10 English utterances from a high school English textbook for testing. The translation accuracy and translation time of the two optimised UNMT models in real translation are shown in Fig. 10.

The translation accuracy and translation time of TCGK+Common coding and TCGK+Bi-LSTM in different English utterances are demonstrated in Fig. 10. Fig. 10(a) and (b) shows the translation accuracy and translation time of TCGK+Common coding and TCGK+Bi-LSTM, respectively. Comparing the translation effects of the two models in ten English utterances, it can be seen that the highest translation accuracies of TCGK+Common coding and TCGK+Bi-LSTM are 0.93 and 0.99, respectively. The shortest translation times of TCGK+Common coding and TCGK+Bi-LSTM take 3.2s and 0.5s, respectively. In addition, the translation accuracies of TCGK +Common coding model has a large change in the accuracy value during the translation process, and its translation elapsed time fluctuates more. Therefore, compared with TCGK +Common coding, TCGK +Bi-LSTM has better translation effect in practical applications.



(a) Translation accuracy and translation time for ten sentences under model TCGK+Common coding

(b) Translation accuracy and translation time for ten sentences under model TCGK+Bi-LSTM

Fig. 10. Translation accuracy and translation time of the two translation models in practice.

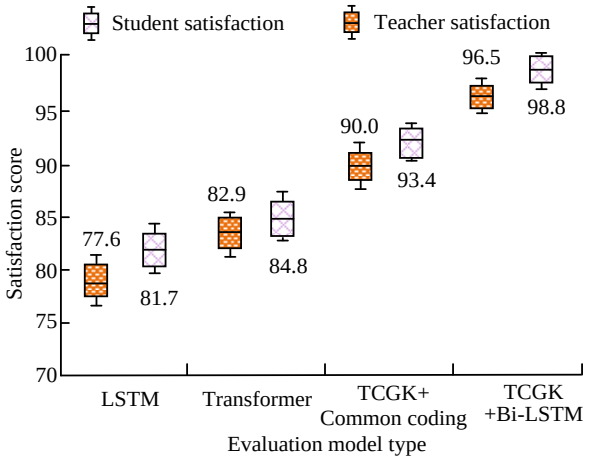


Fig. 11. Satisfaction of university teachers and students with the four translation models.

Fig. 11 shows the satisfaction scores of university students and teachers for the four translation models in practical applications. As shown in Fig. 11, the satisfaction scores of university students for the four translation models LSTM, Transformer, TCGK+Common coding, and TCGK+Bi-LSTM are 77.6, 82.9, 90.0, and 96.5, respectively. The satisfaction scores of university teachers for the four translation models LSTM, Transformer, TCGK+Common coding, and TCGK+Bi-LSTM are 81.7, 84.8, 93.4, and 98.8, respectively. In conclusion, TCGK+Bi-LSTM not only have better translation performance in practical applications, but also have higher satisfaction of university teachers and students for this model.

V. Discussion

In order to improve the accuracy and efficiency of machine translation, this research combines the fusion of Transformer structure and grammar knowledge to optimize the unsupervised neural machine translation model, and finally builds the TCGK+Bi-LSTM translation model. By comparing and analyzing the performance of various types of models as well as their practical application effects, the following discussion is derived from this research.

From the experimental results of error syntax detection accuracy, it is obvious that the TCGK+Bi-LSTM model has better error syntax detection effect compared to LSTM, Transformer, and TCGK+Common coding. The TCGK+Bi-LSTM model outperforms the other three models in terms of test accuracy and F1 value in both the training dataset and the test dataset. The reason behind the high detection accuracy and F1 value of the TCGK+Bi-LSTM model is that the combination of the deep self-attention mechanism of the Transformer structure and the Bi-LSTM network enables the model to better capture long-distance dependencies and complex syntactic structures in sentences. In addition, the TCGK+Bi-LSTM model has a better loss profile compared to LSTM, Transformer, and TCGK+Common coding, which further illustrates that the Transformer structure and the Bi-LSTM network can improve the stability of the model during the training process, which enables the model to obtain more accurate test values.

In addition, although translation accuracy is the primary index of the machine translation task, fast translation speed is also very critical in practical applications. Especially in situations where a large number of translations are required, such as online services or real-time translation applications, efficient translation speed can greatly improve the user experience. The TCGK+Bi-LSTM model also has a significant advantage in translation time compared to the other three models. This is because the introduction of the Transformer structure and the Bi-LSTM network enables the model to process the information features faster, thus achieving fast translation. Finally, the TCGK+Bi-LSTM model was also able to achieve a high level of teacher and student satisfaction in real-world applications, thus proving the value of this technique in real-world applications.

Although this study provides valuable insights into the performance of the proposed model in Chinese-English translation tasks, there are some limitations. First, the experiments were mainly conducted based on specific datasets and specific tasks, and the performance of the proposed model should be further validated on more datasets and multiple language pairs in the future. In addition, although this experiment examined the performance of several models, there are still more existing and emerging modeling approaches that deserve further exploration and comparison. Based on the current findings, future research can further examine the performance of the models on other language pairs or larger datasets. In addition, it can also explore how to further optimize the structure or parameters of the model to improve its performance in specific tasks or scenarios. In summary, this study provides valuable insights into unsupervised neural machine translation models that incorporate Transformer structural and syntactic knowledge, and provides useful directions for future research.

VI. Conclusion

This research utilizes the knowledge of Transformer structure and syntax to construct a new UNMT model that aims to improve the performance and translation accuracy of existing translation models. The results of the study show that the proposed TCGK+Bi-LSTM model model significantly outperforms the other three models in terms of detection accuracy and F1 value on both training and testing datasets. In addition, the TCGK+Bi-LSTM model exhibits higher translation accuracy and translation speed than the TCGK+Common Coding model in real translation tests involving English sentences. Finally, the TCGK+Bi-LSTM model gained high satisfaction among university teachers and students, further validating its effectiveness. Since this study mainly focused on the performance of the model in Chinese-English translation tasks, it does not have comprehensive coverage, and subsequent studies should further extend the scope of the study to examine the performance of the model on more language pairs.

References

[1] Y. Zheng, G. Li, and W. Zhang, “A New Efficient Algorithm Based on Multi-Classifiers Model for Classification,” Int. J. Uncertainty Fuzziness Knowledge Based Syst., vol. 28, no. 1, pp. 25-46, 2020.

[2] M. Heidari, S. Lakshmivarahan, S. Mirniaharikandehei, G. Danala, S. K. R. Maryada, H. Liu, and B. Zheng, “Applying a random projection algorithm to optimize machine learning model for breast lesion classification,” IEEE Trans. Biomed. Eng., vol. 68, no. 9, pp. 2764-2775, 2021.

[3] C. Faustmann, M. Bajzek, H. Hick, J. Edtmayer, and S. Walch, “System models and model classification in tribological system development,” Syst. Eng., vol. 23, no. 6, pp. 783-794, 2020.

[4] S. P. A. Alewijnse, K. Buchin, M. Buchin, S. Sijben, and M. A. Westenberg, “Model-based segmentation and classification of trajectories,” ALGORITHMICA., vol. 80, no. 8, pp. 2422-2452, 2018.

[5] C. Lv, Z. Wu, X. Wang, Z. Dan, M. Zhou, “Ethnicity classification by the 3D discrete landmarks model measure in Kendall shape space,” Pattern Recognit. Lett., vol. 129, no. 1, pp. 26-32, 2020.

[6] J. Lee, H. S. Kim, and T. Y. Kang, “Classification algorithm using halftone features of counterfeit bills and CNN,” J. Forensic Sci., vol. 67, no. 1, pp. 345-352, 2022.

[7] Y. Xiao, Y. Jin, R. Cheng, and K. Hao, “Hybrid attention-based transformer block model for distant supervision relation extraction,” Neurocomputing., vol. 470, no. 22, pp. 29-39, 2022.

[8] Z. Zhang, Y. Wang, and J. Yang, “Text-conditioned transformer for automatic pronunciation error detection,” Speech Communication, vol. 130, no. 10, pp. 55-63, 2021.

[9] Y. Li, Z. Kai, X. Hao, H. Sheng, L. Shen, and S. Yu, “Abnormal Detection Based on Graph Attention Mechanisms and Transformer,” Acta Electronica Sinica, vol. 50, no. 4, pp. 900-908, 2022.

[10] N. Li, Y. Liu, Y. Wu, S. Liu, S. Zhao, and M. Liu, “Robutrans: A robust transformer-based text-to-speech model,” AAAI., vol. 34, no. 5, pp. 8228-8235, 2020.

[11] Y. Xiao, C. Tan, Z. Fan, Q. Xu, and W. Zhu, “Joint entity and relation extraction with a hybrid transformer and reinforcement learning based model,” AAAI., vol. 34, no. 5, pp. 9314-9321, 2020.

[12] S. Ren, Z. Zhang, S. Liu, M. Zhou, and S. Ma, “Unsupervised neural machine translation with smt as posterior regularization,” AAAI., vol. 33, no. 1, pp. 241-248, 2019.

[13] M. Li, P. Y. Huang, X. Chang, J. Hu, and Y. Yang, “Hauptmann A. Video pivoting unsupervised multi-modal machine translation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 3, pp. 3918-3932, 2022.

[14] H. Sun, R. Wang, M. Utiyama, B. Marie, K. Chen, Sumita E, and Zhao T, “Unsupervised neural machine translation for similar and distant language pairs: An empirical study,” ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 20, no. 1, pp. 1-17, 2021.

[15] C. Li, and Z. Mao, “Generative adversarial network-based real-time temperature prediction model for heating stage of electric arc furnace,” Trans. Inst. Meas. Control., vol. 44, no. 8, pp. 1669-1684, 2022.

[16] M. Sajjad, F. Ramzan, M. U. G. Khan, R. Amjad, K. Mahyar, M. F. Suliman, and A. B. Saeed, “Deep convolutional generative adversarial network for Alzheimer’s disease classification using positron emission tomography (PET) and synthetic data augmentation,” Microsc. Res. Tech., vol. 84, no. 12, pp. 3023-3034, 2021.

[17] A. Bozikas, and G. Pitselis, “Incorporating crossed classification credibility into the Lee–Carter model for multi-population mortality data,” INSUR MATH ECON., vol. 93, no. 7, pp. 353-368, 2020.

[18] X. Wang, M. Cheng, J. Eaton, C. Hsieh, and S. F. Wu, “Fake node attacks on graph convolutional networks,” JCCE., vol. 1, no. 4, pp. 165-173, 2022.

[19] W. C. Chang, and I. H. R. Jiang, “iClaire: A Fast and General Layout Pattern Classification Algorithm with Clip Shifting and Centroid Recreation,” IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., vol. 39, no. 8, pp. 1662-1673, 2019.

[20] F. Ay, G. İnce, M E. Kamaşak, and K. Y. Eksi, “Classification of pulsars with Dirichlet process Gaussian mixture model,” MNRAS., vol. 493, no. 1, pp. 713-722, 2020.

Construction of a Security Defense Model for the University's Cyberspace Based on Machine Learning

Wang Bin
Shanxi Technology and Business College, Computer and Information Engineering College
Network and Information Technology Teaching and Research Office
Taiyuan City, Shanxi Province, 030036

***Abstract*—In order to ensure the security of university teachers and students using cyberspace, a machine learning based university cyberspace security defense model is constructed. Adopting a compression perception based data collection method for university cyberspace, the data information collection of university cyberspace is completed through sparse representation, compression measurement, and recovery reconstruction. Combining the advantages of Convolutional Neural Network (CNN) model in spatial feature extraction of data and Long Short Term Memory (LSTM) model in sequential feature extraction of data, extract the features of university network spatial data. After completing the multi feature dimensionality reduction processing of university network data based on the non-negative matrix decomposition algorithm, the feature dimensionality reduction processing results are input into the ConvLSTM-CNN model. After convolution calculation and integration, the security threat detection results of university network space are output. Based on the results of security threat detection, corresponding network attack defense measures are selected to ensure the security of the university's cyberspace. The experimental results show that the average attack interception rate of the model after application can reach 97.6%. It has been proven that building a model can accurately detect security threats to the university's cyberspace and achieve defense against various network attacks in different environments.**

***Keywords*—Machine learning; University's cyberspace; security defense; construction of a model; compressed sensing; non-negative matrix**

I. INTRODUCTION

The development of computer technology has accelerated the process of information popularisation. By utilising the Internet, various services such as electronic banking, online teaching, and video conferencing can be provided to the public, gradually increasing people's dependence on the Internet in their daily lives, studies, and work. Currently, the Internet is also widely used in various universities. The university network has the characteristics of spatial freedom, large distribution scale, and relatively opens, but at the same time, there are also security risks in the university network. The rapid development of the Internet has led to a large number of intrusion behaviours, posing a huge threat to the network security of universities. Therefore, achieving a comprehensive perception and defense of the university's cyberspace attack events is significant for building a complete network security defense system [1]–[3].

The research on secure and reliable defense methods for network-blocking attacks has become an urgent issue to be solved, and many relevant experts and scholars have achieved fruitful research results on this topic. Wang et al. designed a network active security defense system based on the K-means algorithm [4]. However, this method is greatly affected by attackers during maintenance and is easily exploited by them, and its defense ability needs to be improved. Oliveira et al. proposed an intelligent network attack detection and classification method for network-based intrusion detection systems [5]. However, the effectiveness of this method in network security situational awareness is relatively low, and it cannot solve some low perception network attacks. The research in [6] constructed an integrated learning model for detecting botnet attacks in IoT networks. However, this method does not take into account the changes in feature quantities that exist in university networks. Zhang et al. proposed a network security attack detection method for the network Physical system based on deep learning [7]. However, this method addresses a relatively single type of network attack, and its defense effectiveness against different network attacks is not outstanding. Tonkal et al. proposed a machine learning method with neighborhood component analysis for detecting DDoS attacks in software defined networks [8]. Gurumanapalli et al. [9] proposed a state-of-the-art generalized Feistel network-assisted Shannon condition and dynamic key based SSPN (GFS-SSPN) lightweight encryption system for IoT security to achieve high attack resilience. Majeed et al. [10] used intelligent machine learning methods to design IoT assisted drones. This method will provide an intelligent network security system that helps detect network security threats using blockchain. However, due to the processing of multi-dimensional data for network attacks, this method did not undergo dimensionality reduction processing, and the accuracy of network attack detection needs to be improved. Although the above methods have achieved attack defense, they still have issues such as inadequate defense capabilities, low perceptual efficiency, and lack of consideration for feature changes, single attack type resolution, and insufficient dimensionality reduction processing.

Machine learning aims to train computers to perform certain operations through data to train "learning" in datasets and enable computers to make improved decisions in the future. Machine learning is widely used in data mining, which involves searching for unknown or hidden patterns through a large amount of data. Machine learning can be divided into two categories: supervised and unsupervised. In the supervised

learning dataset, all training and test data are assigned a value, that is, a label, which can be a numeric or a character value. Labels of unknown data can be learnt and predicted through machine learning algorithms. In unsupervised learning, data does not have labels and machine learning algorithms are needed to find out the data patterns and predict unknown data. Therefore, in response to the university's network attacks, this paper constructs a security defense model of the university's cyberspace through machine learning, implements effective defense of attack data, and ensures the safe and stable operation of the university's cyberspace. The proposal of this model will provide innovative solutions for university cyberspace security, improving the resilience and accuracy of cyberspace security defense. The main research structure of this article is as follows:

1) Construction of a security defense model for university cyberspace: This section includes several parts: university cyberspace data collection based on compressed sensing, university cyberspace security threat detection based on machine learning, university cyberspace data feature extraction based on ConvLSTM-CNN model, and university cyberspace security defense methods.

2) Experimental analysis: This section verifies the defense effectiveness of the constructed model from different perspectives.

II. Construction of a Cyberspace Security Defense Model for Universities

While operating the security defense model for the university's cyberspace, this paper conducts network security threat defense through three stages: data collection, threat detection, and defense generation. The specific process is as follows:

A. University's Cyberspace Data Collection based on Compressed Sensing

In order to meet the actual application needs of the university's cyberspace [11], this paper implements the university's cyberspace data collection through the principle of compressed sensing. The core goal of the data acquisition method based on the principle of compressed sensing is to ensure the high accuracy of data. Therefore, on the basis of analysing the core theory of compressed sensing and combining it with the specific characteristics of the university's cyberspace monitoring data, it should study practical and feasible data collection methods for the university's cyberspace that meet the needs.

As a whole, the operating status of each node in the university's cyberspace will be interrelated and affect each other [12], which indicates that there must be a strong correlation between the data of each node and data with strong correlation must have a lot of redundancy and data containing redundancy can be expressed from another perspective through simple information. The large quantity and variety of data in universities' online space can be expressed by relatively simple information in a certain space, which shows sparsity in that space. This processing method of transforming data acquisition into information acquisition is one of the core ideas of compressed sensing. Therefore, the collected data from the university's cyberspace can be compressed using the principle of compressed sensing.

The speed and accuracy of monitoring data collection are crucial for the safe operation of a university's cyberspace. The data acquisition method based on the compressed sensing principle includes three processes: sparse representation, compressed measurement and recovery reconstruction, as shown in Fig 1. By studying and improving each process of the collection method, the accuracy of data collection can be further improved.

The starting point of this paper is the co-domain sampling of the university's network spatial data, and its research focuses on the sparse representation of data (the supporting theory of co-domain sampling). According to Fig. 1, improving the sparsity of data can improve the accuracy of data collection methods and reduce the errors introduced when restoring the original monitoring data using the feature data of monitoring quantities; Data recovery requires knowing the compressed measurement sampling value y and sparse basis matrix Ψ at the other end of the communication channel. The amount of data transmitted between the two ends, y and Ψ, determines the speed of the data collection method.

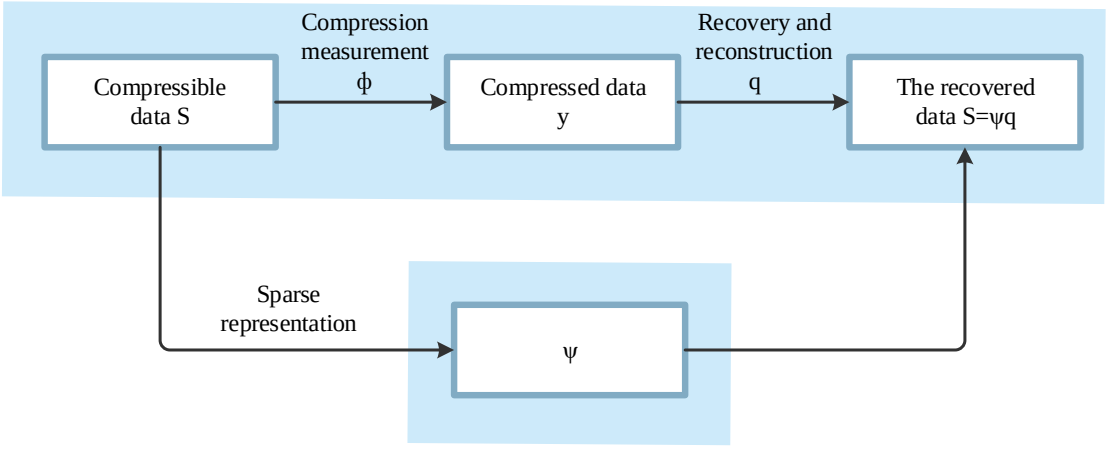


Fig. 1. Compressed sensing flow of monitoring data.

1) Theoretical basis of compression perception algorithm: In compressed sensing [13], the K-means Singular Value Decomposition (K-SVD) dictionary learning algorithm is an adaptive data sparse method. Due to the fact that data collected in a university's cyberspace can always be sparsely represented in an unknown space, and the excellent adaptive characteristics of the K-SVD dictionary enable it to adapt to the collected data in the university's cyberspace through dictionary learning algorithms, so as to effectively find a sparse space suitable for university's cyberspace data and achieve sparse representation of university's cyberspace collected data.

The mathematical model of the K-SVD dictionary learning algorithm is as follows:

$$\begin{cases} D, \theta = argmin\{\|s - D\theta\|_r^2\} \\ s.t. \forall i, \|\theta_i\|_0 \le T_0 \end{cases} \quad (1)$$

where, s represents monitoring data containing different data types within a sampling period; θ is sparse coefficient vector; T_0 is the upper limit of the number of non-zero components in sparse coefficients; the dictionary is $D = [d_1, d_2, \cdots, d_k] \in R^{n \times k}$, where each column d_k represents an atom; r is the norm.

The data s parasitisation steps based on the K-SVD dictionary learning algorithm are as follows.

Initialise the dictionary. Randomly select K data samples from the general data samples as the atoms of dictionary D. This paper takes the monitoring data S as the initial atom of D, and its standardisation process is as follows:

$$d_i^* = \left| \frac{d_i}{\sqrt{\sum_{i=1}^{N} d_i 2}} \right| \quad (2)$$

where, N is any length of compressible discrete real value data.

a) Sparse encoding. Under the condition that the dictionary D is fixed, it can solve the optimisation model of the above equation using the Orthogonal Matching Pursuit(OMP) to obtain the sparse coefficient θ′ of the collected data S.

b) Dictionary update. Update the dictionary column by column. When updating the k -th column d_k of the dictionary, make E_k the error generated by removing d_k from the data, that is:

$$E_k = s - \sum_{j \ne k} d_j \theta_T^j \quad (3)$$

In the equation, θ_T is the jth line of θ. Then perform singular value decomposition on the error to obtain the updated d_k and θ.

c) Repetitive sparse encoding and dictionary updates. If the termination condition is met, the output data is based on the adaptive sparse dictionary D of the K-SVD dictionary learning algorithm.

In the application of data collection in the university's cyberspace, based on the K-SVD dictionary learning algorithm, every atom of the optimal dictionary D of signal S is identical, and the coefficient vector θ decomposed on this dictionary is 1-sparse, ensuring the high sparsity of data S. This shows the feasibility and progressiveness of the K-SVD dictionary learning algorithm applied to a sparse representation of monitoring data in university network space.

2) Compression measurement: For the compression measurement of the university's cyberspace data, data x is compressed to obtain a measurement vector y with a length of $M(M \times N)$, which is expressed as:

$$y = \Phi x \quad (4)$$

Compression measurement no longer uses the method of measuring the university's network spatial data X itself first and then compressing it. Instead, a measurement matrix $\Phi \in R^{M \times N}$ is used to directly compress and measure data X, converting data sampling into sampling of M projections on the measurement matrix Φ.

3) Compression reconstruction: For the restoration and reconstruction of the university's cyberspace data, it is the process of restoring the original data x from vector y. In this process, the measurement vector y is represented as:

$$y = \Theta\theta \quad (5)$$

In the equation, Θ is the perception matrix and satisfies the isometric constraint criterion. The goal of the university's cyberspace data recovery and reconstruction is to find the optimal sparse coefficient vector θ, which can be expressed as an optimisation problem under the l_1 norm, namely:

$$\begin{cases} \theta = \arg\min \|\theta\|_1 \\ s.t. \quad y = \Theta\theta \end{cases} \quad (6)$$

Finally, the optimal recovery value x of the original data can be obtained. And then, the space data collection of the university's network using compressed sensing is realised. Based on the above compression perception, compression measurement, and compression reconstruction processes, data collection and output of university cyberspace security threat detection results are shown in Table I:

TABLE I. SECURITY THREAT DETECTION RESULTS OF UNIVERSITY CYBERSPACE

Connect IP	Time stamp	Detection result
192.168.1.2	2022-10-05 09:30:20	Normal
192.168.1.4	2022-10-05 09:35:10	DDoS attacks
192.168.1.5	2022-10-05 09:40:15	DNS hijacking
192.168.1.7	2022-10-05 09:45:30	Normal
192.168.1.2	2022-10-05 09:50:40	Normal

B. Machine Learning-based Threat Detection for University's Cyberspace Security

1) Theoretical basis of ConvLSTM-CNN model: A CNN is a deep feedforward neural network with a large number of convolution calculation processes. Different from the conventional neural network model, the neurons in each layer of CNN are arranged three-dimensionally. A two-dimensional data's length and width are the neurons' length and width, while the depth represents the third dimension used to activate the data volume. It is precisely because of this characteristic of the CNN model that it can better obtain the spatial features of university network data [14].

At present, CNN models have been widely applied in the field of cyberspace security in universities, such as intrusion detection systems and situational awareness. The structure diagram of a typical CNN model is shown in Fig. 2. Convolutional neural networks are typically composed of input, hidden, and output layers, with hidden layers typically including convolutional, pooling, and fully connected layers.

The data is first input in the input layer, and then the university's network data perceived in Section II (A) is preprocessed. By normalising the data, it is then passed into the hidden layer for calculation.

The first layer to enter is the convolutional layer, which is the core of the CNN model structure. It has multiple convolutional kernels inside and can extract features from the input data. After the convolutional layer, there is usually a pooling layer, which appears alternately in the convolutional neural network. Each convolutional layer corresponds to a pooling layer. The fully connected layer in convolutional neural networks has a similar network structure to conventional neural networks, where each neuron structure constructs connections with all neurons in the previous layer, playing the role of a "classifier" in the entire network, integrating and mapping local sample information into the corresponding space.

Compared with other neural network models, CNN has better automatic feature extraction ability [15] and is widely used in computer vision, natural language processing, cyberspace security and other directions. It has made much progress and can be used for attack classification, image recognition, target segmentation, and other fields. In this paper, we utilised the advantages of CNN models in the perceptual extraction of spatial dimension features and optimised their structure to make them more suitable for modelling traffic time series in the field of cyberspace security situational awareness in universities.

Although the CNN model has obvious advantages in data space feature extraction, it is not good at feature extraction for sequence data such as network traffic. Therefore, it needs to be optimised based on fully utilising its advantages.

LSTM model is a special Recurrent Neural Network (RNN) model which has the ability to store the long-term state of data. Sometimes, due to the long data sequence length, the training process of ordinary RNN models is prone to the problem of vanishing or exploding gradients. The LSTM model performs better in training longer sequence data than this.

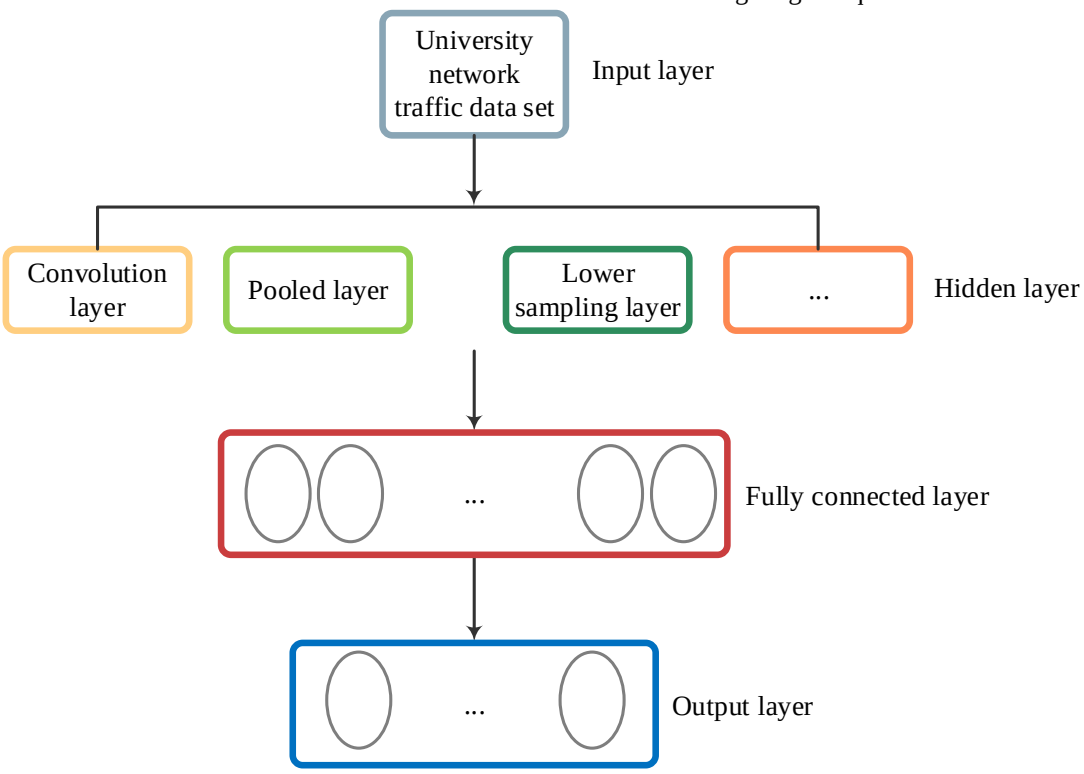


Fig. 2. Schematic diagram of convolutional neural network model.

In order to control the cell state, three gates are designed in the LSTM model to control or delete the added unit state information, namely the forgetting gate, memory gate, and output gate:

a) Forgetting gate: It used to forget or discard part of attribute information, accept a short-term memory output from the previous unit module, and decide which part to retain and forget.

b) Memory gate: It determines the information stored in the cellular state. For the attribute relationships discarded in the forgetting gate, it can find and fill in the corresponding new attribute information in this unit module to supplement the attribute information discarded by the forgetting gate.

c) Output gate: It determines the output value based on the cell state.

The working principle of the LSTM model is shown in Eq. (7):

$$\begin{cases} i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \\ f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \\ c_t = f_t \circ c_{t-1} + i_t \circ tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co} \circ c_t + b_o) \\ h_t = o_t \circ tanh(c_t) \end{cases} \quad (7)$$

where, i_t、f_t、c_t、o_t、h_t represents input gate, forgetting gate, cell state, output gate, and transmission state; $\circ$ is the multiplication of the corresponding elements of the matrix, also known as Hadamard product, and σ is the sigmoid function; W is the weight matrix, and b is the bias top.

The LSTM model, especially for high-dimensional time series data such as network data, can effectively analyse the logical relationships between input features and the complex temporal correlations between information. Meanwhile, compared to the CNN model, the LSTM model can better simulate human thinking patterns and cognitive processes and exhibits better processing ability for complex tasks closely related to time series. Therefore, this paper will use the temporal features extracted from LSTM network data combined with the spatial features extracted from CNN network data as the result of feature extraction for university network data.

2) Multi-feature dimensionality reduction processing of university's network data based on non-negative matrix decomposition algorithm

Due to the multi-dimensional problem of obtaining the features of the university's network data in Section II (B) (1), in order to improve the accuracy of threat detection in the university's network space security, a non-negative matrix decomposition algorithm is used to perform multi-feature dimensionality reduction on university network data [16]. The non-negative matrix decomposition algorithm has the advantages of simple decomposition implementation and small space occupation. The specific process of multi-feature dimensionality reduction for university network data based on the non-negative matrix decomposition algorithm is as follows:

Fuzzy classification of university network data features is carried out, and the initial value of the non-negative matrix decomposition base matrix to be composed of the classification centroid is set [17]. In contrast, the non-negative matrix decomposition rank is the number of classifications. The feature set of university network data is set to $\lambda = [\lambda_1, \lambda_1, \cdots, \lambda_n]$, where n represents the number of samples in the feature set of university network data. After fuzzy l classification, the set is $F = [F_1, F_2, \ldots, F_l]$, where l represents the number of feature classifications of university network data. While the membership degree of the university's network data features to the fuzzy vector is expressed by U_{ij}, the membership degree matrix representation is expressed as $U = [U_{ij}]_{l \times n}$. It should be noted that U_{ij} meets $\sum_{i=1}^{l} U_{ij} = 1$, and the membership degree update function is specified as:

$$U_{ij} = \begin{cases} 0 & \|\lambda_j - z_i\|_2 > 0 \\ 1 & \|\lambda_j - z_i\|_2 = 0 \\ \left[\sum_{a=1}^{l} (\|\lambda_j - z_i\|_2 / \|\lambda_j - z_a\|_2)^{2/(\omega-1)}\right]^{-1} & \|\lambda_j - z_i\|_2 < 0 \end{cases} \quad (8)$$

where, λ_j represents the jth university network data feature; z_i represents the i-th clustering centre; z_a represents the initial clustering centre; ω represents the fuzzy weighted index.

Thus, the cluster centre set is determined as follows:

$$Z_i = \frac{\sum_{j=1}^{n} (U_{ij})^{\omega} \lambda_j}{\sum_{j=1}^{n} (U_{ij})^{\omega}} \quad (9)$$

The objective function is:

$$D = \sum_{i=1}^{l} \sum_{j=1}^{n} (U_{ij})^{\omega} \|\lambda_j - z_i\|_2^2 \quad (10)$$

If the above equation calculates the result $D < \tau$ (threshold), then the iteration stops, and the initial value of the base matrix P in non-negative matrix decomposition is set to be composed of the cluster centre set z_i.

According to the non-negative matrix decomposition algorithm, the feature matrix $\lambda_{m \times n}$ of university network data is decomposed into two non-negative matrices $Q_{m \times n}$ and $P_{m \times n}$, and the product between them infinitely approximates the original non-negative matrix, namely $P_{m \times n} \approx Q_{m \times n}$. Where, $P_{m \times n}$ represents the base matrix; $Q_{m \times n}$ represents the coefficient matrix, using minimising the remaining Frobenius as the objective function, and the expression is:

$$\min_{P,Q} F(P, Q) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\lambda_{ij} - (PQ)_{ij})^2 \quad (11)$$

When, $\lambda = PQ$, the value is 0.

Before the convergence of the objective function, fuzzy classification is used to obtain the centroid initialisation basis matrix, and the matrices P and Q are alternately updated. The matrix iteration rules are expressed as follows:

$$\begin{cases} P_{ia} \leftarrow P_{ia} \sum_j \frac{\lambda_{ij}}{(PQ)_{ij}} Q_{ai} \\ P_{ia} \leftarrow \frac{P_{ia}}{\sum_i P_{ia}} \\ Q_{aj} \leftarrow Q_{aj} \sum_i \frac{\lambda_{ij}}{(PQ)_{ij}} \end{cases} \quad (12)$$

After repeated iterations, the base matrix and coefficient matrix are finally obtained as P_t and Q_t. The coefficient matrix Q_t is used to replace the original sample matrix, thus achieving multi-feature dimensionality reduction of university network data.

3) ConvLSTM-CNN-based threat detection model for university's cyberspace attacks: As mentioned earlier, traditional CNN models cannot fully consider the sequence characteristics of data, and using LSTM models alone cannot fully consider the spatial characteristics of data and the correlation relationships between various features. Therefore, this paper intends to introduce the idea of the ConvLSTM model to model the problem of situational awareness of cyberspace security threats in universities from the dimensions of time and space features. Considering the limited computational performance of the ConvLSTM model and the higher computational cost and longer training time of each layer compared to convolutional layers, this paper proposes a ConvLSTM-optimized CNN model. It applies to the field of cyberspace security threat situational awareness in universities.

ConvLSTM was first proposed in 2015 and has shown good performance in predicting spatiotemporal data with both temporal and spatial features. This model extends the fully connected LSTM to have a convolutional structure in both input-to-state and state-to-state transitions, as shown in Fig. 3, which is the ConvLSTM network architecture diagram.

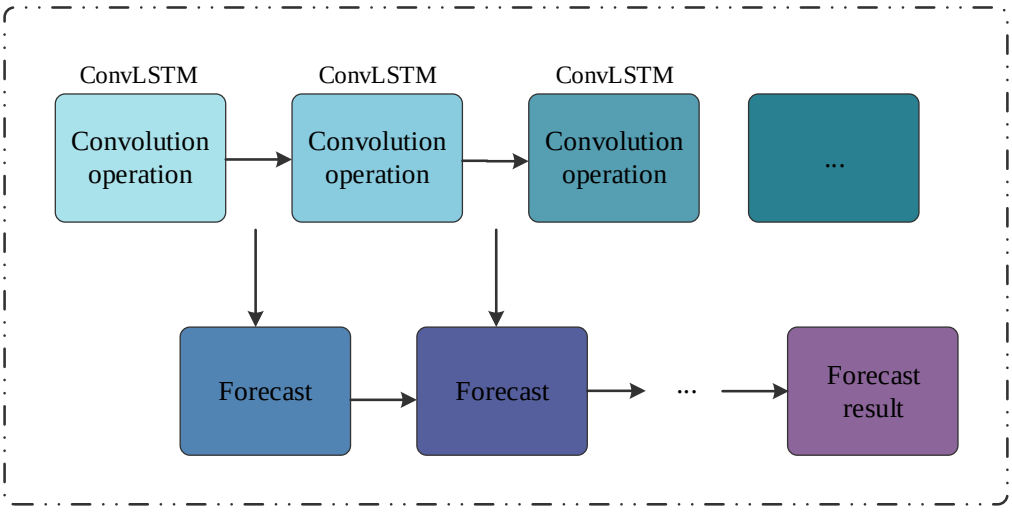


Fig. 3. Internal architecture of ConvLSTM model.

The working principle of ConvLSTM is shown in Eq. (13):

$$\begin{cases} i_t = \sigma(W_{xi} * \chi_t + W_{hi} * H_{t-1} + W_{ci} \circ c_{t-1} + b_i) \\ f_t = \sigma(W_{xf} * \chi_t + W_{hf} * H_{t-1} + W_{cf} \circ c_{t-1} + b_f) \\ c_t = f_t \circ c_{t-1} + i_t \circ tanh(W_{xc} * \chi_t + W_{hc} * H_{t-1} + b_c) \\ o_t = \sigma(W_{xo} * \chi_t + W_{ho} * H_{t-1} + W_{co} \circ c_t + b_o) \\ h_t = o_t \circ tanh(c_t) \end{cases} \quad (13)$$

In the equation, i_t, f_t, c_t, o_t, H_t represents input gate, forgetting gate, cell state, output gate, and transmission state, respectively; $\circ$ is the multiplication of the corresponding elements of the matrix, also known as Hadamard product, and σ is the sigmoid function; W is the weight matrix, and b is the bias top. Unlike LSTM, where χ, c, H, i, f, o represents three dimensions, f and o represent the information of data rows and columns.

For the classification and prediction requirements of spatiotemporal sequence data, such as university's cyberspace data, the ConvLSTM model has the memory storage and computing ability for long-time sequences [18]. Therefore, this paper uses a combination of the ConvLSTM and CNN models to construct the ConvLSTM-CNN model, with the specific structure shown in Fig. 4.

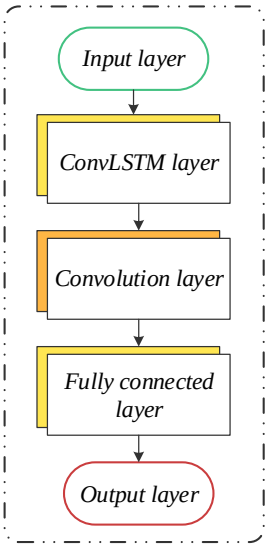


Fig. 4. Structure diagram of ConvLSTM-CNN model.

a) Input layer: First, encode the labels of input data in the input layer, convert them from textual data to continuous numerical variables, fit the data, and then perform standardisation and normalisation processing. Next, convert the data into a specific dimensional matrix that can be accepted by the next ConvLSTM layer.

b) ConvLSTM layer: The first layer to enter after the input layer is the ConvLSTM layer [19], which identifies the time dimension and spatial dimension of data, captures the spatiotemporal characteristics of data, and has a long memory function for solving the problem of serial data modelling with a long time. At the same time, unlike the feedforward fully connected form between the input and each gate in the LSTM model, the model's weight is convolutionally calculated and connected using the convolutional form, which allows for the acquisition of spatial features of the data while considering temporal features.

c) Convolutional layer: After capturing the spatiotemporal features of the data, it enters the convolutional layer for calculation. By utilising the typical sparse connection characteristics of the convolutional layer, it can effectively extract data features while shortening the convergence time of the model [20]. In forward fully connected neural networks, every element in the matrix needs to participate in computation, which undoubtedly increases the computational time and complexity of the algorithm. In this layer, the parameters of each convolutional kernel can be shared with each other, allowing only one operation to be performed on the same parameters in the university network model, which not only improves the runtime but also preserves its advantages in spatial feature extraction [21]. In the design of this model, the number of convolutional layers is set to 3, with a gradually decreasing number of convolutional kernels.

d) Fully connected layer: Located at the last layer of the entire network structure, it is used to integrate the feature information output from the previous layer.

e) Output layer: make classification and probability mapping for the model output and map it to the corresponding labels.

4) Optimisation parameter selection: In a neural network, the key work to enable it to solve the nonlinear problem of a sparse matrix, such as a university's cyberspace security situation awareness, is how to select and use appropriate activation functions, introduce nonlinear factors, retain and map the features of some activated neurons, and remove redundant and irrelevant features in the data, so as to enable neural networks to have hierarchical nonlinear mapping learning capabilities that linear models do not possess [22]. After completing the construction of the neural network model, it is necessary to search for the optimal solution for the model and select a suitable optimiser. Usually, the idea of gradient descent is used to calculate the loss and gradient of the model in order to obtain the parameters of each layer of the university's cyberspace security threat detection model and generate the optimal solution.

a) Activation function: In the structure and calculation process of ConvLSTM, similar to LSTM, there is a tanh operation in the related calculations of its input and output gates, which is used to generate candidate memories. As shown in Eq. (14), the calculation equation for the tanh function is shown. It can be seen that the tanh function requires a fourth power operation during the calculation process, and more power operations may be performed during its backpropagation differentiation, which undoubtedly increases the computational complexity of the ConvLSTM layer. Meanwhile, the tanh function may also cause gradient saturation during model training, which is a major reason for its low computational efficiency.

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{14}$$

where, e is a function.

In order to reduce the computational load of the security threat detection model of the university's cyberspace and solve the gradient saturation problem in ConvLSTM, this paper introduces the linear rectification function (ReLU) as the activation function in ConvLSTM-CNN and CNN layers and Eq. (15) is the calculation equation of ReLU function.

$$\mathrm{ReLU}(x) = max(0, x) \tag{15}$$

In the full connection layer, the activation function [23] is not used by default; that is, each layer only performs simple matrix multiplication. In this layer, the design and use of appropriate activation functions can play the role of adding nonlinear factors to neurons and the entire network structure, making neurons approach nonlinear networks. Therefore, this paper uses Softmax as the activation function in the full connection layer to obtain the probability value of each sample corresponding to the input so as to achieve the purpose of data classification. The calculation equation for the sigmoid function is shown in Eq. (16).

$$\text{soft}\,max(x) = \frac{e^{x}}{\sum_{x=1}^{T} e^{x}} \tag{16}$$

where, T is the number of types of samples.

b) Optimiser: After the establishment of the threat detection model for cyberspace security in basic colleges and universities, multiple iterations is needed to continuously optimise the model, find appropriate parameters, reduce the value of the loss function as much as possible, and improve the accuracy of the threat detection model for cyberspace security in colleges and universities. In order to adjust the parameters of the university's cyberspace security threat detection model during the training process and find the optimal solution, we introduced an optimiser to calculate and update the network parameters that could affect the training and output of the university's cyberspace security threat detection model, so that the results of the university's cyberspace security threat detection model can approximate or achieve the best [24], [25]. In this paper, the Adaptive Time Estimation Method (Adam) is introduced to calculate the adaptive learning rate of each parameter in the university's cyberspace security threat detection model. Adam algorithm

has higher convergence speed and more efficient computational efficiency than other adaptive learning rate algorithms. At the same time, it also has a lower occupancy rate of system computing resources and is more suitable for situations such as sparse gradients or high gradient noise. The Adam algorithm has made improvements such as bias correction and gradient algorithm optimisation on the basis of AdaGrad and RMSProp. Still, it retains its advantage of dynamically and adaptively optimising the learning rate. The calculation process of the Adam algorithm is shown in Eq. (17), which includes storing the exponential decay average value v_t of the historical square gradient and maintaining the exponential decay average value m_t of the historical gradient.

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \quad (18)$$

where, β is the learning rate, and g_t is the fitness function.

The calculation process of the Adam algorithm after introducing the deviation correction step is shown in Eq. (18) to reduce the impact of deviation on the early training of the university's cyberspace security threat detection model.

$$m_t' = \frac{m_t}{1-\beta_1^t}$$
$$v_t' = \frac{v_t}{1-\beta_2^t} \quad (18)$$

After correcting the deviation, as shown in Equation (19), it can adaptively calculate the updated step size η_t from the perspectives of gradient mean m_t' and gradient square v_t'.

$$\eta_t = \eta_{t-1} - \alpha * m_t / (\sqrt{v_t} + \varepsilon) \quad (19)$$

where, α is the Lagrange multiplier and ε is the relaxation variable.

C. Defense Methods for Cyberspace Security in Universities

According to the ConvLSTM-CNN threat detection model in Section II (B), security threats in the university's cyberspace are detected, and the university's cyberspace security defense is carried out through the following methods.

1) Security defense against DNS attack threats: For example, the Domain Name System (DNS) denial of service attack threat is prevented in the network blocking attack detected by the ConvLSTM-CNN threat detection model. DNS denial of service attack threat security defense requires a detailed record of the number of DNS requests and hit rate in the time window. Based on the network attack threat detected by ConvLSTM-CNN, threat detection is performed on DNS attack data [26]. The attack data is divided into a blacklist. Normal data is divided into a whitelist and efficient network space inside the socket is used to transfer data information from the two lists so that it is assigned to filtering work, to filter attacks and redirect them, and implement constraints on the number of DNS requests that each source IP passes through per unit time to achieve security defense. The specific approach is as follows:

a) Prohibit DNS requests from the blacklist;

b) For whitelist lists, IP can be allowed to pass through a large number of DNS requests;

c) For non-whitelist lists, it constrains the number of DNS requests that IP passes through, where the allowed non-frequent DNS requests per unit time (1-2 requests) can be limited by using the *tc* tool. To provide commonly used domain name responses on behalf of servers in cases where there are a large number of DNS requests and to randomly discard some requests that are not included in the DNS frequency list, filters can be set in front of the attacked server. The security defense model steps for network blocking attack threats are shown in Fig. 5.

A filter function is mounted on the NetFilter framework chain of the Linux kernel to discard DNS requests with the source IP on the blacklist. In university's cyberspace data packets, using this framework can achieve the application of custom behaviour. Blocking IP addresses on the blacklist improves filtering efficiency and compensates for the missing functionality of the original DNS protocol [27].

The created queue is set on the output port and constraints university network traffic reasonably based on routing selection. Filters, queues, and classifications together form the *tc* tool. The following is the operation process:

a) Create CBQ queues and apply them to network physical devices;

b) Create the classification in the CBQ queue;

c) Create filters for each category, and the filters need to be created based on the routing;

d) Routing tables can be created by matching filters.

Usually, only one queue needs to be created, with each queue containing a root classification that includes subcategories. The smaller the classification number is, the more effective it becomes. If a certain classification rule is met, the data packet can be sent using that classification, and subsequent classifications will lose effectiveness. This completes the security defense against DNS attacks that threaten the cyberspace of universities.

2) Security defense against DDoS attack threats: The essence of security defense is how to defend against DDoS attack threats detected through the ConvLSTM-CNN threat detection model [28]. In SDN networks, OpenFlow flow tables can be used for implementation. As shown in Fig. 6, in the SDN network, when an attacker controls the botnet to launch a DDoS attack against the server, the attacking university's network space traffic first reaches the edge switch S1 and S1 detects whether there is a flow table matching the flow label in the university's network traffic space. If there is one, it will be forwarded directly according to the flow table rules. If not, it will upload the flow information to the controller and wait for the controller to make a decision. After receiving the flow information, the controller executes the

ConvLSTM-CNN threat detection model to detect the traffic in the university's cyberspace and determines whether the flow is DDoS attack traffic. If it is not, it calculates forwarding and sends a flow table containing routing information to the switch. The switch forwards according to the flow table rules; If so, the controller issues a table containing discarded instruction flows to the switch, and then the switch will discard this DDoS attack. In order to avoid the overflow of switch flow table space caused by excessive traffic in the university network space, a soft timeout time for the university's cyberspace traffic is set. Since DDoS attacks on university network space traffic are usually instantaneous, the soft timeout time is set to 1 second, which means that the flow table will delete itself if it is not matched for more than a second. This completes the security defense against DDoS attacks that threaten the cyberspace of universities.

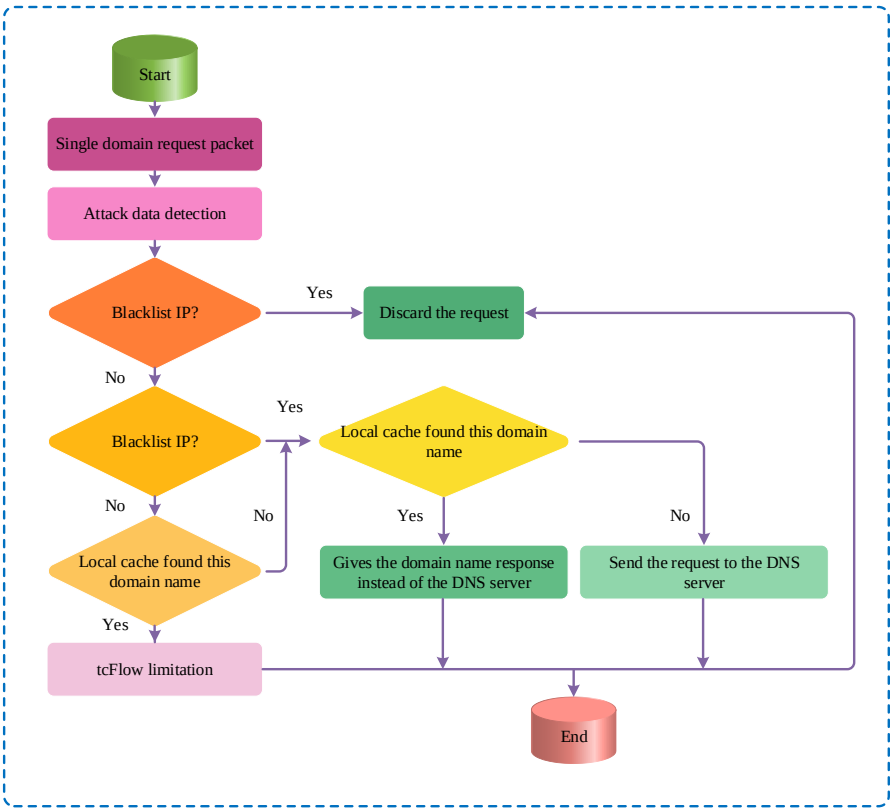


Fig. 5. Steps for defending against network blocking attacks.

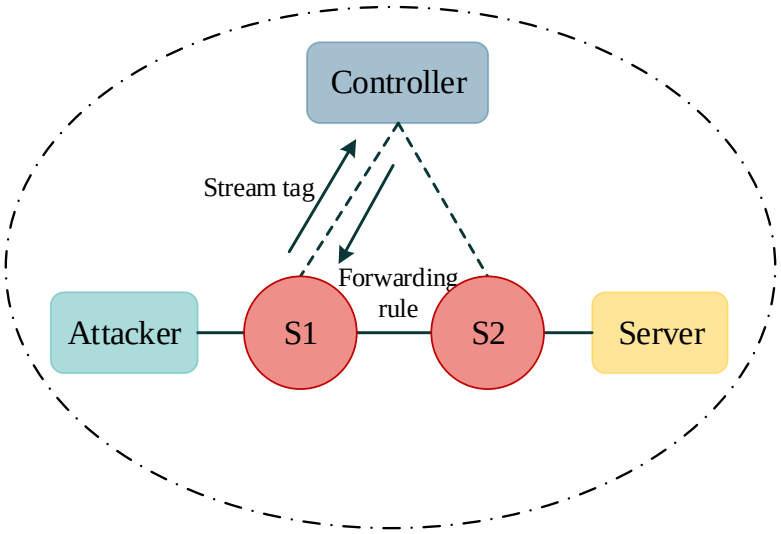


Fig. 6. Network diagram.

III. Experimental Analysis

Taking a certain university's cyberspace as the experimental object, based on big data analysis technology operating parameters, the control centre selects the Mininet+POX platform, simulates the virtual university's cyberspace environment through Mininet, follows the custom remote network interaction process, and runs the POX controller within the platform as the core selection. Based on the connectivity of remote university's cyberspace attacks, six switches are used to connect with each other. Among them, one switch is used to connect to the remote subnet, and another attack host is prepared. The other four are the teaching building, library, experimental building, and cafeteria of a certain university, and the topology of the university's network space attack is completed based on the remaining operating bandwidth of the switch. Using the built-in traffic generation tool Trafgen to attack university network space, the attack host's attack traffic is generated. Based on the attack traffic source IP address and source port number parameters, the attack content in the host file is continuously configured, and iPerf3 is used to generate and configured attack packet-related parameters. The environmental design scheme is described in Fig. 7.

After building the attack topology structure, it simulates the running environment of the university's big data analysis technology using the background traffic in the attack package to call Scapy on the normal host and complete the writing of the main function code of big data analysis technology. Based on the instructions sent and the actual response interactions generated, the traffic in the network space of attacking universities within the topology structure is resolved into multiple protocols that support attack instructions, which can control the attacking host, form identity mapping, generate mapping requests with attacking instructions, extract query ports generated by port access, and use them as fixed identity numerical markers to debug the remote attack process between the attacking host and other hosts.

To ensure the effectiveness of the experiment, the parameters of the proposed method are set as follows:

CNN model parameters: Convolutional kernel size is 3 × 3. The step size is 1, and the activation function is ReLU.

LSTM model parameters: The hidden layer size is 256, and the activation function is tanh.

Non negative matrix decomposition algorithm parameter: Select to retain the first 50 features as the dimensionality reduction result.

CNN model: Convolutional layer 1 has 32 convolutional kernels, with a kernel size of 3x3 and a step size of 1. Convolutional layer 2 has 64 convolutional kernels, with a kernel size of 3×3 and a step size of 1. The activation function is ReLU.

LSTM model: The LSTM layer has 64 hidden layer units.

Fully connected layer: The output layer has two neurons, representing the categories of normal and attack.

Loss function: Using cross entropy loss function.

The selected comparison methods use the parameters during their testing period, and will not be listed here.

The DDoS attack and DNS attack are used to verify the security defense capability of the model in this paper against the network space security of colleges and universities. At the same time, the network active security defense model based on the K-means algorithm in reference [4], the network security defense model based on the improved particle swarm optimisation algorithm in reference [5], the network security defense model oriented to digital transformation based on endogenous security framework in reference [6], the network security defense decision-making model based on timing game in reference [7], and the network data resource security defense model based on cloud computing in reference [8] are tested as a comparison method for the model in this paper. The test results are shown in Table II.

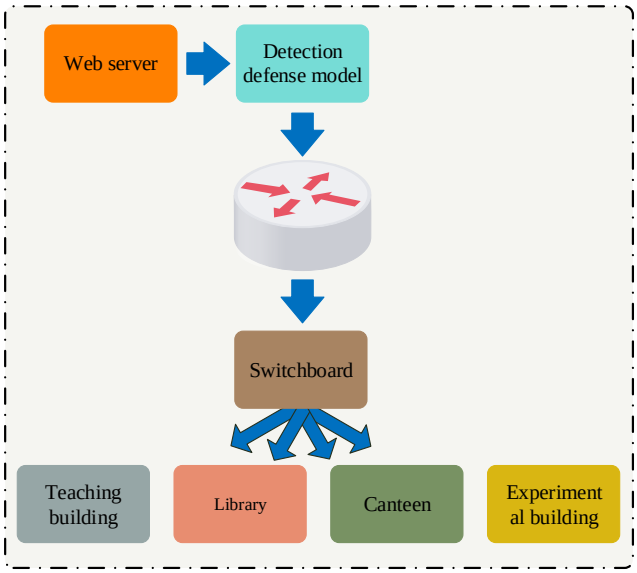


Fig. 7. Design of security defense model environment for cyberspace attacks in universities.

TABLE II. SECURITY DEFENSE RESULTS OF DIFFERENT ENVIRONMENTS IN UNIVERSITY NETWORK SPACE

Host		Teaching building	Library	Canteen	Experimental building
Model in this paper	DDoS attack	Success	Success	Success	Success
	DNS attack	Success	Success	Success	Success
Reference [4] Model	DDoS attack	Failure	Success	Success	Failure
	DNS attack	Failure	Failure	Failure	Failure
Reference [5] Model	DDoS attack	Failure	Failure	Success	Failure
	DNS attack	Success	Failure	Failure	Failure
Reference [6] Model	DDoS attack	Failure	Failure	Failure	Success
	DNS attack	Success	Success	Success	Success
Reference [7] Model	DDoS attack	Failure	Success	Success	Failure
	DNS attack	Success	Success	Success	Failure
Reference [8] Model	DDoS attack	Success	Success	Success	Failure
	DNS attack	Success	Failure	Failure	Failure

From Table II, it can be seen that in the simulated cyberspace attack environment outside a certain university's teaching building, library, cafeteria, and experimental building, the models in reference [4] and the reference [5] only successfully defended two times when they attacked DDoS and DNS cyberspace eight times, while the other six times failed. The defense effect of the models in reference [6], reference [7], and reference [8] are better than the first two models, the defense against threat attacks is above 50%, but it has not achieved the desired security defense effect. And through the model in this paper, the defense against DDoS and DNS network space attacks is all successful, indicating that the model in this paper can achieve security defense in network space under different network attacks in different environments and has strong security defense capabilities.

The experiment sets the attack host traffic value between 200-1200MB to ensure the accuracy of the test results. In this attack traffic mode, the fixed controller has 18 running windows. Based on the running window values and attack traffic parameters, it completes the statistical work of the actual intercepted traffic data of the defense software. Then it calculates the attack interception rate of the defense software.

Equation $F = (AR + TN) \div (TA + TN + FA + AR) \times 100\%$ is used to represent numerical relationships. Among them, F represents the interception rate of defense software attacks; AR represents the number of attack data correctly intercepted by the defense software, TN indicates successful tagging of attack data traffic; TA represents the number of intercepted windows supported by the controller, FA represents the set attack traffic value; traffic interception rate statistics for defense software attacks can be conducted based on different traffic attacks. The results are shown in Table III.

Based on the numerical relationship between different attack interception values, statistical analysis is conducted on the attack interception rates of the models in reference [4], reference [5], reference [6], reference [7], reference [8], and the defense software in this paper. The average is read for the attack interception rates under different supply scales. From Table III, it can be seen that the average attack interception rates of the defense software in reference [5] model and reference [6] model are 77.7% and 77.2%, respectively, indicating weak defense capabilities against university cyberspace security; The average attack interception rates of the defense software for reference [4] model, reference [7] model, and reference [8] model are 85.2%, 85.4%, and 87.9%, respectively, indicating an improvement in defense capabilities; The average attack interception rate of the defense software designed in this model is 97.6%, which is the strongest defense capability for university cyberspace security compared to the other five models.

Based on two types of network attack methods, DDoS attack and DNS attack, the congestion status of the university's cyberspace is tested for six attack defense models. The network space attack defense models of reference [4], reference [5], reference [6], reference [7], and reference [8] are used as comparative models. Fig. 8 shows the network congestion rate results of the six models for the university's cyberspace attack defense.

TABLE III. INTERCEPTION RATES OF UNIVERSITY'S CYBERSPACE ATTACKS USING DIFFERENT MODELS (%)

Host attack traffic	Reference [4] Model	Reference [5] Model	Reference [6] Model	Reference [7] Model	Reference [8] Model	Model in this paper
200	82.1	79.2	75.6	82.6	89.2	97.6
400	86.2	77.5	76.8	84.3	86.9	98.9
600	85.6	79.4	79.5	85.9	88.4	96.4
800	87.1	75.3	77.9	87.5	86.1	97.7
1000	85.2	76.1	78.2	86.7	89.4	98.3
1200	84.9	78.7	75.3	85.3	87.5	96.9

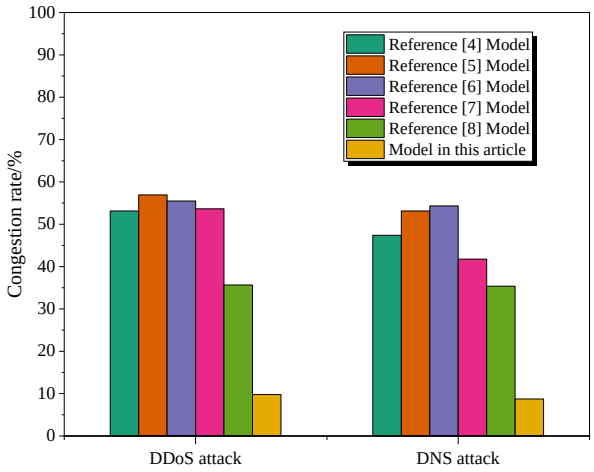


Fig. 8. Comparison of defense congestion rates of different defense methods.

Analysing Fig. 8, it can be seen that after being defended by six different network space security defense models, the congestion rate of DDoS attacks and DNS attacks is less than 60%. However, for the two different attacks, DDoS attacks and DNS attacks, the congestion rate of the model in this paper is the lowest, below 10%. The university's cyberspace security defense model proposed in this paper can effectively reduce the congestion rate of the university's cyberspace and protect the security of the university's cyberspace; the model in this paper can defend against different types of attack methods and has excellent defense capabilities against network space blocking attacks in universities.

In order to verify the effectiveness of the university's cyberspace security threat detection model in this paper and compare the impact of different numbers of feature quantities on detecting the university's cyberspace security threats, the number of feature quantities is set to 5, 10, 15, 25, 25, 30, 35, 40, and 45. The security threat detection accuracy of the university's cyberspace is obtained on different numbers of feature quantities, and the detection results are shown in Fig. 9.

From Fig. 9, it can be seen that through the feature extraction of the university's cyberspace using this model, as the number of features continues to increase, the model's accuracy in detecting security threats in the university's cyberspace has also improved. When the number of features is 35, the model has the highest accuracy in detecting security threats in the university's cyberspace, at around 0.98. However, when the number of features exceeds 35, the detection accuracy of the model in this paper for threats to the security of the university's cyberspace has begun to decline. Therefore, the model in this paper is used to reduce the dimensionality of the features in the university's cyberspace, setting the feature amount to 35, laying the foundation for the next step of the university's cyberspace security detection and improving detection accuracy.

In order to further verify the detection performance of the model in this paper for university's cyberspace security threats, DDoS attacks, DNS attacks, and DoS attacks are set as attack sources. The detection accuracy of the models in reference [4], reference [5], reference [6], reference [7], reference [8], and the model in this paper for university's cyberspace security threats are used as evaluation indicators and compared. The detection results are shown in Table IV.

From Table IV, it can be seen that the average detection accuracy of the model in reference [8] for DDoS attacks, DNS attacks, and DoS attacks is 0.76, which is relatively low in detection accuracy; The average detection accuracy of the models in reference [4], reference [5], reference [6], and reference [7] for three different network attacks is 0.82, 0.83, 0.80, and 0.80, all of which are above 0.8. Compared with the model in reference [8], the detection accuracy has been improved; the model in this paper has the highest detection accuracy for these three different types of network attacks, with an average detection accuracy of 0.98. The detection accuracy for each network attack is 0.97 or above, indicating that the model in this paper is the most effective in detecting security threats in the university's cyberspace and ensuring the security of the university's cyberspace.

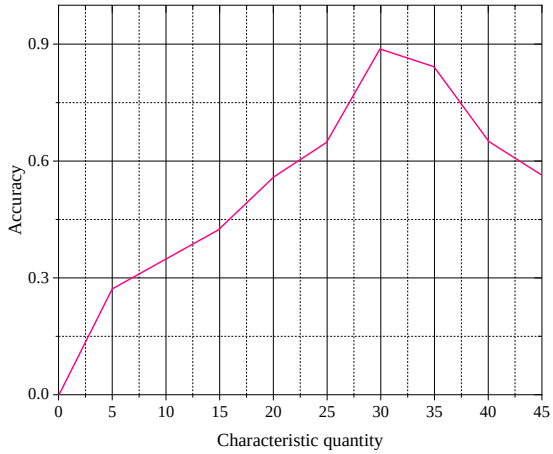


Fig. 9. University network security detection accuracy of different characteristic quantities.

TABLE IV. ACCURACY OF SECURITY DETECTION IN UNIVERSITY'S CYBERSPACE BY DIFFERENT MODELS

Model	DDoS	DNS	DoS
Reference [4] Model	0.85	0.79	0.83
Reference [5] Model	0.84	0.80	0.85
Reference [6] Model	0.79	0.81	0.81
Reference [7] Model	0.83	0.76	0.82
Reference [8] Model	0.73	0.77	0.79
Model in this paper	0.97	0.98	0.98

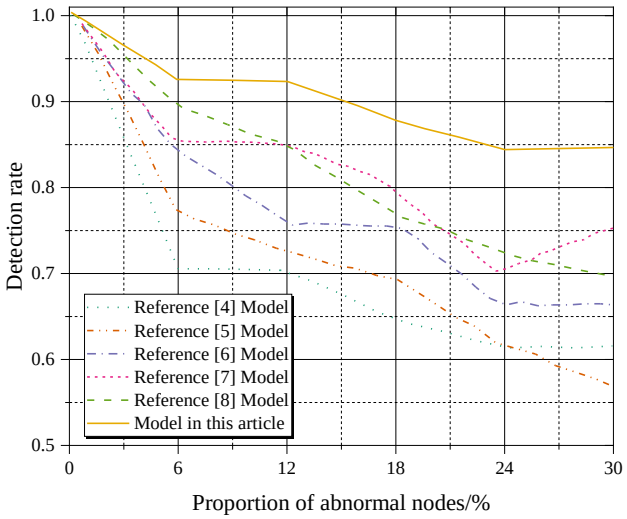


Fig. 10. Comparison of data detection rates collected by different models.

The core goal of the data collection model based on the compressed sensing principle is to ensure the high accuracy of the data. Distinguishing whether there is a threat to the security data nodes in the university's cyberspace is an important indicator to measure the model in this paper. Based on the model oriented towards the university's cyberspace security threat data in this paper, the detection rate of the algorithm is defined as the proportion of the detected university's cyberspace security threat data nodes to the total data nodes. In order to verify the effectiveness of the model proposed in this paper in detecting threats to the university's cyberspace security, 200 sensor nodes are set up in the experiment. The detection rates of the model proposed in this paper are compared with those of models in reference [4], reference [5], reference [6], reference [7], and reference [8] under different abnormal node ratios. The results are shown in Fig 10.

As shown in Fig. 10, as the proportion of nodes posing security threats to the university's cyberspace increases, the detection rates of the models in references [4], [5], [6], [7], [8] and this paper all show a downward trend. However, the detection rates of this paper are all above 0.9, while the detection rates of other models are all below 0.9. In comparison, the model in this paper has the highest detection rate. This model can effectively distinguish the types of efficient cyberspace security threat data nodes, provide decision-making support for data reliability collection, and improve the detection rate of cyberspace security threats in universities. Based on the above experimental results, it can be seen that the machine learning based university cyberspace security defense method proposed in this article has achieved significant results in practice. Compared with previous studies, the method proposed in this paper not only enhances existing research results, but also provides innovative solutions and improvement strategies.

Firstly, the method proposed in this article has achieved certain results in the defense of cyberspace security in universities. The model in this article combines data collection based on compressed sensing, security threat detection based on machine learning algorithms, and optimization parameter selection, providing comprehensive security defense support for university cyberspace.

Secondly, the method results presented in this article demonstrate novelty that was not discovered in previous studies. By introducing a compression aware data collection method, this article extracts effective information from large-scale network data, avoiding the related problems of storing and transmitting all the original data in traditional methods. Meanwhile, by using different machine learning models and feature selection methods, this article can accurately detect and prevent security threats in university cyberspace from multiple perspectives.

Finally, the research findings of this article provide new ideas and insights for the field of cyberspace security in universities. By effectively combining machine learning algorithms with knowledge in the field of network security, this method not only improves security defense capabilities, but also improves the accuracy and efficiency of security threat detection. These innovative achievements not only have significant academic significance, but also provide specific guidance and reference for the practice of cyberspace security in universities, and provide new solutions for network security protection in practical applications.

IV. Conclusion

The security threats and defense issues of university cyberspace are not only related to campus security but also closely related to the daily lives of teachers and students. There is a large amount of data, such as network traffic, log information, and system signals, in the cyberspace of universities. This study proposes a new method for university cyberspace security defense through the comprehensive application of compressed sensing based data collection method and deep learning model. After experimental verification, the model has achieved significant results in attack interception rate and provided effective protection for the security of university cyberspace. These results demonstrate the innovation and feasibility of the model, providing new ideas and insights for research and practice in the field of cyberspace security in universities. Due to the large amount of attack data, more types of data will be added to the model in the future, such as network traffic data, log data, user behavior data, etc., to improve the breadth and accuracy of security detection.

V. Data Availability

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation."

VI. Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work."

Acknowledgment

This work supported by Shanxi Provincial Universities Philosophy and Social Sciences Research Project 2022W170, Research on the Logical and Practical Path of the Construction of University Cyberspace Security.

REFERENCES

[1] C. Yin, "Application of Virtual Private Network Technology in University Network Information Security," in *Journal of Physics: Conference Series*, IOP Publishing, 2021, p. 042071.

[2] S. Ding, Z. Zhang, and J. Xie, "Network security defense model based on firewall and IPS," *Journal of Intelligent & Fuzzy Systems*, vol. 39, no. 6, pp. 8961–8969, 2020.

[3] Y. Niu, W. Du, and Z. Tang, "Computer Network Security Defense Model," in *Journal of Physics: Conference Series*, IOP Publishing, 2022, p. 012041.

[4] H. Wang and W. Li, "DDosTC: A transformer-based network attack detection hybrid mechanism in SDN," *Sensors*, vol. 21, no. 15, p. 5047, 2021.

[5] N. Oliveira, I. Praça, E. Maia, and O. Sousa, "Intelligent cyber-attack detection and classification for network-based intrusion detection systems," *Applied Sciences*, vol. 11, no. 4, p. 1674, 2021.

[6] Q. Abu Al-Haija and M. Al-Dala'ien, "ELBA-IoT: an ensemble learning model for botnet attack detection in IoT networks," *Journal of Sensor and Actuator Networks*, vol. 11, no. 1, p. 18, 2022.

[7] J. Zhang, L. Pan, Q.-L. Han, C. Chen, S. Wen, and Y. Xiang, "Deep learning-based attack detection for cyber-physical system cybersecurity: A survey," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 3, pp. 377–391, 2021.

[8] Ö. Tonkal, H. Polat, E. Başaran, Z. Cömert, and R. Kocaoğlu, "Machine learning approach equipped with neighbourhood component analysis for ddos attack detection in software-defined networking," *Electronics (Basel)*, vol. 10, no. 11, p. 1227, 2021.

[9] K. P. Gurumanapalli, and N. Muthuluru, "Feistel Network Assisted Dynamic Keying based SPN Lightweight Encryption for IoT Security," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, p. 377-392, 2021.

[10] R. Majeed, N. A. Abdullah, and M. F. Mushtaq, "IoT-based Cyber-security of Drones using the Nave Bayes Algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, p. 422-427, 2021.

[11] I. M. Ali, "LP Based Integration of Computer Network and Security in University College," in *Journal of Physics: Conference Series*, IOP Publishing, 2021, p. 012028.

[12] G. Sharma, "Secure Remote Access IPSEC Virtual Private Network to University Network System," *Journal of Computer Science Research*, vol. 3, no. 1, pp. 16–27, 2021.

[13] E. Balestrieri, P. Daponte, L. De Vito, F. Picariello, S. Rapuano, and I. Tudosa, "A Wi-Fi internet-of-things prototype for ECG monitoring by exploiting a novel compressed sensing method," Acta Imeko, vol. 9, no. 2, pp. 38–45, 2020.

[14] X. Xu et al., "Crack detection and comparison study based on faster R-CNN and mask R-CNN," Sensors, vol. 22, no. 3, p. 1215, 2022.

[15] J. Lu, L. Tan, and H. Jiang, "Review on convolutional neural network (CNN) applied to plant leaf disease classification," Agriculture, vol. 11, no. 8, p. 707, 2021.

[16] J. Zhang, X. Zhang, and L. Jiao, "Sparse nonnegative matrix factorization for hyperspectral unmixing based on endmember independence and spatial weighted abundance," Remote Sens (Basel), vol. 13, no. 12, p. 2348, 2021.

[17] P. Lu and W. Chen, "Vertex centrality of complex networks based on joint nonnegative matrix factorization and graph embedding," Chinese Physics B, 2022.

[18] A. Gallardo-Antolín and J. M. Montero, "On combining acoustic and modulation spectrograms in an attention LSTM-based system for speech intelligibility level classification," Neurocomputing, vol. 456, pp. 49–60, 2021.

[19] A. Agga, A. Abbou, M. Labbadi, and Y. El Houm, "Short-term self-consumption PV plant power production forecasts based on hybrid CNN-LSTM, ConvLSTM models," Renew Energy, vol. 177, pp. 101–112, 2021.

[20] Y. Ma, L. Wu, and Z. Li, "A novel face presentation attack detection scheme based on multi-regional convolutional neural networks," Pattern Recognit Lett, vol. 131, pp. 261–267, 2020.

[21] J. Kim, J. Kim, H. Kim, M. Shim, and E. Choi, "CNN-based network intrusion detection against denial-of-service attacks," Electronics (Basel), vol. 9, no. 6, p. 916, 2020.

[22] B.-Z. Han and W.-X. Huang, "Active control for drag reduction of turbulent channel flow based on convolutional neural networks," Physics of Fluids, vol. 32, no. 9, p. 095108, 2020.

[23] X. Liu and X. Di, "TanhExp: A smooth activation function with high convergence speed for lightweight neural networks," IET Computer Vision, vol. 15, no. 2, pp. 136–150, 2021.

[24] Y. Liu, J. Wang, H. He, G. Huang, and W. Shi, "Identifying important nodes affecting network security in complex networks," Int J Distrib Sens Netw, vol. 17, no. 2, p. 1550147721999285, 2021.

[25] I. Khan, W. Farrelly, and K. Curran, "A Demonstration of Practical DNS Attacks and their Mitigation Using DNSSEC," International Journal of Wireless Networks and Broadband Technologies (IJWNBT), vol. 9, no. 1, pp. 56–78, 2020.

[26] C. Zhou, Y. Yu, S. Yang, and H. Xu, "Intelligent immunity-based security defense system for multi-access edge computing network," China Communications, vol. 18, no. 1, pp. 100–107, 2021.

[27] R. Kadhim and M. Gaata, "A hybrid of CNN and LSTM methods for securing web application against cross-site scripting attack," Indones. J. Electr. Eng. Comput. Sci, vol. 21, pp. 1022–1029, 2020.

[28] S. Janarthanam, N. Prakash, and M. Shanthakumar, "Adaptive learning method for DDoS attacks on software defined network function virtualization," EAI Endorsed Transactions on Cloud Systems, vol. 6, no. 18, pp. e6–e6, 2020.

Investigating Efficiency of Soil Classification System using Neural Network Models

Pappala Mohan Rao[1], Neeli Koti Siva Sai Priyanka[2], Kunjam Nageswara Rao[3], SitaratnamGokuruboyina[4]
Research Scholar, Department of CS&SE, Andhra University College of Engineering, Andhra University, Visakhapatnam, India[1]
MTech, Department of CS&SE, Andhra University College of Engineering, Andhra University, Visakhapatnam, India[2]
Professor, Department of CS&SE, Andhra University College of Engineering, Andhra University, Visakhapatnam, India[3]
Research Scientist, Institute of Bioinformatics and Computational Biology (Recognized as SIRO), Visakhapatnam, India[4]

***Abstract*—Soil is a vital requirement for agricultural activities providing numerous functionalities restoring both abiotic and biotic materials. There are different types of soils, and each type of soil possesses distinctive characteristics and unique harvesting properties that impact agricultural development in various ways. Generally, farmers in the olden days used to analyse soil by looking at it visually while some prefer laboratory tests which are time-consuming and costly. Testing of soil is done to analyse the features and characteristics of the soil type, which results in selecting a suitable crop. This in turn results in increased food productivity which is very beneficial to farmers. Hence, to recognize the soil type an automatic soil identification model is proposed by implementing Deep Learning Techniques. It is used to classify the soil for crop recommendation by analysing accurate soil type. Different Convolution Neural Networks have been applied in the proposed model. They are VGG16, VGG19, InceptionV3 and ResNet50.Among all those techniques it is analysed that better results were obtained with ResNet50 having an accuracy of about 87% performing Multi-classification that is Black soil, Laterite Soil, Yellow Soil, Cinder soil & Peat soil.**

Keywords*—Agricultural; convolution neural network; soil classification deep learning; VGG16; VGG19; InceptionV3; multi-classification; ResNet50

I. Introduction

Soil stands as a vital agricultural resource, housing a plethora of nutrients essential for crop cultivation. Each region possesses its unique soil composition, giving rise to diverse soil types worldwide. Serving as Earth's outer layer, soil encompasses various minerals, organic matter, living organisms, and water. Acting as a bridge between the planet's internal layers and its surface, soil plays a pivotal role in facilitating plant growth. The availability of nutrients and water in the soil determines the growth of a plant, making it a crucial factor for analysis. In the current era, effective plant management holds paramount importance to ensure human sustainability. The global population have been steadily increasing, thus there is a growing need to enhance food, fabric, and medicine production. Consequently, improving the agricultural sector becomes imperative, as it stands as a primary source capable of meeting these escalating demands.

The growth of different types of crops is influenced by diverse factors such as the availability of nutrients, water, and oxygen in the soil, balancing both environmental and physical conditions to achieve a good yield. Among all these resources the current work concentrates on the different classifications of soil to identify the specific category of soil which ultimately determines a selective crop. Each region is determined with different types of soils like black soil, peaty soil, alluvial soil, red soil, desert soil, forested soil, laterite soil, and many more. For soil classification various Deep Learning techniques are implemented as they are categorized based on the considered soil images. To classify soil types, a variety of features such as hue, saturation, texture, colour, intensity, and other relevant characteristics are extracted and utilized.

Deep learning is an architecture comprising a large number of layers that enables the transformation of raw data into meaningful features. This process is often referred to as feature engineering in the context of deep learning models. Various types of CNNs are used for classification. Some of the existing models are implemented using both Machine Learning and Deep Learning Techniques [1]. The author implements a model using K-NN Classifier and SVM Classifier on different types of soil image datasets [2] [9] [14]. The author implements a model using an SVM classifier. The research in [3] [13] proposed a technique designed using an SVM classifier and ResNet50, CNN while the author in [4][17], employs a range of neural network models, including CNNs, DBNs, LSTMs, Multilayer Perceptron, and Autoencoders. These architectures serve various purposes such as image processing and feature extraction within the study. The study in [5] [16] proposed a technique implementing an, Network Model (CNN), SVM classifier and StoolNet. The research in [15] [6], the author implemented a Network Model based on Convolutional Neural Networks (CNN). This model was designed to filter images using five different types of masks: spot, wave, level, ripple, and edge. These masks are utilized as filtering techniques in image processing, helping to extract specific features or enhance certain characteristics in the images. The application of these masks within a CNN framework suggests that the study focuses on leveraging convolutional operations to analyses and process images for a particular purpose, such as texture analysis or pattern recognition [10]. A model has been proposed by the author that combines Naive Bayes and Artificial Neural Network (ANN) techniques. This hybrid approach is utilized to analyze physical parameters like water content and dry density, as well as soil parameters such as internal friction angle, cohesion. Simultaneously in [6] the study employs a Gabor Filter for edge detection, followed by Std, skew, Mean, and kurtosis and other sinusoidal measures

to retrieve images. The classification task involves using the CNN algorithm for land and soil images. Additionally, a Visual Transformer for image classification is introduced, showcasing superior accuracy compared to CNN, SVM, ResNet-50 and subsequent machine learning models as stated in study [3].

Therefore, in this paper, a model developed is determined using VGG16, InceptionV3, VGG19 and ResNet50associated with Transfer Learning resulting in better and more accurate soil classification. The main objective of this model is to do multi-classification that is yellow soil, Laterite Soil, Cinder soil, Black Soil and Peat soil. The general workflow of the proposed model is explained and continued with the Literature Survey in Section II and Section III provides a diagrammatic approach for the developed model. Section IV covers the experimental evaluation. Section V presents the conclusion of the proposed paper.

II. Literature Survey

Rahman Zaminur proposed a system that utilized the K-NN Classifier, a machine learning algorithm, for classifying soil texture. The research recommended Support Vector Machine as the most effective classifier. The approach involved Bootstrap resampling and a stacked decision tree ensemble classifier. The study encompassed nine different soil types and also explored alternative algorithms such as Artificial Neural Network and GAtree [1].

Navya and Vijay E V proposed a system the image classification process, employing diverse methods including, Sub-pixel Classification, Artificial Neural Network Classification and Maximum Likelihood Classification. They opted for SVM due to its versatility and suitability for comparison purposes. The research utilized a range of pre-processing methods to identify patterns, leading to improved classification analysis [2].

Jagetia Aaryan underscored the significance of precise soil classification, substantiated by multiple parameters such as void ratio, moisture content, liquid limit, clay content, specific gravity and plasticity. Employing the Visual Transformer, an advanced technique for image classification, resulted in impressive accuracy rates of 98.13% during training and 93.62% during testing [3].

Prabhavathi V's research is cantered on Utilizing deep learning algorithms to classify soil, particularly highlighting an inventive deep learning model. The investigation delves into a range of deep learning algorithms, encompassing CNNs, DBNs, LSTM, Autoencoders and Multilayer Perceptron CNNs, in particular, exhibit remarkable accuracy when it comes to Identifying soil through the analysis of hyperspectral bands derived from satellite data and categorizing aggregates using stereo-pair images [4].

Srivastava Pallavi research proposal investigates soil classification techniques through computer vision and the utilization of image processing. colour and soil texture are determined employing methods such as the Munsell colour chart, elutriation, pipette, decantation, The model incorporates StoolNet, which attains a remarkable 100% accuracy in classifying burozem soil and yellow soil [5].

Aparna Yerrolla examines the International Soil Reference and Information Centre dataset, which includes various soil images belonging to different classes. Feature extraction is performed using the Gabor filter to capture attributes like entropy, standard error, and mean. The analysis places significant importance on soil colour extraction. Texture characteristics are extracted using the Laws mask method, involving filtering images with five different mask types: edge, ripple, level, wave and spot. The application of Convolutional Neural Networks (CNNs) with three-layered hidden layers enables the algorithm to effectively classify land images and soil, distinguishing between categories such as Clayey peat, Silty sand, Clay, Humus clay, Peat, Sandy clay and Clayey sand [6].

Barkataki Nairit introduced a deep CNN model designed for automatic soil classification through non-invasive techniques such as ground penetrating radar (GPR). A fabricated dataset was created through the use of gpr-Max for both training and validation. Through a 5-fold cross-validation, the model demonstrated exceptional performance, achieving an impressive accuracy rate of 97% in classifying seven distinct soil types based on ground penetrating radar (GPR) B-Scan images [7].

Khullar Vikas puts forward an effective soil classification system by leveraging deep learning techniques. The study incorporates a diverse set of algorithms, including Random Forest, KNN, Ada-Boost, SVV Machine, Quadratic Discriminant Analysis, Logistic Regression, Decision Tree, Extra Trees, Gaussian Naïve Bayes, and Histogram Gradient Boosting. Additionally, the use of VGG16 and InceptionResNetV2 deep networks for soil classification enhances the categorization process, yielding robust and dependable results that outperform prior state-of-the-art methodologies [8].

Greema S Raj describes a survey done on soil classification using different techniques. Decision tree predictions are made using a binary tree model, known for its speed and accuracy. Naïve Bayes classifiers utilize the Bayes theorem for predicting unrelated features within a class. Parameter estimation is performed using maximum likelihood or Bayesian methods. SVM is a heuristic algorithm employed for supervised learning, determining the optimal hyperplane to separate two classes [9].

Ladan Samadi's research focuses on soil classification using machine learning algorithms, namely ANN and Naïve Bayes. Neural networks are effective in establishing relationships between input variables and target parameters. The Unified Soil Classification System (USCS) is employed for soil classification based on and particle size analysis Atterberg limits. Naïve Bayes and ANN algorithms are utilized for classification, considering particle size analysis and Atterberg limits. The objective is to develop an Artificial Neural Network model that predicts soil classification based on soil conditions and collected data on soil mechanics parameters [10].

Rakesh Kr Dwivedi's Deep learning process, using K-means clustering, aids farmers in classifying soil based on its texture, clay, silt, sand concentrations, and pH value. Soil

image classification utilizing machine learning involves three steps: image segmentation, feature extraction, and classification [11].

Abhinav Pandey utilizes a deep CNN methodology for satellite image classification, concentrating on the classification of soil types using chemical and physical properties as criteria. Using the DGX-2, the trained deep CNN achieves an average accuracy of 0.67 and a maximum accuracy of 0.80 in five runs when utilizing images with vegetation removed. However, accuracy decreases to 0.41 due to vegetation fluctuation over time. The desert soil class shows the highest confidence, while the black soil class exhibits the lowest. The average classification accuracy during testing is 0.72, indicating the effectiveness of the model in classifying different soil types from satellite images [12].

The existing methods were implemented mostly on one type of soil from various soil images, based on their respective parameters. This research works aims to classify different types of soil like black soil, yellow soil, laterite, cinder and peat soil with more efficient CNN models.

III. Proposed Model

The suggested model uses the input photos to determine the type of soil. Several varieties of convolution neural networks, which are a part of deep learning techniques, are used in this instance to automatically classify data. Several unseen layers are stacked upon each another in a specific order to create CNN, feed-forward neural network which is multi-layered, which extracts features that are displayed as patterns. The pre-processed input data is then sent through many CNN types, such as ResNet50, InceptionV3, VGG16, and VGG19, to further aid in the classification of the picture of soil under consideration. Fig. 1 illustrates the suggested model's workflow.

The collection of images in the dataset is divided into four categories: images of black soil, yellow soil, peat soil, cinder soil, and laterite soil. It is gathered from offline and internet sources. When putting them into the model, they are mainly divided into Train and Test data images and are pre-processed utilizing techniques for image transformation. Pre-processing procedures, which involve the implementation of Data Augmentation Techniques, are executed in this context. During data augmentation, specific circumstances are applied, such as rotating the image at a ninety-degree angle and modifying it horizontally and vertically. This is done in order to capture more photographs and view it from various perspectives. Images are resized simultaneously to change their aspect ratio and size to a standard 220x220x3 format.

Every image has a specific label applied to it and is mapped for additional classification. As a consequence, pre-processed train and test image data are obtained. Thus, the acquired trained dataset is used to train the model, and the test dataset is used to assess the model.

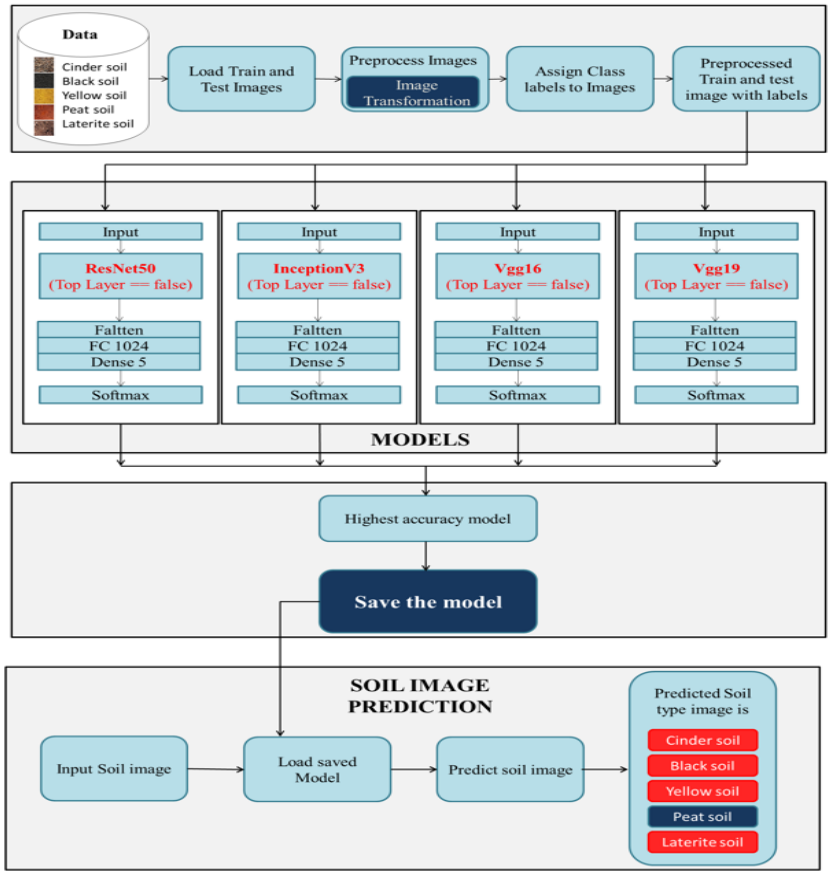


Fig. 1. Workflow of the proposed model.

The acquired data undergoes processing through several CNNs, namely Inception V3, ResNet50, VGG16, and VGG19. Each neural network's individual layer operates on the input data, extracting distinct patterns crucial for classifying the particular soil type under examination. CNNs stand out as highly efficient architectures for image recognition and classification owing to their pattern recognition capability, significantly aiding in the prediction and classification of input image data and yielding the highest achievable accuracy. Standard CNNs typically comprise three fundamental types of layers, often referred to as "building blocks": convolutional, pooling, and fully connected layers. The initial two layers, convolution and pooling, primarily focus on feature extraction, while the third, the fully connected layer, translates these identified features into the final output, such as a classification. Within a CNN, which functions as a sequence of mathematical operations, the convolution layer plays a pivotal role, providing a specialized version of linear operations. Digital images are represented as an array of integers or a two-dimensional (2D) grid capturing pixel values. A kernel, serving as an optimizable feature extractor, is applied at each position across the image, allowing for the extraction of essential features.

We use Transfer learning technique to Connect base model with fine-tuned model Fully Connected Layers. Subsequently, layer freezing is carried out employing the trained data images for the model's first run. When knowledge is taken from an established model and applied to a newly proposed model, transfer learning plays a significant role. Every item listed above The Transfer Learning Mechanism is the foundation upon which CNNs operate. The epochs, learning rate, optimizer and batch size are among the many parameters that are changed for each Convolutional Neural Networks, to classify the photos of dirt, save the model and run it.

Plots of loss percentages and accuracy have been made in accordance with the outcomes of the designed model. One indicator used to assess the model's performance in an understandable manner is accuracy. A metric called loss indicates how well the model performs following each optimization step. In order to anticipate the provided input soil input, we must load the saved model weights concurrently with the input picture path of the testing image, and then predict the input image using the loaded model. Class label assigned determines the index of output, which is based on the index of the maximum element within an array along a specified axis.

Algorithm: Implementation of Model:

Input: Images of different kinds of soil

Output: Prediction of image (Peat, Yellow, Laterite, Black and Cinder Soil images)

Step 1.Importing all training and testing images corresponding to soil categories into the dataset.

Step 2. Preprocessing the Images.

- Enhancing all training and testing images through the application of data augmentation techniques.

Step 3.Associating class labels with their respective images by creating a mapping.

Step 4.Generating the pre-processed dataset after performing all necessary transformations, including data augmentation, and mapping class labels with their corresponding images.

Step 5.Utilizing ResNet50, InceptionV3, Vgg16, and Vgg19 models as the base models for further analysis or processing.

Step 6.Integrating a fine-tuned model into Vgg16, InceptionV3, ResNet50, and Vgg19 by customizing the top layers. Steps

- Configuring the model with specific parameters such as epochs, batch size and learning rate to train and optimize the neural network.
- Incorporating a flattened and fully connected layer into the base model architecture.
- Appending a dense layer with five units (for the number of classification classes) and applying the “SoftMax” activation function to the model.

Step 7. Implementing transfer learning by connecting the base model with fine-tuned fully connected layers to leverage the pre-trained features and optimize the model for the specific classification task.

Step 8. Freezing the layers to maintain their pre-trained weights and prevent them from being updated during the initial execution of the model.

Step 9. Training the model by fitting the training data images to it.

Step 10. Saving the model weights after training for future use or further analysis.

Step 11. Plotting accuracy and loss percentages to visualize the performance of the designed model during training and evaluation phases.

Step 12. The classification of images has been successfully completed using the trained model.

1) VGG16: The utilization of Very Deep Convolutional Networks in Large-Scale Image Recognition showcases the impact of network depth on accuracy within extensive image

identification scenarios. The primary focus is a comprehensive analysis of networks with incrementally deeper layers, employing an architecture using (3x3) convolution filters which is relatively small. This study reveals that increasing the depth to a range of 16–19 layers significantly enhance performance compared to current configurations. Our team participated in the 2014 ImageNet Challenge, leveraging these discoveries, leading to our team securing top two places in the localization and classification tracks, respectively. This success underscored the adaptability of our representations across diverse datasets, consistently yielding state-of-the-art results. To encourage continued exploration of deep visual representations in computer vision, our two most impactful ConvNet models available for research purposes.

2) VGG19: The depth of convolutional networks affects their accuracy in large-scale image identification contexts. It employs an architecture with (3x3) convolution filters, just like VGG19, and shows that going deeper to 16–19 weight layers can yield a discernible improvement over the current configuration. When using this VGG19, we have more hidden layers, which yields the best outcomes.

3) InceptionV3: Convolutional networks are the foundation of most state-of-the-art computer vision systems for various workloads. Since its debut in 2014, very deep convolutional networks have made considerable progress in a number of benchmarks. We investigate strategies for scaling up networks that leverage factorized convolutions and aggressive regularization to make the most efficient use of the extra processing. While it's often true that larger model sizes and increased computational resources lead to improved performance in various tasks, provided there is a sufficient amount of labelled data for training, we are exploring methods to efficiently scale up neural networks. This involves applying strong regularization techniques and optimizing convolutions through appropriate factorization, ensuring that the additional computation is used as efficiently as possible, these methods are compared to the state of the art using ILSVRC 2012 automation challenge validation set and find notable improvements: For single frame evaluation, a network with less than 25 million parameters and a computational cost of 5 billion multiply-adds per inference produced errors of 21.2% top-1 and 5.6% top-5. Using an ensemble of four models with multi-crop assessment, we report a 17.3% top-1 error and a 3.5% top-5 error on the validation set and 3.6% error on the test set.

4) ResNet50: Deep residual networks have emerged as a category of highly extensive architectures, demonstrating exceptional accuracy and attractive convergence behaviours. The analysis delves into the propagation formulations fundamental to the residual building blocks, suggesting that both forward and backward signals can be seamlessly transmitted from one block to any other block, specifically when utilizing identity mappings as skip connections followed by activation after addition. Numerous ablation experiments confirm the significance of these identity mappings, which, in turn, serve as the inspiration behind our introduction of a novel residual unit. This new unit not only enhances generalization but also streamlines the training process.

Deep residual networks represent a class of highly extensive architectures known for their exceptional accuracy and favourable convergence behaviours. This analysis explores the propagation formulations within the residual building blocks, particularly when employing identity mappings as skip connections along with post-addition activation. It indicates the direct transferability of both forward and backward signals between various blocks. Numerous ablation experiments further affirm the significance of these identity mappings. As a result, this serves as the driving force behind our proposal for a novel residual unit, aiming to streamline training processes and enhance generalization.

IV. Experimental Analysis

In this paper, the system has been tested with 4 deep-learning keras API models using different optimization techniques with 203 soil images of five categories: Laterite Soil, Peat Soil, Black Soil, Yellow Soil and Cinder Soil.

There are 47 tests done where nine images for each category all the images in the dataset are divided into eight batches. Each image has its own 30 iterations along with various features optimizers that belong to Adam and SGD. There are 156 training images of are done for each model.

For the VGG16 and VGG19 systems recorded mean accuracies are 100% & 100% and loss error rates are 0.0032 & 0.0019, for InceptionV3 system recorded a mean of 87% and the mean loss value is 0.4747.

By using the SGD optimizer technique with the ResNet50 System recorded highest accuracy is 100% and the loss error rate is 0.0013 using the Adam optimizer technique with the ResNet50. From all the above models ResNet50 with the Adam optimizer system is a bit higher than the remaining models.

To compute accuracy, we utilize the confusion matrix, which consists of four categories: True Positives, True Negatives, False Positives, and False Negatives. This matrix enables the calculation of various valuable metrics.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

TABLE I. Measuring Accuracy using the Presented Techniques

Models	VGG16	VGG19	Inception V3	ResNet50
Validation Accuracy	1.0000	1.0000	0.8750	1.0000
Validation Loss	0.0032	0.0019	0.4747	0.0013
Training Accuracy	0.8871	0.9355	0.5645	0.8750
Training Loss	0.9862	0.5171	1.0490	0.9059

The graphs below illustrate the accuracy and loss metrics obtained from different models, including VGG16, InceptionV3, VGG19 and ResNet50.

As illustrated from Fig. 2, the accuracy values ranging from 0.0 to 1.0 is represented on X-axis and the number of epochs (100) executed in the model is represented on Y axis. Throughout the study, 100 epochs were conducted, and the optimal accuracy was achieved at 26 epochs, accompanied by a loss rate of 0.0032. The blue line on the graph corresponds to the training data, while the orange line represents the validation data. It is important to note that 20% of the images were randomly selected from each class for testing purposes, ensuring a representative evaluation of the model's performance.

As illustrated from Fig. 3, the accuracy values ranging from 0 to 70 is represented on X-axis and the number of epochs (100) executed in the model is represented on Y axis. The graph illustrates the loss corresponding to the accuracy of the VGG16 model. The optimal result was achieved at 26 epochs with a minimal loss rate of 0.0032. The blue line in the graph represents the training data, and the orange line represents the validation data. It's important to note that 20% of the images were randomly selected from each class for testing purposes, ensuring a representative evaluation of the model's accuracy and loss.

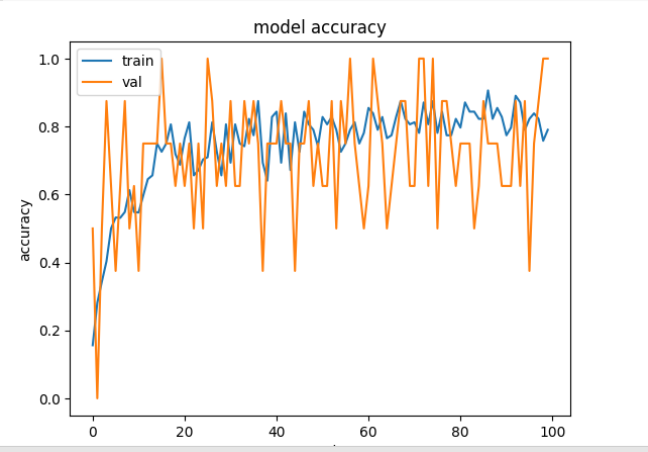


Fig. 2. Accuracy of VGG16.

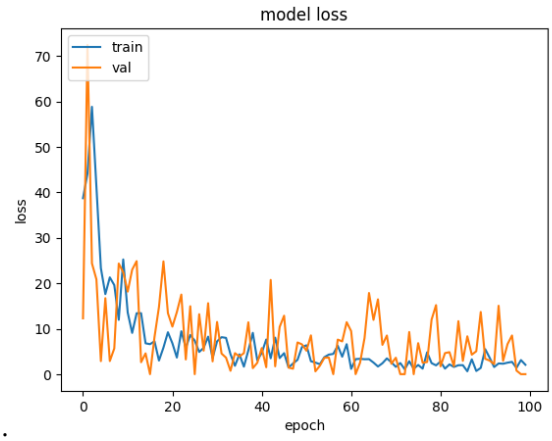


Fig. 3. Loss of VGG16.

As illustrated from Fig. 4, the accuracy values ranging from 0.3 to 1.0 is represented on X-axis and the number of epochs (100) executed in the model is represented on Y axis. In the course of this study, 100 epochs were conducted, and the highest accuracy was achieved at 58 epochs, with a minimal loss rate of 0.0019. The blue line on the graph corresponds to the training data, while the orange line represents the validation data. It's worth noting that 20% of the images were randomly selected from each class for testing purposes, ensuring a representative evaluation of the model's performance.

The blue line is representative of the training data, while the orange line signifies the validation data. Notably, for testing, 20% of images were randomly selected from each class.

As illustrated from Fig. 5, the accuracy values ranging from 0 to 60 is represented on X-axis and the number of epochs (100) executed in the model is represented on Y axis. The graph illustrates the loss corresponding to the accuracy of the VGG16 model. The lowest loss rate was achieved at 58 epochs, with a value of 0.0019. The blue line in the graph represents the training data, while the orange line represents the validation data. It's important to note that 20% of the images were randomly selected from each class for testing, ensuring a reliable evaluation of the model's accuracy and loss.

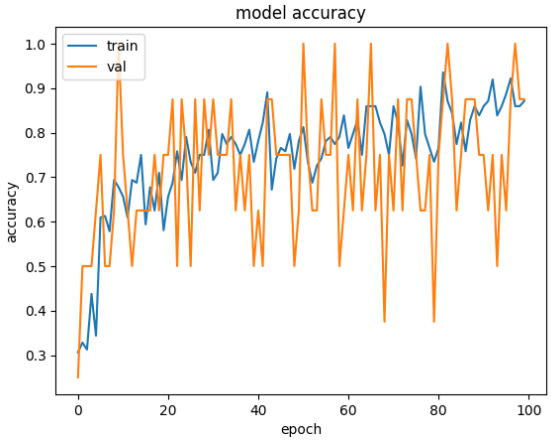


Fig. 4. Accuracy of VGG19.

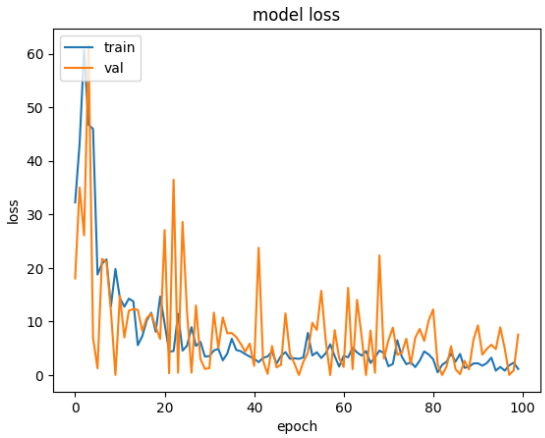


Fig. 5. Loss of VGG19.

As illustrated from Fig. 6, the accuracy values ranging from 0.0 to 0.8 is represented on X-axis and the number of epochs (100) executed in the model is represented on Y axis. During the study, the model was trained for 100 epochs, and the highest accuracy, 87%, was achieved at 52 epochs with a corresponding loss rate of 0.4747. The blue line on the graph represents the training data, while the orange line represents the validation data. It's important to note that for testing purposes, 20% of the images were randomly selected from each class, ensuring a representative evaluation of the model's performance.

The blue line corresponds to the training set, while the orange line represents the validation set. Notably, for testing purposes, 20% of images were randomly selected from each class.

As illustrated from Fig. 7, the accuracy values ranging from 0 to 80 is represented on X-axis and the number of epochs (100) executed in the model is represented on Y axis. The graph illustrates the loss corresponding to the accuracy of the VGG16 model. The lowest loss rate was achieved at 52 epochs, with a value of 0.4747. The blue line on the graph corresponds to the training data, while the orange line represents the validation data. It's important to note that for testing purposes, 20% of the images were randomly selected from each class, ensuring a representative evaluation of the model's accuracy and loss.

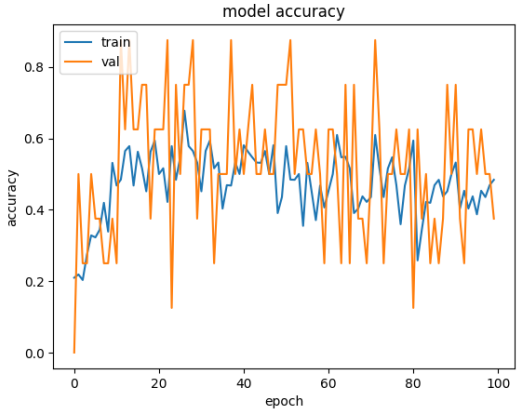


Fig. 6. Accuracy of inceptionV3.

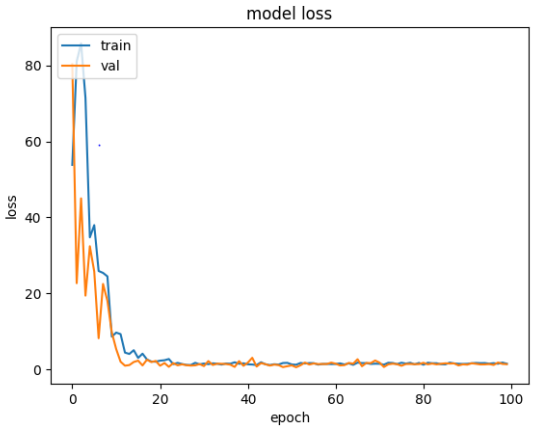


Fig. 7. Loss of inceptionV3.

As illustrated from Fig. 8, the accuracy values ranging from 0.4 to 1.0 is represented on X-axis and the number of epochs (100) executed in the model is represented on Y axis. Throughout the study, the model had been trained for 100 epochs, and the highest accuracy, achieved at 22 epochs, was recorded at 0.0013 loss rate. The blue line on the graph corresponds to the training data, while the orange line represents the validation data. It's important to note that 20% of the images were randomly selected from each class for testing purposes, ensuring a representative evaluation of the model's performance.

The training data is represented by the blue line, while the validation data is indicated by the orange line. Notably, 20% of images from each class were randomly chosen for testing.

As illustrated from Fig. 9, the accuracy values ranging from 0 to 40 is represented on X-axis and the number of epochs (100) executed in the model is represented on Y axis. The graph illustrates the loss corresponding to the accuracy of the VGG16 model. The lowest loss rate was achieved at 22 epochs, with a value of 0.0013. The blue line in the graph illustrates the training data, while the validation data is depicted by the orange line. It's important to note that for testing purposes, 20% of the images were randomly selected from each class, ensuring a representative evaluation of the model's accuracy and loss.

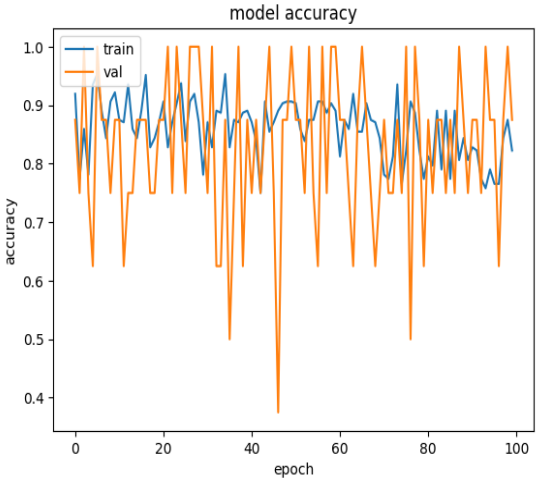


Fig. 8. Accuracy of ResNet50.

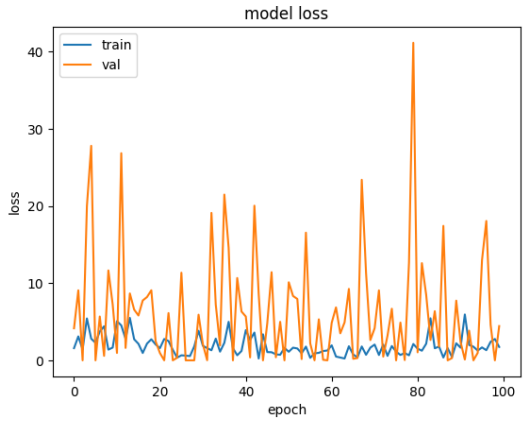


Fig. 9. Loss of ResNet50.

Out of the listed models, namely InceptionV3,VGG16, ResNet50 and VGG19 the ResNet50 model stands out, exhibiting the highest accuracy with minimal loss. Furthermore, in addition to accuracy and loss metrics, the output includes Soil Image and the subsequent soil classification as described below.

Fig. 10 shows the output of the model which is name of the soil here the classification of soil type is Yellow Soil, list of values in array and the image which is given as input to the model with height 220 and width 220.. Out of the listed models such as VGG16, VGG19, InceptionV3, ResNet50 , the ResNet50 got highest accuracy with less loss. So model taken ResNet50 for execution and identified as Yellow Image.

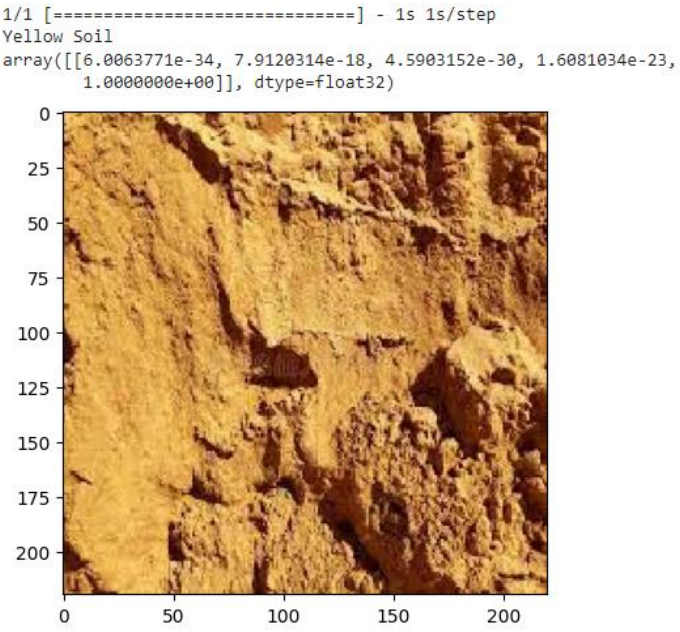


Fig. 10. Soil classification.

ResNet50 achieves high accuracy and minimal loss due to its 50-layer architecture utilizing a bottleneck design for building blocks. The bottleneck residual block incorporates 1×1 convolutions, called a "bottleneck," reducing parameters and matrix multiplications. This design enables faster training of each layer. Unlike the traditional two-layer approach, ResNet50 employs a stack of three layers, contributing to its superior performance.

V. Conclusion

Soil classification holds significant importance as it aids in the analysis of soil nutrients and minerals. This analysis enables precise crop management, leading to enhanced productivity and meeting the growing food demands. Deep Learning Techniques are employed to address challenges encountered in the manual soil classification process. This project emphasizes the utilization of Deep Learning and Image Processing for classifying soils, focusing on key soil characteristics such as particle size, texture and color. These techniques aim in replacing the traditional manual soil inspection methods. The proposed model incorporates four different Convolutional Neural Networks (CNNs) – ResNet50, VGG16, VGG19, and InceptionV3 – for multi-class classification, categorizing soils into Yellow, Peat, Cinder, Laterite and Black Soils respectively. Among these neural networks, ResNet50 outperformed the others, achieving a minimal error rate of 0.13% and a flawless accuracy of 100%. Although the model was tested with a limited dataset, expanding the dataset could potentially improve accuracy further. Additionally, incorporating pH values into the dataset and integrating crop recommendations could enhance the model's capabilities in extracting soil nutrients and minerals. This model can be enhanced by using large dataset, to get nutrients and minerals from soil by adding pH values to the dataset and also for crop recommendation.

References

[1] Shraddha Shivhare , Kanchan Cecil "A Review on Automatic Soil Classification in Digital Image Processing" International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429 Volume 9 Issue VIII Aug 2021

[2] Vijay E V, Navya Ch, Abdul Shabana Begum, Rajaneesh D, Mahesh Babu B , "Soil Classification Using Image Processing and Modifies SVM Classifier" , International Journal of Research in Advent Technology, Vol.8, No.9, September 2020 E-ISSN: 2321-9637

[3] Aaryan Jagetia, Umang Goenka , Priyadarshini Kumari , Mary Samuel."Visual Transformer for Soil Classification", 2022 IEEE Students Conference on Engineering and Systems (SCES), July 01-03, 2022, Prayagraj, India

[4] Prabhavathi V , Kuppusamy P "A study on Deep Learning based Soil Classification", 2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)

[5] Pallavi Srivastava, Aasheesh Shukla, Atul Bansal, "A comprehensive review on soil classification using deep learning and computer vision techniques", Multimedia Tools and Applications (2021) 80:14887–14914.

[6] Yerrolla Aparna, Dr. Giddaluru Somasekhar, Nuthanakanti Bhaskar, "Analytical Approach for Soil and Land Classification Using Image Processing with Deep Learning", 2023 2nd International Conference for Innovation in Technology (INOCON) Bangalore, India. Mar 3-5, 2023

[7] Nairit Barkataki , Sharmistha Mazumdar , P Bipasha Devi Singha "Classification of soil types from GPR B Scans using deep learning techniques", 2021 6th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), August 27th & 28th 2021.

[8] Vikas Khullar1 ,Sachin Ahuja2 , Raj Gaurang Tiwari3 , Ambuj Kumar Agarwal "Investigating Efficacy of Deep Trained Soil Classification System with Augmented Data", 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions (ICRITO) Amity University, Noida, India. Sep 3-4, 2021.

[9] Greema S Raj, Lijin Das S , "SURVEY ON SOIL CLASSIFICATION USING DIFFERENT TECHNIQUES", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072.

[10] Ladan Samadi, Hanan Samadi, "Soil Classification Modelling Using Machine Learning Methods", See discussions, stats, and author profiles for this publication, March 2022.

[11] Rakesh Kr Dwivedi , Neeraj Kumari , Ashish Bishnoi, Rajendra Prasad Pandey, "Soil Identification and Classification using Machine Learning: A Review", Proceedings of the SMART–2022, IEEE Conference ID: 55829 11th International Conference on System Modeling & Advancement in Research Trends, 16th–17th, December, 2022 College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India.

[12] Abhinav Pandey, Devesh Kumar, Debarati B. Chakraborty , "SOIL TYPE CLASSIFICATION FROM HIGH RESOLUTION SATELLITE IMAGES WITH DEEP CNN" , 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS**,**INSPEC Accession Number **:** 21228955,DOI**:** 10.1109/IGARSS47720.2021.9554290.

[13] K. Srunitha and S. Padmavathi, “Performance of SVM classifier for image based soil classification,” in International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES),2016, pp. 411-415.

[14] A.I. Khan and S. Al-Habsi, Machine learning in computer vision. Procedia Computer Science, vol 167, pp 1444-1451, 2020.

[15] Srivastava, P., Shukla, A. Bansal, A,“ A comprehensive review on soil classification using deep learning and computer vision techniques,” in Multimed Tools Appl 80, 14887aˆ14914 (2021).

[16] K. Sharma and S. Kumar, “Soil Classification Characterization Using Image Processing,” in Second International Conference on Computing Methodologies and Communication (ICCMC), 2018, pp. 885-890, doi: 10.1109/ICCMC.2018.8488103.

[17] P. A. Harlianto, T. B. Adji and N. A. Setiawan, “Comparison of machine learning algorithms for soil type classification,” in 3rd International Conference on Science and Technology - Computer (ICST), 2017, pp. 7-10, doi: 10.1109/ICSTC.2017.8011843.

An Empirical Study: Automating e-Commerce Product Rating Through an Analysis of Customer Review

Uvaaneswary Rajendran[1], Salfarina Abdullah[2], Khairi Azhar Aziz[3], Sazly Anuar[4]
Department of Software Engineering and Information System, Faculty of Computer Science and Information Technology
University Putra Malaysia[1, 2, 3]
Industrial Automation Section, UNIKL Malaysia France Institute (UNIKL-MFI)[4]

***Abstract*—e-Commerce today is a remarkable experience. However, finding and purchasing a right quality product based on numerous product reviews and manual rating in the e-commerce websites utilize much time among the consumers. This paper presents the problems faced by the consumers when buying products in e-commerce websites and a solution to solve the problems. Thus, the idea of an automated product rating system would be very useful for the consumers in which it rates the products automatically based on the reviews given by the buyers. To do this, a technique called Sentiment Analysis is used. It also ranks the products in order based on the product rating that is generated automatically. It would provide a way for the consumers to purchase their desired product within minutes. Surveys and interviews were conducted to find out the problems faced by consumers when purchasing a product online through e-commerce websites. There was also research conducted to study the product rating and product review section on the current e-commerce websites. To conclude, this automated product rating system eventually eases the consumers' effort and time from reading numerous reviews and trusting inaccurate product rating to find a best quality product for them.**

Keywords*—*e-commerce website; sentiment analysis technique; manual product rating; automated product rating; product review

I. Introduction

As technology advances, everything is becoming more digital and automated. In this digitalized world, all of us now can buy anything on e-commerce websites and have it delivered without having to go to a store physically. There is almost nothing that we cannot find on e-commerce websites. However, judging a product solely based on its pictures and reviews is difficult and requires extensive research. Online buyers will always look for reviews of a product before buying to avoid purchasing low quality products. For them to look for the right product to be purchased, they must look for the products from different shops and compare those products. Reading numerous reviews and selecting the best product from a sea of similar items is a time-consuming process. Every review must undergo a thorough examination. The buyers had to scroll through and analyze thousands of reviews to buy a product online. Not only that, even when there is a rating provided for the product, it is not accurate enough as the buyer can just rate simply and not everyone knows the reason behind the rating that is given. Some reviews and ratings that are available in certain e-commerce websites are not even related. The reviews may be written in a positive way, but the rating given may be too low and vice versa. This makes the other buyers confused to look whether to refer to the reviews or the ratings. They would be frustrated in looking for that kind of unmatching reviews and ratings. Furthermore, the arrangement of the products is unstructured. Even though a product has many good reviews or ratings, it is shown at the bottom of the page. We are not sure how many people would scroll till the bottom to find a perfect product for them. They will only be going a certain mile to find the products and read the reviews. So, by the time the buyer does not reach the product that has a good review, they are losing the chances of buying the good ones. Thus, an advanced Sentiment Analysis for Product Rating system is developed to speed up finding the best products online using Sentiment Analysis technique. In addition, the system is also able to detect the hidden sentiments in the reviews and rate the product accordingly in which it will be done automatically by the system. Using the rating that is given for each review, an overall rating is assigned for each product. Based on the overall rating, it will rank the products from the highest to lowest product rating when displaying the products. By doing this, the true customer sentiment will be shown in the form of rating, and the correct and best product can be purchased by the consumer easily. To understand better, an effort was made to look for pertinent papers, research papers, reports, and documents that are related to the topic.

II. e-Commerce

Electronic commerce refers to the buying, selling, and exchanging of goods and services over the internet. e-commerce is a contemporary business that caters to the needs of organizations and the customers able to purchase products when the prices drop, when the product quality improves and the delivering services pace enhances [1]. Moreover, Abdullah et al. [2] portrayed e-commerce as the use of Internet, computer and shared software technology to exchange the product details and visuals, offers and purchasing information and any other information that requires to be communicated to customers, suppliers or the society. Online store sales are roughly equivalent to the physical stores and this trend would continue without a stop [3]. To add on to the previous point, Franco & Regi [4] claimed that the customers can easily compare the products that they wanted to be purchased

without burdening themselves to drive from shop to shop. Armando & Alberto [5] mentioned that there was a massive shift in spending towards e-commerce. The shopping rate is also hugely elevated. However, the customers always wanted and always rely on the reviews or ratings that a product has before they decide to purchase the product. This is because when buying products online, the product cannot be touched or sensed physically, therefore the only way to make the customers trust and buy a product is the feedback or reviews given by the other customers who have bought the product earlier. Nellutla et al. [6] mentioned that a typical customer goes to the website, selects a product, inspects the prices and ratings, reads the reviews, and then proceeds with the transaction. However, e-commerce has its own set of issues. One of them is the rating of the products. Thus, we can tell that e-commerce is consistently growing and becoming more significant to businesses as technology constantly advances and it is something that needs to be taken advantage of and implemented. Customers are playing a vital role in purchasing products online as their honest reviews or the information that they are spreading are eventually influencing other customers who want to purchase a product online.

III. Product Rating and Review

Product rating and review are becoming more important in the customer decision making process nowadays. Buyers are relying on the reviews and ratings that are posted by the other buyers who have bought the product. As a result, the customer ratings and reviews have the power to significantly influence the sales of the product [7]. According to Fawzy et al. [8], the product rating is critical in ascertaining whether a product or service is in a good quality as publicized and can be trusted. In addition, Fan & Fuel in [9] mentioned that 94% of customers read reviews before deciding to do any online purchasing. To add on, Wang & Chen [10] also stated that reviews in e-commerce websites are a valuable resource and play a vital role in purchasing a product. Customers decide either to proceed with the transaction or drop down based on the reviews in the e-commerce websites. In addition, Lackermair et al. [11] stated that 104 of customers in Germany reported that roughly 85% of the customers view and read all the product reviews very frequently before purchasing a product. To conclude, ratings and reviews aid the customers as they help them to get a better and clear idea of a product before buying it. Before deciding whether a product is worth purchasing, the customers analyze the reviews and ratings in the comment section and make themselves clear. Unfortunately, the product ratings are inaccurate to resemble the quality of the products as some of the customers simply put some random ratings without having a proper rationale behind it.

IV. Sentiment Analysis Technique

To solve the problems such as unmatched product ratings with the reviews and inaccurate product ratings, Natural Language Processing (NLP) was decided to use as it has the power of measuring sentiment from a text. NLP is a branch of machine learning (more specifically the branch of Artificial Intelligence) that is concerned with a computer's ability to comprehend, analyze, manipulate, and possibly generate human language [12]. In NLP, there is a technique called Sentiment Analysis which suits well to achieve the objective of this topic. To use the Sentiment Analysis technique, some other NLP techniques or processes needed to be used so that the objective of Sentiment Analysis can be achieved (will be discussed in the upcoming section). Based on the research that is made, there are no other techniques that are used to analyze a review. The only technique that has been introduced to analyze the sentiment of a review is Sentiment Analysis. There are many reasons why Sentiment Analysis is decided to be used for analyzing the product reviews and generating product rating in accordance with the reviews. Those are described in the next subsection.

V. Sentiment Analysis in e-Commerce Website

Nowadays, people place a lot of importance on online shopping because it allows them to complete their purchases faster and with less effort. According to Jha et al. [13], Sentiment Analysis is used to ascertain what customers think of the product. This helps other customers in making decisions on buying a product. Textual reviews, star ratings, and emojis are all used to express opinions. Thus, Sentiment Analysis is used to examine the vast amounts of data that assist retailers or the buyers in meeting their objectives. Sentiment Analysis, a cognitive process for eliciting a user's feelings and emotions. As the internet-based applications have increasingly evolved, it has resulted in a huge number of personalized reviews for many types of information on the Web. Sentiment Analysis is also known as a mighty tool for extracting relevant and needed information including aggregating the sentiments of the reviews for the users. It requires a training set for its performance and its quality is utmost important in evaluating a text accurately [14]. Besides that, Jabreel et al. [15] stated that the main goal of Sentiment Analysis is to predict whether a text's overall sentiment is positive, negative, or neutral. It has numerous variations. One method is to assign a rating scale from the reviews, such as 1 = "worst" to 5 = "best". According to Vyas & Vijayasundaram [16], big data from customer reviews, e-commerce sites and other sites are nearly impossible to manage. An automated system that calculates the overall tendency of belief and intensity to units like agencies, manufactured items, events, and their components is Sentiment Analysis. Sentiment Analysis is necessary for a greater understanding of the product. Furthermore, the simplest way to analyze the reviews is by calculating the feedback rating using Sentiment Analysis with word count. The rating can be predicted by the customer feedback. After receiving the sentiment analysis output, the customer can read all the feedback as fast and efficiently as possible in terms o [17].

Not only that, Haque et al. [18] mentioned that there are so many reviews that an effective method of analysis is needed. Customers must read through thousands of reviews manually to purchase a product in e-commerce. The volume of reviews is stored like a mountain which requires some effective classifier to identify valuable information from text. Sentiment Analysis can be useful in determining customer behavior by analyzing and examining customer reviews in e-commerce. Customers express their feelings by providing a subjective judgment about the products in e-commerce [19].

Additionally, Sentiment Analysis aids in categorizing the unstructured text as positive, negative, or neutral, which summarizes customer opinions and helps us better understand how other customers feel about a given product and retailer [20]. To summarize, Sentiment Analysis plays an essential role in analyzing the hidden and true sentiment of customers based on the feedback or reviews given by them.

VI. Comparison on Existing e-Commerce System

A comparison on some existing e-commerce websites such as Shopee, Lazada and Amazon in some aspects of their product ratings are also made. Those are shown in Table I.

TABLE I. Comparison of the Existing e-Commerce Websites with Product Rating Feature

Criteria to assess	Shopee	Lazada	Amazon	Proposed system
Existence of product rating feature	Yes	Yes	Yes	Yes
Automated product rating based on product review	No	No	No	Yes
Technique used in product rating	No technique used	No technique used	No technique used	Sentiment analysis will be used to analyse the true sentiment of the customers based on the product review given by them
Matchiness of product rating and product review	Not match sometimes	Not match sometimes	Not match sometimes	Will be matched as the product rating will be automated based on the product review given
Arrangement of products based on product rating	Random	Random	Random	Will be orderly arranged automatically based on the product ratings (from high to low)

VII. Method

Interviews and surveys were conducted to find out the problems faced by consumers when purchasing a product online through e-commerce websites.

A. Interview

The interview was conducted with two frequent users of e-commerce websites who are X, regular user of Shopee and she is a student who's running a small jewelry business and another interviewee is Y, active user of Lazada and she is a businesswoman who's having her own boutique. The interviews were conducted by asking open-ended questions without providing any options. This is because answering open-ended questions will reveal the truth of interviewees' problems so that all their pains and challenges of purchasing products online can be understood well. By using qualitative method, the interviewees' problems were clear-cut. It creates an opportunity to understand the interviewees in depth. Not only that, but the behavior of the interviewees was also easily observed during the interviews. This was helpful to understand them especially their pains and gains because those were expressed through their expressions and their body languages. In addition, 'WH' questions were asked to the interviewees to uncover deeper meaning and to identify the problems that the interviewees had but never realized its existence. Furthermore, it was effective enough to immerse in the interviewees' experience by purchasing products online through the e-commerce website together with them. The prepared questions were asked and there were some extra questions added based on the response given by the interviewees. The online interview sessions were conducted for 20 to 45 minutes.

B. Survey

Besides, a simple online short survey was conducted through Google form with a set of questionnaires. Since this issue involves e-commerce website users which are also known as consumers, the survey has been conducted among the people who actively use e-commerce websites to purchase their desired products. Based on the survey, there are a total of 20 respondents that have answered the questionnaire.

C. Sentiment Analysis Process

Since it involves the Sentiment Analysis technique, there are some processes that have performed so that it automatically analyzes the reviews for the sentiment of the consumers and rate according to it. Those processes are as below.

D. Data Collection

Dataset was collected from the Kaggle platform. The dataset is about clothing reviews with its product ratings and some relevant information. It is the raw data. And it is in the .csv format.

E. Data Reading

After collecting the dataset from the Kaggle platform, the data has to be read for the upcoming process. Then, the number of rows, the number of columns and the number of data were read. For this project, the amount of data for every star-rating has been read and displayed which will be useful for the upcoming processes.

F. Data Cleaning

Data cleaning was done so that the data will be accurate, consistent and complete. Data cleaning or cleansing is the process of cleaning datasets by accounting for missing values, removing outliers, and smoothing noisy data (removing the meaningless data, renaming the meaningless column names and removing duplicated data) [21]. At the end, only the product reviews column and product ratings column were retained.

G. Data Preprocessing

Data preprocessing is the process of transforming raw data into a useful, understandable format. Raw data usually has inconsistent formatting, human errors, and can also be incomplete [22]. In data preprocessing, firstly, tokenization process was implemented. Tokenization is a process that will break the raw text into small chunks which helps in understanding the context or developing the model for the NLP [23]. It also helps in interpreting the meaning of the text by analyzing the sequence of the words. Then, special characters such as punctuations were eliminated as those characters are less important for training and testing models later on. Next, stop words were removed in which it removes the words that occur commonly across all the documents in the corpus. Typically, articles and pronouns are generally classified as stop words. Moreover, stemming process was also performed. It is also known as data filtering. It is a process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.

H. Feature Extraction

Feature extraction is a process that will convert text into a matrix (or vector) of features. For this, we will be using a technique called TF-IDF technique in which it stands for term frequency-inverse document frequency. It highlights a specific issue which might not be too frequent in our corpus but holds great importance [24]. The TF– IDF value increases proportionally to the number of times a word appears in the document and decreases with the number of documents in the corpus that contain the word. It is composed of two sub-parts, which are Term Frequency (TF) and Inverse Document Frequency (IDF). Term Frequency (TF) specifies how frequent a term appears in the entire document while the Inverse Document Frequency (IDF) is a measure of whether a term is rare or frequent across the documents in the entire corpus. The formula of TF-IDF is:

$$\text{TF}(t, d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d} \quad (1)$$

$$\text{IDF}(t) = \log\frac{N}{1+df} \quad (2)$$

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) * \text{IDF}(t) \quad (3)$$

where, d refers to a document, N is the total number of documents, df is the number of documents with term t.

TF-IDF is word frequency scores that highlight the words that are more interesting. The scores have the effect of highlighting words that are unique in a given document.

I. Data Balancing

Data balancing is used to balance the imbalanced data. There are two methods to be used to solve the imbalanced data. One is a random under sampling method, and another is a random oversampling method. Random under sampling method aims to randomly choose and eliminate samples from the majority class, thereby reducing the number of examples in the majority class in the transformed data whereas random oversampling involves choosing random examples from the minority class with replacement and adding multiple copies of this instance to the training data, so it's possible that a single instance will be chosen more than once [25].

J. Model Building (Sentiment Classification Analysis)

There are several Machine Learning based classification algorithms available using supervised and unsupervised learning approaches. The supervised methods make use of a large number of labeled training documents while the unsupervised methods are used when it is difficult to find these labeled training documents. For this proposed system, supervised learning method is chosen as it has a labeled dataset and a training process, therefore, several models were used such as Gaussian Naïve Bayes Classifier, Multinomial Naïve Bayes Classifier, Bernoulli Naïve Bayes Classifier, Random Forest Classifier, Logistic Regression Classifier, Decision Tree Classifier, Support Vector Classifier and K-neighbors Classifier. All these mentioned models were decided to use as they are capable of scalability. Using all these models, accuracy of each model was then identified. From there, the highest accuracy model was chosen and the dataset using the chosen model was trained.

K. Sentiment Analysis

With the highest accuracy model, Sentiment Analysis is then carried out to analyze the sentiments in each review and provide appropriate rating to it.

VIII. Results and Discussion

The results of the interviews and survey are discussed in this section.

A. Interview

In the first interview, X mentioned that she uses Shopee because it is user-friendly. In Shopee, she normally browses the same products in distinct shops and compares the price and the quality of the products by looking at the reviews and ratings. She often does online shopping in Shoppe (twice a week) to buy things for herself and for her business. She said it takes almost two to three days to view all the product features and compare them with other sellers because the same product from other shops has different kinds of reviews. Before she decides to purchase a product, she normally goes through the product features, product rating and the product reviews. As for her, purchasing a product online consumes her time as it takes time to compare the products with different shops and choose the right product. She also mentioned that she will not purchase a product if she found the product has a huge number of negative reviews and even if it has an equal number of positive and negative reviews. Furthermore, she found some of the ratings are displayed as high but portrayed bad about the product when reading the reviews. She said that seeing and reading a ton of reviews are really time consuming, as a result, she will not be in the mood to shop most of the time.

In the second interview, Y said that she uses Lazada because it is easier for her to use, and it was recommended by her friend. In Lazada, she usually buys her desired products, reads reviews of her desired product before purchasing and sees the ratings of the product. Y mentioned that she does online shopping in Lazada once a week mainly to buy some

necessary items for her business. She said it takes almost a day to make a purchasing decision. This is because she always trusts the people's views, thus, she will always read the reviews before purchasing anything and she also said that some of the products have a ton of reviews which is time consuming. Before she wants to buy anything, she usually looks for people's thoughts in the review section. As for her, purchasing a product online consumes more time because she must go through a lot of reviews. She also mentioned that if she found the product has a huge number of negative reviews and even it has an equal number of positive and negative reviews, she normally will not buy the product in that shop and will find another shop that sells the same product which has more positive reviews. In addition, trusting too much on the product reviews including the product rating before purchasing a product became a bigger problem for her because she often confused with the product reviews and ratings as they both will not be correlated and she does not know which one to trust, therefore, she will move to other shop to purchase the same product. Furthermore, she found that there are unmatching reviews and ratings of a product in which the rating is low for positive reviews and vice versa. She said that seeing and reading a ton of reviews are really time consuming. As a result, her excitement of buying the product will be lessened in that shop and will decide to shop in different shops.

B. Survey

There are roughly 15 questions in the survey, but the result of this survey will be displayed only for some questions that are important for achieving the objective of this project. Those are as below:

Fig. 1 shows a bar chart of the e-commerce websites that are used by the respondents daily. Based on this bar chart, 18 of the respondents use Shoppe websites, 20 of the respondents use Lazada websites and 11 of them use Amazon websites. Since Shopee, Lazada and Amazon are the three highest e-commerce websites used by the respondents, that's why a comparison of those three e-commerce websites was made in the previous section (see Table I).

Fig. 2 shows a bar chart of the activities that are carried by the respondents in the e-commerce websites. It is clearly be seen that reading other users' reviews of a product and looking at the rating of a product have the highest counts which means the respondents always look for the reviews and ratings when they shop.

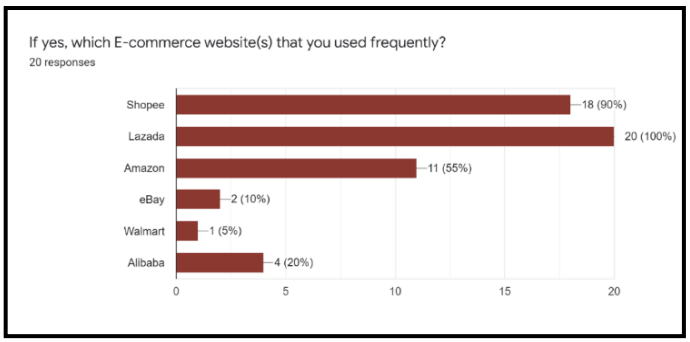


Fig. 1. Frequently used e-commerce website(s).

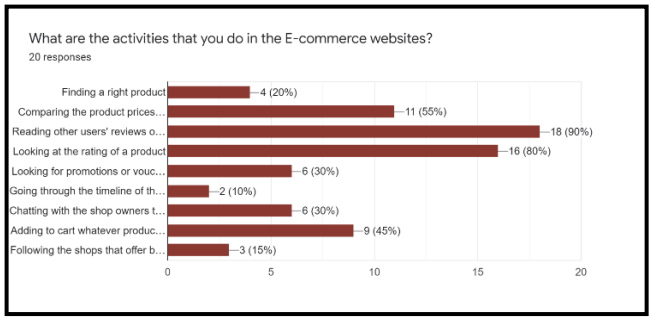


Fig. 2. Activities carried out in e-commerce websites.

Fig. 3 shows a pie chart on the time taken to make a purchasing decision while shopping online by the respondents. Around 40% of the respondents took more than five hours to decide while 25% of the respondents took more than 1 day to decide and followed by the rest. From here, it is obvious that spending more than five hours or one day is too much for a person to decide on a product to purchase. Even the respondents would not spend that much time when they shop physically. It would take only 10-15 minutes to buy physically. Tons of reviews to be read might be one of the reasons why most of the respondents spend too much of their time to make decision.

Fig. 4 shows a bar chart on the criteria(s) that are considered by the respondents before they proceed with purchasing a product. From this chart, reading the product reviews and looking at the product ratings have the same count which is 19 and both are the highest among the other criteria. Thus, it is crystal clear that reviews and ratings are playing a major role when purchasing a product in e-commerce websites.

Fig. 3. Time taken to make a purchasing decision.

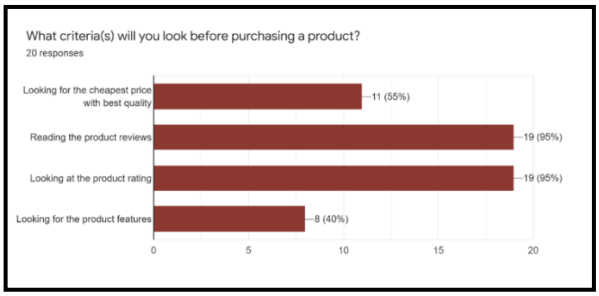


Fig. 4. Criteria(s) that are considered before purchasing a product.

Fig. 5 shows a pie chart on the time taken the respondents took to read all the related reviews of a product. 60% of the respondents took more than one hour to read the reviews.

There are also some respondents who voted for 10 – 30 mins and 30 mins - 1 hour. This is mainly because of the number of reviews a product has. The more the reviews a product has, the more the time taken to read the reviews one by one.

Fig. 6 shows a pie chart to know whether the respondents experience unmatched product reviews and ratings. All the respondents voted for 'Yes' which means all of them experienced seeing unmatched product reviews together with manual product ratings. This is meant by, either they have seen a good review with low product rating or a bad review with a high product rating which both do not make sense at all. The reviews and ratings should be related with each other to avoid creating distrust among the other consumers who firstly visited a particular shop.

Fig. 7 and Fig. 8 show pie charts on difficulty and time consumption in finding and purchasing a right product. All 20 respondents agreed that finding and buying the best product for them is really time consuming and very difficult. This is maybe because finding a product that has the highest product rating on an unordered products arrangement consumes more time and a mixture of positive and negative reviews of all the similar products from different shops make the consumers think in depth to proceed with the purchasing decision.

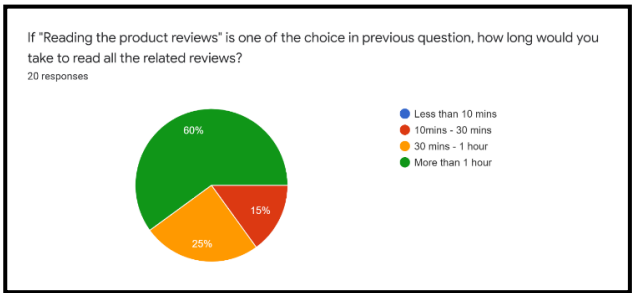


Fig. 5. Time taken to read all the reviews.

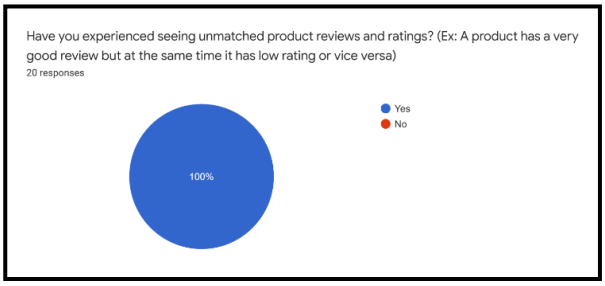


Fig. 6. Experience of seeing unmatched product reviews and ratings.

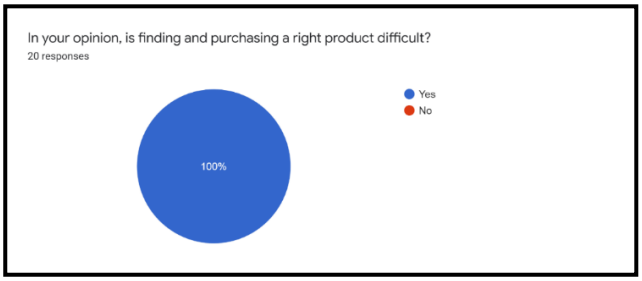


Fig. 7. Difficultness in finding and purchasing a product.

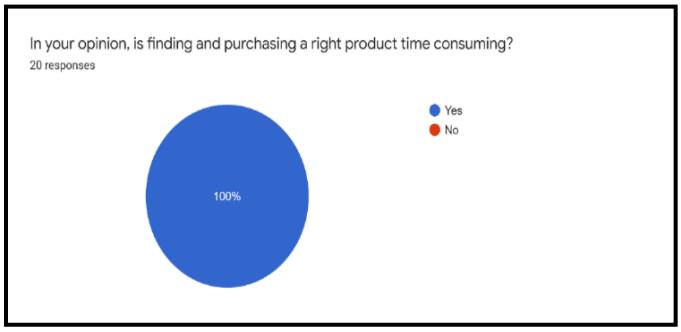


Fig. 8. Time consumption in finding and purchasing a product.

C. Sentiment Analysis Process

The result of each process mentioned in the previous section is discussed as below.

D. Data Collection

Clothing Reviews dataset was collected from Kaggle platform in .csv format that has the columns named product review and product rating.

E. Data Reading

The number of rows, the number of columns and the number of data counts of every rating value were read using pandas library (python)

F. Data Cleaning

The null values, duplicated values and unwanted columns are removed to make the data accurate and consistent. Additionally, the column names were renamed with meaningful names. All these were done using the drop (columns = […]) method, rename (columns= […]) method, dropna () method and drop_duplicates () method.

G. Data Pre-Processing

In this process, label encoder is used to encode the rating value of 1,2,3,4,5 to 0,1,2,3,4,5. Then, using the NLTK library, stop words are imported so that it can remove the stop words found in the dataset. All the unwanted characters like special characters were removed. And made the reviews all in small characters. Moreover, the review was also tokenized into small chunks. And finally, the suffixes and prefixes of a word in the review were removed and only holds the root word.

H. Feature Extraction

Feature extraction technique was used to find the frequency and the importance of the words that are present in the corpus. Here, the TF-IDF technique was implemented from the sklearn library by importing TfidfVectorizer from the feature extraction.text module.

IX. Data Balancing

Upon visualizing the product review with its rating in the dataset, it is found that the number of product reviews for all the product ratings (1-star to 5-star) is not balance as shown in Fig. 9.

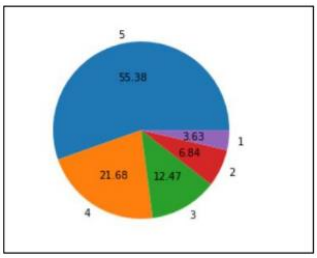


Fig. 9. Pie chart of the number of product reviews for all the product ratings in Clothing review dataset.

This is known as imbalanced data. When the machine learning models are trained using this imbalanced data, the models will be well trained for only 5-star rating reviews as it holds a bigger amount of product reviews compared to the other star ratings. In this case, even a bad review could make the models to interpret the review as a good one and could create a good rating since the product rating of 1-star, 2-star, 3-star have a smaller amount of review data compared to 5-star reviews and it will not be trained as how 5-star product review do. Thus, out of two methods that were mentioned in the previous section which are random oversampling and random under sampling, the random under sampling method is chosen because the random under sampling method gives the accurate result compared to the random oversampling method. This was identified by running both methods and observed the accuracy of the product rating that was generated by the system upon analyzing the reviews that were submitted in the review section.

A. Model Building (Sentiment Classification Analysis)

To find out which classifier suits the dataset the best, we implemented our dataset on multiple classification models namely Multinomial Naive Bayes Classifier, Bernoulli Naive Bayes Classifier, Gaussian Naive Bayes Classifier, Support Vector Classifier, K Nearest Neighbour Classifier, Decision Tree Classifier, Logistic Regression Classifier and Random Forest Classifier.

```
For  Support Vector Classifier
Accuracy -  0.4776910616993079
Time taken -  10.018787860870361

For  K-Neighbors
Accuracy -  0.31438668826387867
Time taken -  10.443702697753906

For  Multinomial Naive Bayes
Accuracy -  0.518038580474157
Time taken -  0.8072564601898193

For  Gaussian Naive Bayes
Accuracy -  0.35016934177587 98
Time taken -  8.906440496444702

For  Bernoulli Naive Bayes
Accuracy -  0.5131792077749964
Time taken -  2.075514316558838

For  Decision Tree
Accuracy -  0.36769253423648945
Time taken -  6.655981779098511

For  Logistic Regression
Accuracy -  0.4919746723604771
Time taken -  1.396794080734253

For  Random Forest
Accuracy -  0.4659107642467972
Time taken -  6.154578685760498


Multinomial Naive Bayes holds the highest accuracy
Accuracy:  0.518038580474157
```

Fig. 10. Accuracy and time taken of each classifier.

The accuracy and time taken of each machine learning model were identified to train the model using each classifier that was mentioned, as shown in Fig. 10. In Fig. 10, it is clear that the Multinomial Naïve Bayes model holds the highest accuracy compared to other models.

B. Sentiment Analysis

The highest accuracy model, which is Multinomial Naïve Bayes model, was selected to analyze the sentiments on each review given and provide an appropriate rating to it.

X. Conclusion

It is true that reading numerous reviews, finding a right product for a longer time, and having unmatched reviews and ratings are the top issues among the consumers and there is a dire need to solve them. These are the key reasons to have an automated product rating system and different from the existing e-commerce websites. It uses Sentiment Analysis technique as it has an ability to analyze the hidden sentiment of the consumers in the review section. In addition, this system can also rank the products based on the rating in which it will reduce the time taken for the consumers to scroll the page down till they find the right one for them. On top of that, the unmatched reviews and ratings can be avoided as the rating will be given automatically based on the reviews, hence helps the consumers shorten their time searching for the best products within minutes.

Future Works

A fundamental challenge in e-commerce research comprehends the idea of Artificial Intelligence (AI). According to our research, e-commerce product ratings based on customer reviews have primarily been taken into consideration by researchers. Researchers can contribute AI to e-commerce research in the future. This will assist in precisely analyzing data and forecasting e-commerce activity through the use of AI platform algorithms. Having this clarification would help avoid misunderstandings between AI and business analytics and intelligence in e-commerce. Additionally, it would make it easier to distinguish between AI as a social actor and AI as a computing technology capable of cognitive tasks. Context is a second fundamental problem with AI research in e-commerce. Based on the output of an e-commerce system that is relevant in the real world, an AI system would be able to interpret the message that the user communicated or sought using the same data. Hence, rather than making broad assertions about product prices, researchers would need to work with practitioners to better understand and define contexts of inquiry. Lastly, to provide more information about the methodology, such as sample size, demographics, and questions asked during interviews and surveys.

References

[1] F. Salehi, B. Abdollahbeigi, A. C. Langroudi, and F. Salehi, "The Impact of Website Information Convenience on e-commerce Success of Companies," *ScienceDirect*, vol. 57, pp. 381-387, Oct 2012.

[2] H. Abdullah, I. Ismail, A. Alnoor and E. Yaqoub, "Effect of perceived support on employee's voice behaviour through the work engagement: a moderator role of locus of control," *International Journal of Process Management and Benchmarking*, vol. 11, no. 1, pp. 60-79, Jan. 2020.

[3] A. Nogoev, S. Mohseni, A. P. Yazdanifard and B. Samadi, "The Evolution and Development of e-Commerce Market and e-Cash," in *International Conference on Measurement and Control Engineering 2nd (ICMCE 2011). At: USA*, Oct. 2011.

[4] D. C. Franco and B. Regi, "Advantage and challenges of e-commerce customers and businesses: In indian perspective," *International Journal of Research-Granthaalayah*, vol. 4, Mar. 2016.

[5] R. L. C. Armando and V.-M. J. Alberto, "Disruption in the consumer decision making? Critical analysis of the consumer's decision making and its possible change by the COVID19," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 4, pp. 1468-1480, Apr. 2021.

[6] A. P. Nellutla, M. Hudnurkar, S. S. Ambekar and A. D. Lidbe, "Online Product Reviews and Their Impact on Third Party Sellers Using Natural Language Processing," *International Journal of Business Intelligence Research*, vol. 12, no.1, pp. 26-47, Jan. 2021.

[7] W. W. Moe and M. Trusov, "The Value of Social Dynamics in Online Product Ratings Forums," *Journal of Marketing Research*, vol. 48, no. 3, pp. 444-456, Jun. 2011.

[8] M. Fawzy, R. Sharuddin and W. Zulkifly, "e-commerce adoption and an analysis of the popular e-commerce business sites in Malaysia," *Journal of Internet Banking and Commerce*, vol. 23, no. 1, Apr. 2018.

[9] Fan and Fuel, "No online customer reviews means BIG problems in 2017," 2017.

[10] Z. Wang and Q. Chen, "Monitoring online reviews for reputation fraud campaign," *ScienceDirect*, vol. 195, May. 2020.

[11] G. Lackermair, D. Kailer and K. Kanmaz, "Importance of Online Product Reviews from a Consumer's Perspective," Advances in Economics and Business, vol. 1, no. 1, pp. 1-5, 2013.

[12] "What is natural language processing (NLP)?," *IBM Cloud Education, Jul. 2020*.

[13] B. K. Jha, S.G.G and V. K. R, "Sentiment Analysis for e-Commerce Products Using Natural Language Processing," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 5, pp. 166-175, May 2021.

[14] M. D. D, S. C and A. Ganesh, "Sentiment Analysis: A Comparative Study On Different Approaches," *ScienceDirect*, vol. 87, pp. 44-49, 2016.

[15] M. Jabreel, N. Maaroof, A. Valls and A. Moreno, "Introducing Sentiment Analysis of Textual Reviews in a Multi-Criteria Decision Aid System," *applied sciences*, vol. 11, no.1, Dec. 2020.

[16] V. Vyas and U. Vijayasundaram, "Approaches to Sentiment Analysis on Product Reviews," *ResearchGate*, Jun. 2018.

[17] Dr. S. P and G. K, "Mining of Customer Review Feedback Using Sentiment Analysis for SmartPhone Product," *Turkish Journal of Computer and Mathematics Education, vol. 12*, no. 10, pp. 5515-5523, 2021.

[18] T. U. Haque, N. N. Saber and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," in *2018 IEEE International Conference on Innovative Research and Development (ICIRD). At: Bangkok, Thailand*, Jun. 2018.

[19] P. K. Sari, A. Alamsyah and S. Wibowo, "Measuring e-Commerce Service Quality from Online Customer Review using Sentiment Analysis (Case Study: Tokopedia)," *Journal of Physics: Conference Series*, vol. 971, no. 1, 2018.

[20] A. H. Alamoodi, O. S. Albahri and B. Bahaa, "Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review," *Expert Systems with Application*, vol. 167, Oct. 2020.

[21] F. Ridzuan and W. M. N. Wan Zainon, "A Review on Data Cleansing Methods for Big Data," *The Fifth Information Systems International Conference 2019*, vol. 161, pp. 731-738, 2019.

[22] V. Agarwal, "Research on Data Preprocessing and Categorization Technique for smartphone Review Analysis," *International Journal of Computer Applications*, vol. 131, no. 4, pp. 30-36, Dec. 2015.

[23] R. Friedman, "Tokenization in the Theory of Knowledge," *Encyclopedia*, vol. 3, no. 1, pp. 380-386, Mar. 2023,

[24] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25-29, Jul. 2018.

[25] S. Kotsiantis, D. Kanellopoulos and P. E. Pintelas, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering*, vol. 30, pp. 25-36, Nov. 2005.

Secure IoT Routing through Manifold Criterion Trust Evaluation using Ant Colony Optimization

Afsah Sharmin[1], Rashidah Funke Olanrewaju[2], Burhan Ul Islam Khan[3*], Farhat Anwar[4], S.M.A. Motakabber[5], Nur Fatin Liyana Mohd Rosely[6], Aisha Hassan Abdalla Hashim[7]

Department of ECE, Kulliyyah of Engineering, International Islamic University Malaysia, Kuala Lumpur, Malaysia[1, 2, 4, 5, 7]
Department of CST, Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia[3]
Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia[6]

***Abstract*—The paper presents a simplified yet innovative computational framework to enable secure routing for sensors within a vast and dynamic Internet of Things (IoT) environment. In the proposed design methodology, a unique trust evaluation scheme utilizing a modified version of Ant Colony Optimization (ACO) is introduced. This scheme formulates a manifold criterion for secure data transmission, optimizing the sensor's residual energy and trust score. A distinctive pheromone management is devised using trust score and residual energy. Concurrently, several attributes are employed for constraint modeling to determine a secure data transmission path among the IoT sensors. Moreover, the trust model introduces a dual-tiered system of primary and secondary trust evaluations, enhancing reliability towards securing trusted nodes and alleviating trust-based discrepancies. The comprehensive implementation of the proposed integrates mathematical modeling, leveraging a streamlined bioinspired approach of the revised ACO using crowding distance. Quantitative results demonstrate that our approach yields a 35% improvement in throughput, an 89% reduction in delay, a 54% decrease in energy consumption, and a 73% enhancement in processing speed compared to prevailing secure routing protocols. Additionally, the model introduces an efficient asynchronous updating rule for local and global pheromones, ensuring greater trust in secure data propagation in IoT.**

Keywords*—Internet of things (IoT); secure IoT routing; manifold criterion trust evaluation; ant colony optimization (ACO); bioinspired computing; pheromone management

I. Introduction

Sensors are an integral part of the IoT landscape. They acquire environmental data and perform real-time transmission over the hosted IoT network [1]. These compact electronic devices can sense various environmental attributes, such as chemical composition, motion, sound, light, pressure, humidity, and temperature, depending on their application or the environment in which they are deployed [2–5]. In the context of IoT, sensors acquire information from the physical world and forward it to a sink node for analysis [6]. This acquired information might be used for research purposes or to trigger specific actuators for automated actions [7–10]. However, despite their capabilities, these sensors often have limited processing and computing abilities, and ensuring their extended lifespan remains challenging [11].

Among the various issues associated with sensors, security is the most critical concern for IoT sensors [12]. The first issue is data privacy, ensuring data is stored securely and transmitted to the intended terminal without unauthorized access [13]. The second pertains to vulnerabilities in IoT devices; sensors often fall prey to cyber-attacks due to unpatched vulnerabilities, outdated software, or weak passwords [14]. The third challenge relates to malware attacks in the form of Trojans, worms, and viruses, which can hinder data transmission, steal data, or corrupt device functionalities [15]. Physical security represents the fourth concern; unauthorized access to a sensor can lead to data tampering, malware introduction, or functionality disruption [16]. The fifth challenge involves Distributed Denial-of-Service (DDoS) attacks, where overwhelming traffic incapacitates sensors, disrupting their communication or services [17]. The sixth pertains to Man-in-The-Middle attacks, where attackers can intercept and potentially modify or steal data [18]. Numerous studies have proposed security mechanisms for IoT to address these vulnerabilities [19–23], but comprehensive solutions that tackle all these challenges remain elusive, with each approach having its strengths and weaknesses.

Securing routing in IoT has become a complex endeavor in today's landscape, marked by an ever-increasing array of known and emerging threats [24, 25]. One core challenge arises from the use of resource-constrained devices in IoT, characterized by limited battery life, memory, and computing power. This limitation precludes the deployment of robust cryptographic algorithms on such devices [26]. The dynamic topology of IoT, where devices can spontaneously join or leave the network, further complicates security protocols. Incorporating diverse IoT devices with specific service requirements complicates implementing universal security protocols. Challenges also arise from issues with mobility and localization; accurate localization information is hard to obtain, and the legitimacy of mobile nodes is difficult to verify. The need for standardization in IoT devices and security protocol management further complicates matters. Most current IoT security research is conducted in controlled environments, distinct from real-world scenarios. Thus, the reliability of their applications in practical deployments remains to be determined.

The manuscript is structured as follows: Section II delves into current methodologies for secure routing in IoT, emphasizing various bioinspired approaches and their

*Corresponding Author

This research was supported by the INTI IU Research Seeding Grant Phase 1/2023 initiative under Project No: INTI-FDSIT-01-01-2023.

contributions. Section III outlines the research problem identified from insights gleaned from these existing methodologies. The proposed methodology to address these issues is detailed in Section IV. An analysis of the results is presented in Section V, while Section VI concludes the paper.

II. Existing Approaches

Over time, several schemes have been developed to investigate secure routing within IoT. Among these, trust-based schemes have emerged as an essential tool for a simplified defense against security breaches [27, 28]. Liu *et al.* [29] devised a secure aggregation model for a Wireless Sensor Network (WSN), ensuring increased trust when operated with a single mobile sink node. Additionally, some research has focused on optimizing trust factors through bioinspired approaches. For instance, Mangalampalli *et al.* [30] employed a whale optimization scheme for enhancing task scheduling. Muzammal *et al.* [32] introduced a trust-based protocol to counter blackhole attacks in static and mobile IoT environments. However, these approaches often need to pay more attention to the dynamic nature of IoT networks and may not be efficient in real-time scenarios.

Awan *et al.* [33] embraced a blockchain-based model for secure routing in WSN, aiming to refine trust management. Notably, their model incorporated the Rivest Shamir Algorithm to safeguard data propagation. However, blockchain's inherent latency issues could limit its practicality in specific IoT setups. Nagaraju *et al.* [34] combined energy optimization with a traditional hybrid approach for secure IoT routing within heterogeneous WSNs. While energy-efficient, such models might compromise on real-time response. Bakhtiari *et al.* [35] proposed a two-way trust strategy using Bayesian learning automata for fog computing in IoT. It improves efficiency, reduces latency, and enhances trust calculations compared to existing methods. However, it's sensitive to initial trust accuracy, potentially impacting performance in dynamic networks. Also, implementing this two-way trust management strategy may introduce increased computational complexity as a potential limitation. Concurrently, Rakesh and Sultana [36] designed a neural network-empowered quantum scheme for mobile sink selection, employing the sailfish optimization approach for enhanced route security. However, this approach might demand more computational resources, impacting resource-constrained IoT devices. Additionally, its effectiveness may vary based on network conditions and the presence of malicious nodes, with reliance on initial trust assessments potentially limiting performance in dynamic environments. Gladkov *et al.* [37] championed a novel routing technique merging the residual number of redundant systems with a secret sharing scheme. However, the complexity associated with such hybrid approaches might lead to higher computational overheads. These diverse strategies enhance security in IoT and WSNs, but careful consideration of their computational requirements and adaptability to dynamic networks is crucial during implementation. Balancing security with resource constraints remains an ongoing challenge in these technologies.

Ramaswamy and Norman [38] introduced a trust model targeting network longevity and internal attacks in IoT. The exploration of bioinspired methods for secure IoT routing has seen algorithms like ACO applied in secured routing (Wang [39], Saleem & Ahmad [40]), decentralized traffic management (Nguyen and Jung [41]), electric vehicle selection (Ajinappa and Prabhakar [42]), and malware detection (El-Ghamry *et al.* [43]). While ACO's adaptive nature is commendable, it may struggle with large-scale, dynamic IoT environments due to its iterative nature. Particle Swarm Optimization (PSO), as documented by Alterazi *et al.* [44], Lin *et al.* [45], and Rajeshwari & Ramakrishnan [46], has also been harnessed for secure data transmission. Other notable bioinspired techniques include the Mayfly Optimization Algorithm (MOA) by Janani and Ramamoorthy [47], the Dragonfly Algorithm (DA) by Hosseinzadeh *et al.* [48], and Glowworm Swarm Optimization (GSO) by Selvaraj *et al.* [49]. Although these bioinspired strategies offer unique solutions, their scalability and adaptability in diverse IoT ecosystems might be limiting factors. The summary of strength and weakness of the reviewed literature is presented in Table I to state that existing methodologies towards secure routing in IoT is associated with beneficial features as well as shortcomings, which are required to be addressed in future upcoming series of research work.

TABLE I. Summary of Existing Schemes

Author	Problem	Methodology	Advantage	Limitation
Liu *et al.* [29]	Security in WSN, IoT	Trust-based secure data aggregation	• Trust-based secure data aggregation mechanism. • Real-time and accurate data acquisition. • Robust aggregation tree algorithm for efficient data gathering. • Enhanced network performance, including improved accuracy and reduced delay.	• Possibility of contradiction in indirect trust. • Scalability challenges in large-scale Industrial Internet of Things (IIoT) settings. • High implementation complexity. • Limited consideration for mobile sensor nodes. • Need more focus on energy efficiency.
Mangalampalli *et al.* [30]	Task scheduling with trust	Whale Optimization	• Enhanced task scheduling efficiency. • Reduced makespan. • Lower energy consumption. • Improved quality of service. • Increased trust in cloud service providers.	• Parameter sensitivity. • Implementation complexity. • Scalability issues. • Input data dependency. • Workload-driven performance variations. • Setup-specific fine-tuning. • Lack of real-world validation.

Muzammal *et al.* [31,32]	Resisting routing attacks	Trust-based protocol	• Improved security against Routing Protocol for Low-Power and Lossy Network (RPL) attacks. • Tailored for mobile IoT environments with trust and mobility metrics. • Superior performance in packet loss rate, throughput, and topology stability. • Meets consistency, optimality, and loop-freeness requirements. • Better throughput.	• Increased computational resource demands. • Sensitive to network size and setup. • Limited real-world IoT testing. • Protocol implementation complexity. • Requires further assessment in highly dynamic IoT scenarios. • Relies on trust and mobility metrics' accuracy.
Awan *et al.* [33]	Secure data transmission	Trust model, Asymmetric encryption	• High delivery ratio. • Enhanced security through blockchain-based authentication. • Extended network lifespan and reduced packet loss. • Effective malicious node detection and removal. • Secure routing based on residual energy and trust. • High packet delivery ratio in simulations.	• Increased key size. • Blockchain dependency may add complexity. • Scalability issues in more extensive networks. • Lack of real-world validation. • Computational resource demands. • Limited scope beyond trust assessment. • Latency potential in real-time applications. • Ongoing trust monitoring is required.
Nagaraju *et al.* [34]	Security, energy issues	Multipath link routing, hybrid protocol	• Satisfactory network lifetime. • Improved energy efficiency in heterogeneous WSNs. • Enhanced network lifetime. • Secure routing for confidential IoT data. • Load balancing capability. • Improved data storage capacity.	• Cannot sustain dynamic threats. • Limited real-world validation. • Dependence on specific routing protocols. • Possible sensitivity to network dynamics. • Lack of consideration for scalability in large-scale deployments.
Bakhtiari *et al.* [35]	Trust issues in Fog and IoT	Bayesian-based learning automata	• Faster response time. • Improved energy consumption. • Efficient network usage. • Reduced latency. • Enhanced trust management.	• Induces complexity for an extensive network. • Limited real-world validation. • Possible sensitivity to network dynamics. • Dependence on specific trust management methods. • May require fine-tuning for different IoT applications.
Rakesh and Sultana [36]	Trust issues	Sailfish optimization, Neural Network	• Can mitigate multiple attacks. • Improved node reliability with the introduction of a mobile sink. • Secure routing implemented using the sailfish optimization algorithm. • Data encryption for increased data security.	• Consumes higher memory to retain trust values. • Limited real-world validation and scalability considerations. • Network dynamics and computational resource dependencies. • Complexity in implementation. • Emphasis on specific optimization algorithms.
Gladkov *et al.* [37]	Routing reliability	Secret sharing scheme	• Highly adaptive. • Enhanced speed and reliability in data transmission. • Improved security through Secret Sharing Schemes (SSS). • Enhanced fault tolerance and reliability. • Adaptive multipath secured transmission for route attack prevention.	• No consideration of energy constraints. • Possible complexity in implementing SSS and RRNS. • Dependency on adaptive multipath secured transmission. • Scalability needs to be fully addressed. • Limited evaluation in dynamic heterogeneous networks.
Ramaswamy [38]	Energy issues	Trust model	• Secured clustering	• Extensive analysis is needed further.
Wang [39], Saleem and Ahmad [40], Nguyen and Jung [41], Ajinappa and Prabhakar [42], El-Ghamry *et al.* [43]	Security issues in IoT	ACO	• Higher accuracy	• Inferior convergence speed.
Alterazi *et al.* [44], Lin *et al.* [45], Rajeshwari [46]	Trust/security in IoT	PSO	• Faster performance	• High dimensional space issue needs to be addressed.

Janani and Ramamoorthy [47]), Hosseinzadeh *et al.* [48], Selvaraj *et al.* [49]	Secure data transmission in IoT	MOA, DA, GSO	• Flexible performance	• Highly iterative scheme leading to complexity.

From the highlights of the methodologies, their advantages, and limitations in the table above, it can be noted that existing secured routing methodologies are associated with various shortcomings in the perspective of deployment in an IoT environment. Following are some of the significant research problems related to the existing methods towards secure routing in IoT:

- Issues in Trust-based IoT Security: Various trust-based secure routing schemes are formulated in existing systems. Most of these techniques offer better throughput; however, the majority are also witnessed by their non-sustainability towards dynamic threats. Furthermore, the existing trust management schemes are usually designed considering the predefined information of adversaries. This makes the model robust in one environment but not applicable if the adversary environment is altered.
- Imbalance between Energy and Security Demands: Existing approaches favor energy or security retention. The security schema presented in existing schemes uses various sophisticated operations that can offer more security but at the cost of uncertain resource consumption. There has yet to be a benchmarked study model to prove this in the presence of the dynamic environment of IoT.
- Usage of Bioinspired Approach: The adoption of bioinspired approaches towards secure routing is less abundant in archives of technical publications. However, the available publications on the use of bioinspired strategies contribute towards findings of the secured path by adopting the varied cognitive principles of organisms to attain optimal security. Unfortunately, the issues related to premature convergence and higher sensitivity towards parameters by conventional bioinspired approaches have not yet been addressed. Moreover, there is no report of any study model where novel features of organisms' cognitive behavior have been attempted to be modified and investigated.
- Non-inclusion of Constraint in Trust Management: The existing trust management scheme needs to be more reportedly designed considering the restricted resources in IoT, e.g., processing power, energy, etc., which makes it quite challenging even to execute the sophisticated security protocols in sensors. This yet-to-solve challenge acts as a potential impediment to computing and establishing trust between all the entities and IoT devices in large and dynamic environments. For secure routing to occur correctly, it is necessary to formulate a better form of constraint modeling with a clear definition of the attributes that need to be added to the existing system.
- The tradeoff between Model Effectiveness and Scalability: Scalability is a different set of problems in IoT to be resolved. It demands a smart and highly planned processing for secure routing in IoT. Unfortunately, none of the existing study models has been designed considering its optimization parameter or problem space mapping with the large environment of an IoT with more interconnected devices. For this purpose, the outcome of model effectiveness doesn't match the sensors' scalable performance in increased traffic flow over an IoT.

From the above-stated highlights of the identified research problem, it can be inferred that trust modeling is one of the complex issues that demand the inclusion of various intrinsic and extrinsic attributes and effective constraint modeling. Apart from this, it is necessary to optimize the problem space to increase the change of optimal outcome of the secured path. Bioinspired algorithms are a potential alternative to address this problem, but they demand a novel inclusion of characteristics that can balance security and resource demands. The proposed solution addresses all these research issues, and its associated methodology towards implementation is discussed next.

III. Research Problem

The rapid proliferation of the IoT heralds both transformative opportunities and notable security challenges. With its expansive network of devices, IoT underscores the urgency for reliable, secure routing. Emerging as a promising avenue, trust-based secure routing strategies aim to buttress IoT's security framework. However, IoT's dynamic and resource-limited nature has often rendered traditional trust-based methods inadequate.

A significant limitation of extant trust-based IoT security methodologies is their inability to adapt to evolving threats, often designed around known adversarial models. The quest to harmonize energy conservation with security amplifies this challenge, as current systems lean towards one, often sacrificing the other. This compromise becomes more palpable when high-security measures, despite their efficacy, devour substantial resources.

The contemporary research landscape needs to display bioinspired strategies for secure IoT routing. Drawing inspiration from natural systems to fortify security, these approaches appear promising. Nevertheless, prevalent bioinspired models grapple with issues like early convergence and parameter sensitivity. This highlights a pressing need to refine and adapt these methods, tailoring them for IoT's unique challenges.

Further complicating the scenario is the design of trust management systems. The intrinsically limited resources of IoT devices, in terms of energy and processing power, hinder the execution of comprehensive security protocols. This challenge intensifies in the sprawling IoT ecosystems, where fostering trust amongst various devices is vital and daunting.

Scalability further accentuates these problems. While effective in a controlled setting, an approach might need to be revised under IoT's expansive and interconnected structure, especially when encountering unexpected traffic surges.

While trust modeling offers a promising foundation for IoT security, its practical deployment is beset with multifaceted challenges. This study aspires to address these gaps, employing the bioinspired ACO technique to architect an adaptive and secure IoT routing paradigm adeptly poised to navigate the intricate challenges of IoT.

IV. Materials and Methods

The prime agenda of the proposed study model is to introduce a novel computational model that ensures robust trust-based security while transmitting data among the sensor nodes in IoT. The prime basis of the proposed study model is based on the fact that the severity degree of security threat for sensor nodes deployed in an IoT environment is comparatively higher in contrast to conventional WSN. Therefore, the security aspect of WSN deployed in IoT is subjected to improvement by balancing the demand for increased security along with energy consumption. Hence, various criterion-based schemes is implemented, harnessing ACO along with constraint consideration towards secure trust evaluation.

From the exhibited methodology in Fig. 1, it can be noted from the declaration of the proposed scheme that it uses a manifold criterion modeling towards trust-based secure routing in an IoT environment. Adopting an optimization agenda towards manifold criteria is simultaneously challenging due to the surfacing possibilities of various conflicts. Moreover, unlike managing unit criterion-based routing improvement strategy, the multiple criterion schemes usually seeks to obtain compromised outcomes, yet another sub-optimal solution towards secure routing. Therefore, the proposed system constructs a manifold criterion-based trust modeling in IoT to address the issue of energy drainage and security threats as follows:

$$arg_{max}\varphi(t) = [\varphi_1(t)\varphi_2(t)] \quad (1)$$

In Eq. (1), the manifold criterion function $\varphi(t)$ is represented by two sub-criteria, i.e., $\varphi_1(t)$ and $\varphi_2(t)$, representing the mean criterion for remnant energy and the mean value of trust of the sensor nodes associated with the routing path in IoT. Two processes follow for this purpose:

- The above expression is required to be satisfied for its remnant energy score $R_i(t) > 0$ for the i^{th} the sensor at t instance of time such that acquired data by the sensor is $AD_{ij}(t) > 0$ from the i^{th} sensor to the j^{th} sensor while forwarding data $FD_{ji}(t) > 0$ from the j^{th} sensor to the i^{th} sensor.
- This task is carried out considering $0 < t < t_{max}$. From Eq. (1), it can be noted that the proposed scheme considers
  - $\varphi_1(t)$ as a function for mean remnant energy,
  - $\varphi_1(t)$ as a function for trust score that is considered towards the selection of an optimal path,
  - Multiple constraints as a maximal time of communication t_{max}, the quantity of forwarding data $FD_{ji}(t)$, the quantity of acquired data $AD_{ij}(t)$, and remnant energy $R_i(t)$.

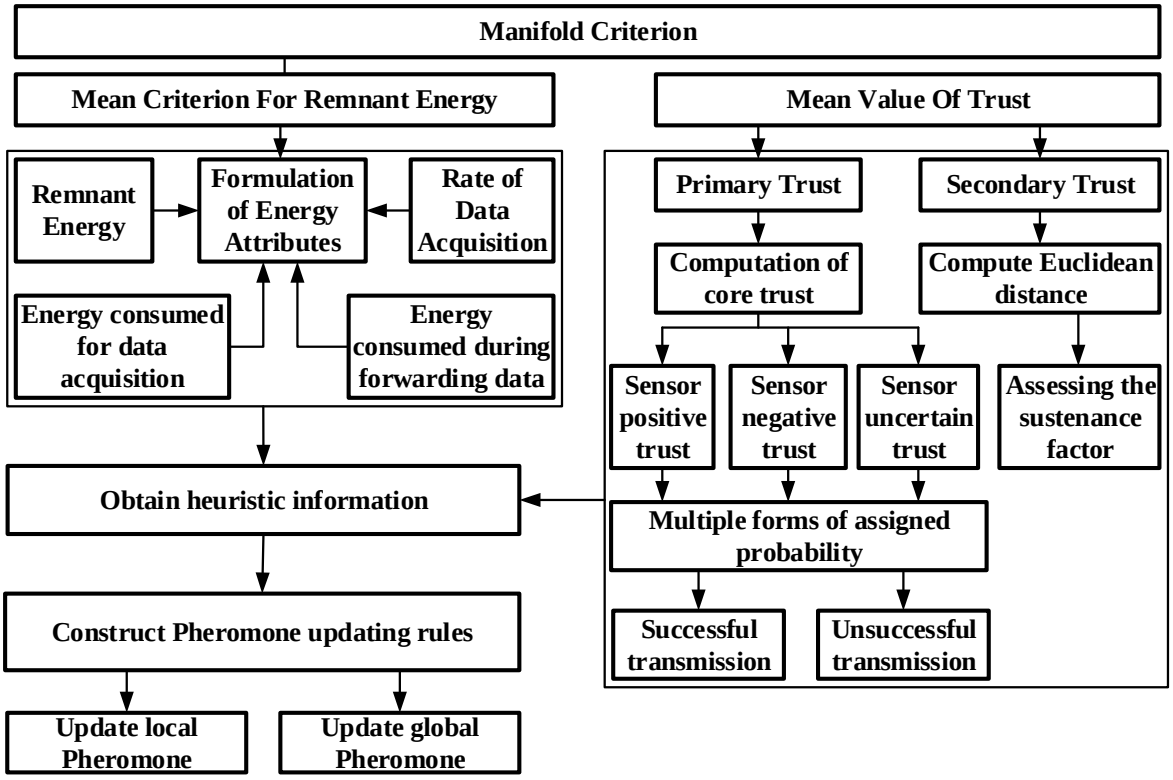


Fig. 1. Proposed methodology.

The complete formulation of the proposed study is based on two essential functions toward meeting the objectives, i.e., $\varphi_1(t)$ and $\varphi_2(t)$ towards energy and trust, respectively. It also includes four conditions of constraint. It should be noted that adopting these two essential functions is deployed for opting for the best path for data propagation, where the performance of nodes is selected as an indicator. This section further elaborates on the formulation of the proposed scheme. Further discussion of the formulation of the proposed modeling is carried out in subsequent sections.

A. Formulation of Manifold Criterion

Energy is one of the essential attributes considered in the proposed scheme whose consumption is generally recorded for multiple events of a sensor being in an idle state, sleep state, or either in receiving or transmission state. The formulation of energy attribute $R_i(t)$ is carried out as follows:

$$R_i(t) = \eta_1 + \eta_2 \tag{2}$$

To obtain information on the consumed energy of IoT nodes, any conventional mechanism can be adopted considering software and hardware components. The proposed scheme considers this an energy input value and contributes towards further optimized routing. Besides energy, the proposed method chooses to initiate its design implementation considering any conventional trust evaluation scheme that can vary based on application and use case. However, the system relies not much on device-specific trust computation mechanisms but more on behavioral analysis based on historical reports of security breaches, violations /adherence to security protocols, and compliance with access control protocols. Like the consumed energy metric, the proposed scheme also considers this trust score to act as an input to its model toward working in the direction of proof-of-concept. In Eq. (2), the computation of energy attribute $R_i(t)$ is carried out using two sub-entities, i.e., η_1 and η_2.

$$\eta_1 = R_j(t-1) - \lambda_{ij}(t).R_j^{\lambda}(t) \tag{3}$$

As shown in Eq. (3), the formulation of η_1 is carried out by differentiating the product of $\lambda_{ij}(t)$ and $R_j^{\lambda}(t)$ from $R_j(t-1)$, where the entities $R_j(t)$ and $R_j(t-1)$ represent remnant energy of the j^{th} sensor at instant t and $(t-1)$, respectively. The entity $\lambda_{ij}(t)$ represents the data acquisition rate from the i^{th} sensor to the j^{th} sensor. In contrast, $R_j^{\lambda}(t)$ represents the energy consumed for data acquisition for the j^{th} node at the t-instance.

$$\eta_2 = \lambda_{ij}(t).R_j^{C}(t) - R_j^{C}(t) \tag{4}$$

On the other hand, the formulation of η_2 shown in (4) is carried out by differentiating energy dissipated by the j^{th} sensor, $R_j^{C}(t)$ from the product of the data acquisition rate from the i^{th} sensor to the j^{th} sensor, $\lambda_{ij}(t)$ and energy consumed while forwarding data for the j^{th} sensor, $R_j^{C}(t)$. Therefore, the criterion function for mean energy associated with the path of routing considering α number of sensors in IoT can be represented as:

$$\varphi_1(t) = \sum {R_j(t)}/{\alpha} \tag{5}$$

In Eq. (5), the suffix j resided between (1, α). After formulating the energy attribute, the next task is developing the trust attribute of a sensor deployed in an IoT environment. The notion of trust attribute represents the degree of consistency in data packet transmission and receiving by the sensors when exposed to different severity levels of attacks in IoT (e.g., DDoS). The trust attribute is computed based on the evidential traces furnished by the neighboring sensors or by observing trust attributes from the adjacent sensors. Therefore, the proposed scheme formulated two types of trust attributes, i.e., primary and secondary.

1) Evaluation of primary trust: The primary trust is the value obtained directly from the i^{th} sensor to the j^{th} sensor. In contrast, the secondary trust is the value obtained from the i^{th} node to different relay sensors, and then it reaches the j^{th} sensor. The initial assessment towards the trust modeling is carried out for primary trust that considers that every possibility of evaluation of the sensor's trust consists of information associated with the degree of consistency associated with data transmission, rate of transmission, and rate of acquiring the data packet. The proposed scheme considers the evaluation of the core trust score γ_{ij} of the target i^{th} sensor on the j^{th} sensor to be assessed at t-instance of time as follows:

$$\gamma_{ij}(t) = \left[spt_{ij}(t), snt_{ij}(t), sut_{ij}(t)\right] \tag{6}$$

In Eq. (6), the computation of core trust γ_{ij} is carried out based on sensor positive trust, sensor negative trust, and sensor uncertain trust represented by $spt_{ij}(t)$, $snt_{ij}(t)$, and $sut_{ij}(t)$, respectively. It can be noted that all these three types of trust variables are equivalent to probability factors associated with multiple forms of assigned probability $\omega_{ij}^{c}(\pi)$, $\omega_{ij}^{c}(-\pi)$, and $\omega_{ij}^{c}(\pi,-\pi)$, where π represents a possible secured route for data propagation. It will eventually mean that the variable $\gamma_{ij}(t)$ represents the summation of the success rate of data acquiring, i.e., $\lambda_{ij}^{s}(t)$. In contrast, the success rate of data forwarding $F_{ij}^{s}(t)$ and $\beta_{ij}^{s}(t)$ denotes consistency in data packet transmission. Therefore, the expression in Eq. (6) can now be rewritten as:

$$\gamma_{ij}(t) = [\omega_{ij}^{c}(\pi), \omega_{ij}^{c}(-\pi), \omega_{ij}^{c}(\pi,-\pi)] \tag{7}$$

where, $\omega_{ij}^{c}(\pi) = [\theta_1.\lambda_{ij}^{s}(t) + \theta_2.F_{ij}^{s}(t) + \theta_3.\beta_{ij}^{s}(t)]$

$\omega_{ij}^{c}(-\pi) = [\theta_1.\lambda_{ij}^{us}(t) + \theta_2.F_{ij}^{us}(t) + \theta_3.\beta_{ij}^{us}(t)]$

$\omega_{ij}^{c}(\pi,-\pi) = [1 - spt_{ij}(t) - snt_{ij}(t)]$

In Eq. (7) in its expanded form, the power variable of the expression, i.e., s and us, represents successful and unsuccessful transmission in the IoT environment, while the variables θ_1, θ_2, and θ_3 represent weight values associated with different modes of transmission, which are subjected to training using first-order iterative optimization to arrive at the local value of the defined function. With the aid of Eq. (7), the formulation of the primary trust PT_{ij} between the i^{th} sensor and j^{th} sensor can be carried out as follows:

$$PT_{ij}(t) = [P1_{ij}(t), P2_{ij}(t), P3_{ij}(t)] \tag{8}$$

$$= [\omega_{ij}(\pi), \omega_{ij}(-\pi), \omega_{ij}(\pi, -\pi)]$$
$$= h_1 + h_2$$

In Eq. (8), the variables $P1$, $P2$, and $P3$ are primary values of trust, which are nearly similar to the notion of the variables $spt_{ij}(t)$, $snt_{ij}(t)$, and $sut_{ij}(t)$, while the variables h_1 and h_2 are empirically represented as follows:

$$h_1=\left[(1-\eta).\gamma_{ij}(t)\right]$$
$$h_2=\left[\eta.\gamma_{ij}(t)(t-1)\right] \quad (9)$$

From Eq. (9), the variable η further represents the temporal attribute of adaptivity used to evaluate the significance of heuristic data over the existing routing data to assess the trust among the sensors while deployed in an IoT environment.

2) Evaluation of secondary trust: The next part of the implementation is associated with formulating secondary trust $ST_{ij}(t)$, which is essentially meant to process any form of conflict. The empirical formulation of the secondary trust $ST_{ij}(t)$ is expressed as:

$$ST_{ij}(t) = [S1^l_{ij}(t), S2^l_{ij}(t), S3^l_{ij}(t) \quad (10)$$
$$= [\omega^l_{ij}(\pi), \omega^l_{ij}(-\pi), \omega^l_{ij}(\pi, -\pi)]$$

From Eq. (10), it can be seen that the formulation of secondary trust bears a similar strategy as noted in Eq. (8) for primary trust computation, including a variable l representing the common sensor node that resides within the transmission region for both the i^{th} and j^{th} sensors. A closer look into Eq. (10) for secondary trust evaluation will show that it is an n-ary circled times operation between two primary trusts, i.e., $ST_{ij}(t) = PT_{i,1}(t) \otimes PT_{1j}(t)$. The variables $S1^l_{ij}(t)$, $S2^l_{ij}(t)$, and $S3^l_{ij}(t)$ represent secondary positive trust, secondary negative trust, and secondary uncertain value of trust, respectively. These are also corresponding to $\omega^l_{ij}(\pi)$, $\omega^l_{ij}(-\pi)$, and $\omega^l_{ij}(\pi, -\pi)$.

Fig. 2 highlights the pictorial representation of primary and secondary trust evaluation. As the secondary trust evaluation is carried out by the other relay nodes, which are neither source nor destination nodes, there is always a possibility of evolving contradiction in the trust computation. Therefore, the next part of consecutive implementation is associated with evaluating the conflicts in the secondary trust. For this purpose, the proposed scheme constructs a reference matrix towards assessing the equivalency of the trust attributes of secondary trust where the variable μ represents the steadiness score between two secondary trusts and hence $\mu_{a,b}$ indicates the steadiness score for a^{th} and b^{th} relay sensors with a Euclidean distance of $dis_{a,b}$ between two secondary values of trust of ST^{lb}_{ij} and ST^{la}_{ij}, where la and lb represent common a and b sensors between i^{th} transmitting sensor and j^{th} receiving sensor. Empirically, it will be designated as follows:

$$dist_{ab} = |sqrt\left[\left(ST^{lb}_{ij} - ST_{ij}\right)^2 + \left(ST^{la}_{ij} + ST_{ij}\right)^2\right]| \quad (11)$$

After evaluating distance in Eq. (11), the next task is to assess the sustenance factor of secondary trust σ_a, which is obtained by summing up all steadiness scores, i.e., $\mu_{a,b}$. Further, towards attaining secondary trust value, there is a need for one more dependable attribute, i.e., indicative weight g_a, which is obtained from the standard weight of secondary trust (i.e., $ST^{l)}_{ij}$, i.e., O_a. This variable of normal weight of ST^l_{ij} is obtained by dividing sustenance score σ_a with the cumulative sustenance score of secondary trust, i.e., $\sum \sigma_a$. Therefore, the suggested indicative weight of secondary trust, i.e., g_a, is obtained by dividing the normal weight O_a with $arg_{max}(O_a)$. Thus, the final empirical expression towards attaining the value of secondary trust is as follows:

$$ST^l_{ij}(t) = [b_1, b_2, b_3] \quad (12)$$

In Eq. (12), the computation of secondary trust $ST^l_{ij}(t)$ is carried out using three dependable attributes, i.e., b_1, b_2, and b_3 representing $[g_a.S1^l_{ij}(t)]$, $[g_a.S2^l_{ij}(t)]$, and $[1 - g_a.(S1^l_{ij}(t) + S2^l_{ij}(t))]$ respectively.

$$ST^l_{ij}(t) = [(g_a.S1^l_{ij}(t)), (g_a.S2^l_{ij}(t)), (1 - g_a.(S1^l_{ij}(t) + S2^l_{ij}(t)))] \quad (13)$$

Finally, the primary and secondary trust (shown in Eq. (13)) values are combined to yield the joint trust score. It should be noted that the computation of primary trust is carried out directly between the two communicating nodes. In contrast, the secondary trust is evaluated using associated neighboring nodes, as shown in Fig. 2.

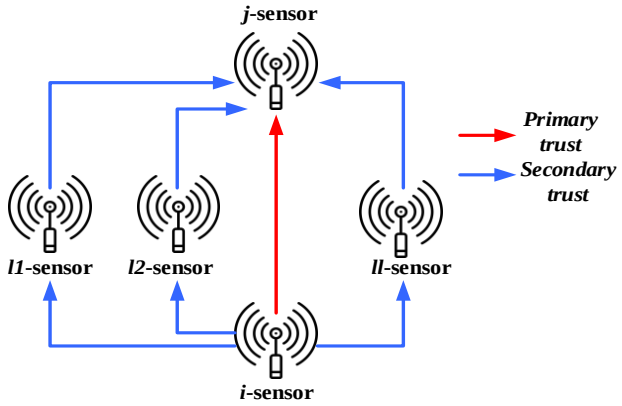


Fig. 2. Primary and secondary trust evaluation.

B. Formulation of Proposed ACO

After the trust computation, the next part of implementing the proposed scheme consists of applying the ACO scheme to arrive at an optimal solution for enhancing network lifetime and accomplishing a higher degree of resiliency from varied threats in the IoT environment. It should be noted that the proposed scheme introduces dual criteria and optimal solutions to amend the conventional ACO scheme to be used in secured routing in IoT. The novelty of the proposed ACO approach is that it constructs logic of manifold heuristic data and manifold pheromones to accomplish the discussed concept of manifold criterion in IoT routing by sensor nodes. The proposed scheme makes use of the idea of crowding distance to boost the algorithm to yield optimal solution diversity. The idea of the proposed ACO approach using crowding distance is to perform

an optimal selection of solutions required for the next generation to map with secure and energy-efficient routing among the sensors in IoT. If the crowding distance is more for a specific set of solutions, then the algorithm considers that set of solutions further for the next generation. Therefore, the proposed ACO approach offers a diversified and optimal solution to accomplish both secured trustworthy routes for data transmission with higher residual energy. According to the proposed scheme, the first set of operations is towards selecting the j^{th} sensor by k number of ants in the form of the i^{th} sensor present in the next hop. This computation offers the probability of change from one state to another. The proposed scheme constructs dual heuristic information that is associated with remnant energy and trust value, which can be empirically exhibited as follows:

$$\sigma_{ij}^{1}(t) = R_j(t) \text{ and } \sigma_{ij}^{2}(t) = \pi_j(t) \quad (14)$$

From Eq. (14), it should be noted that the first heuristic information $\sigma_{ij}^{1}(t)$, i.e., $R_j(t)$, is evaluated considering energy dissipated by the data acquired divided by the initialized energy of the sensor. In contrast, the second heuristic information $\sigma_{ij}^{2}(t)$ depends on a set of candidate routes for optimal data transmission performance with higher trust. The proposed scheme implements two types of pheromone information to fit the model with two heuristic information. The idea is to jointly study the value of trust and the associated remnant energy of a sensor. When an ant constructs a candidate solution using Eq. (15), the scheme instantly updates the pheromone associated with each heuristic. The following are the empirical expressions:

$$\left[\rho_{ij}^{1}(t), \rho_{ij}^{2}(t)\right] = [\chi_1, \chi_2] \quad (15)$$

In Eq. (15), the first attribute of pheromone, i.e., $\rho_{ij}^{1}(t)$, is associated with the criterion for remnant energy, i.e., χ_1. In contrast, the second attribute of pheromone, i.e., $\rho_{ij}^{2}(t)$, is related to the criterion for the mean value of trust, i.e., χ_2. It should be noted that the maximization of the pheromone score is relative to the mean quantity of resource availability and trust score associated with the propagation routes. The inference is that increased resource availability on the communication routes can only ensure a higher degree of pheromone retention. The conclusive remark of this logic is that nodes with reported higher value of trust and resources will eventually have higher feasibility to be opted as participating routing nodes. However, this empirical expression is in abstract form, and it demands more clarity in terms of the actual updating mechanism of pheromone in ACO, which is given as follows:

$$\rho_{ij}^{n}(t+1) = funct(e, \rho_{ij}^{n}) \quad (16)$$

In Eq. (16), the updating of local pheromone ρ_{ij}^{n} is stated considering the unique function $funct$ of coefficient of pheromone evaporation e mainly. The function $funct$ performs extractions of two operators using arguments of e and ρ_{ij}^{n}, i.e., i) $(1-e).\rho_{ij}^{n}(t)$ and ii) $e.\Delta\rho_{ij}^{n}(t)$, where the values of suffixes i and j are associated with π, while the value of power variable n is (1, 2) owing to consideration of two heuristic information. The variable $\Delta\rho_{ij}^{n}(t)$ is incremented, further proportional to the heuristic information of mean trust and energy score. The prominent inference of this implementation concept is that the maximum score of the function χ_1 associated with mean remnant energy will mean higher retention of pheromone on the routing path. It will also mean that if the value of the function χ_2 associated with the mean trust value is found to be maximal, then it will mean that a large quantity of pheromone will be retained. Hence, the routing operation in IoT will always choose only those sensors with a higher value of remnant energy and higher trust. The implementation concept of updating the local pheromone is completed when the data reaches the sink node, followed by further updating of enhanced crowding distance by sub-optimal solutions in the sink node. The system finally yields a backward-moving ant from the forward-moving ant upon completion of the local updating operation of the pheromone.

Finally, the proposed scheme performs updating of the global pheromone by subjecting all the sub-optimal solutions presented by the sink node. The backward-moving ant in the proposed ACO approach carries out this task of updating the sub-optimal solution. An empirical expression for this global pheromone updating is similar to that of Eq. (12) only with the difference of $\Delta\rho_{ij}^{n}(t)$ equivalent to $1/\varsigma$, where the variable ς represents the quantity of the sensors traversed by the k number of backward-moving ants considering a set of sub-optimal solutions as routing paths to update the global pheromone.

The contributions of this methodology can be seen by two significant results: (a) The primary contribution of the proposed method is to design and develop a simplified yet robust secure data propagation scheme in IoT. It is simplified as it doesn't consist of any sophisticated mechanism or involve a higher number of complex processing routing schemes. It is robust as the method can realize the dynamic vulnerabilities present in links connecting IoT nodes without any dependencies on the apriori information of an attacker. The scheme is highly secured as the formulation of the system is carried out considering manifold criteria in the form of an essential function associated with trust and resources of IoT nodes as well as various practical constraints related to time, forwarding, and receiving data, and remnant resources of IoT nodes, (b) The second prime contribution is associated with the mechanism of deploying an ACO approach where novel pheromone management is presented. The credibility of the data propagation and exchange among the IoT nodes is carried out by proposed ACO-based routing, where the selection of cost-effective and secured routes is based on heuristic and pheromone information of manifold type. It should be noted that the proposed scheme considers heuristic information and pheromone information derived from remanent information and the trust score of IoT nodes. The main contributions of the study can be briefly summarized as:

1) The proposed scheme balances trust computation with the selection of optimal nodes possessing substantial residual energy, ensuring they have the requisite resources for extensive secure data propagation within IoT.

2) The trust evaluation mechanism of the proposed system is scalable and viable for both compact and expansive IoT environments. Its resilience is evident as it can function even

in the presence of unknown-origin attackers. This resilience stems from the fact that, regardless of an attacker's strategy, the proposed scheme ensures all standard sensors compute an optimal solution that remains inscrutable to potential intruders.

3) The proposed ACO approach addresses and rectifies the traditional issues of slow convergence and parameter sensitivity, which plague conventional ACOs. By providing a broader problem scope, it aptly aligns with the expansive nature of IoT.

The ensuing section delves into the results derived from implementing the proposed scheme.

V. RESULT

This section presents the results achieved after implementing the proposed model. Since the proposed implementation introduces a novel ACO-based secure routing method, emphasis has been placed on investigating data transmission performance. Additionally, the trust management scheme has been executed for the system. The primary objective of the result analysis is to establish an extensive test environment using variable performance metrics. This is done to gauge the impact of the proposed secure routing conducted by sensors within the IoT environment. The results are then analyzed to provide insights into how the model's performance compares to existing secure routing schemes.

A. Assessment Strategy

The entire implementation is conducted in MATLAB on a standard 64-bit Windows machine. The simulation environment selected for the experiment spans an area of 1000x1000 m^2, with the specific simulation parameters detailed in Table II. This environment replicates a smart city setting where numerous clusters of wireless sensor nodes are interconnected, facilitating data aggregation with a predetermined energy level. While the initial energy assigned to each node is 10 J, it can be adjusted based on the specific IoT application in use. Thus, the proposed simulation environment offers considerable flexibility, accommodating modifications to parameter values to fit various scenarios.

TABLE II. SIMULATION PARAMETERS ADOPTED FOR ASSESSMENT

Parameters	Values
No. of Sensors	500-1000
Initialized energy	10 J
Rate of data transmission	400 kbps
Data packet size	5000 byes
Communication Radius	200 m
Antenna	Omni-directional
MAC	802.11
Simulation Time	100 s

To gauge the effectiveness of the proposed scheme, a benchmarked analysis against existing secure routing schemes in the IoT environment is essential. This comparative analysis examines specific performance metrics across the proposed and existing secure routing schemes. The conventional secured routing schemes selected for comparative analysis in IoT include:

- Routing Protocol for Low-Power and Lossy Network (RPL): This standard IoT routing scheme is tailored for networks with lossy features and low-powered nodes. Its secure variant, Secure RPL (SRPL), employs authentication of messages and encryption operations to ensure data freshness, integrity, and confidentiality [32].
- Routing using 6LoWPAN: Another conventional IoT routing method closely aligned with RPL, this protocol leverages the IPv6 scheme for routing. It adopts 6LoSec, primarily designed to guard against replay attacks and ensure data integrity and confidentiality [50].
- Secured routing using Zigbee Cluster Library (ZCL): Zigbee's prevalence in IoT-based wireless communication systems is notable. However, ZCL is designed to provide secure routing exclusively for Zigbee-based networks. This scheme uses authentication and encryption to ensure authentic communication, data integrity, and confidentiality [51].
- Secure routing using Constrained Application Protocol (CoAP): This protocol facilitates data transmission for resource-constrained devices within the IoT framework. It incorporates Datagram Transport Layer Security (DTLS) to provide authentication, data integrity, and confidentiality [52].

The aforementioned secure routing schemes and proposed method have been implemented in comparable test environments and under similar simulation parameters. Moreover, the conventional ACO and PSO algorithms have been employed to evaluate the performance enhancements ushered in by the proposed scheme relative to existing bioinspired approaches.

B. Discussion of Result

The initial performance metric examined is network throughput, calculated as the volume of data packets transmitted from one node to another within the IoT environment over a specific time. A detailed examination of the proposed secure routing scheme reveals that it encompasses various mathematical and logical operations based on the manifold criterion-based model. This design ensures a robust defense against manipulation or unauthorized access. Given these intricacies, potential delays or added computational demands might impact throughput. Consequently, benchmarking based on throughput provides a clear insight into whether the inherent security operations compromise network performance and data transmission. Additionally, thorough throughput assessment can highlight potential bottleneck areas within the routing path.

Fig. 3 presents the average throughput observed during a series of evaluations over specified simulation duration. It indicates that the proposed scheme outperforms existing methods in terms of throughput. The RPL protocol emerges as the next best performer in current systems, primarily due to its dynamic path selection capabilities. However, it needs help to balance traffic load during dynamic events within the IoT landscape. Following RPL, conventional ACO, PSO, and

CoAP algorithms rank next in terms of throughput. It's worth noting that traditional ACOs have a limitation, needing help to provide optimal solutions amidst heavy traffic flow. PSO faces a similar challenge, grappling with increased memory dependencies as traffic surges, reducing throughput. The CoAP protocol, because it utilizes User Datagram Protocol (UDP), is prone to packet loss, primarily when data packets are transmitted in an unordered sequence within IoT.

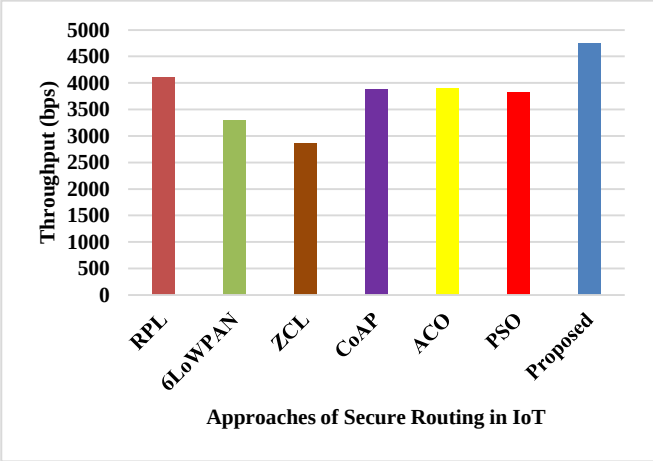


Fig. 3. Comparative analysis of throughput.

Moreover, 6LoWPAN and ZCL underperform in throughput within the IoT environment due to complexities arising from header compression. In contrast, the proposed system remains unaffected by such challenges. Its constraints concerning energy and trust, combined with its modeling, are adept at identifying vulnerabilities and pinpointing alternative optimal paths for propagation based on the manifold criterion. Consequently, even without encryption, the proposed scheme achieves superior secure data transmission throughput compared to conventional secure IoT routing strategies that predominantly depend on encryption and authentication.

The subsequent performance metric evaluated is a delay, calculated as the latency encountered during the data bit transmission across the network from one sensor to another within the IoT framework. Evaluating delay is paramount for secure data transmission, given the diverse applications and services housed within the IoT environment. It's essential to highlight that the proposed scheme incorporates several mathematical procedures for trust calculation, where local and global parameters play pivotal roles. Conversely, most extant secure routing schemes lean heavily on encryption and message authentication to ensure data security. Thus, it becomes imperative to ascertain that these intrinsic security processes don't detrimentally influence network performance by augmenting delay. For optimal network efficiency, it's crucial to maintain low delay, as extended latency often signals network bottlenecks or areas compromised by security vulnerabilities, consequently impinging on data transmission durations.

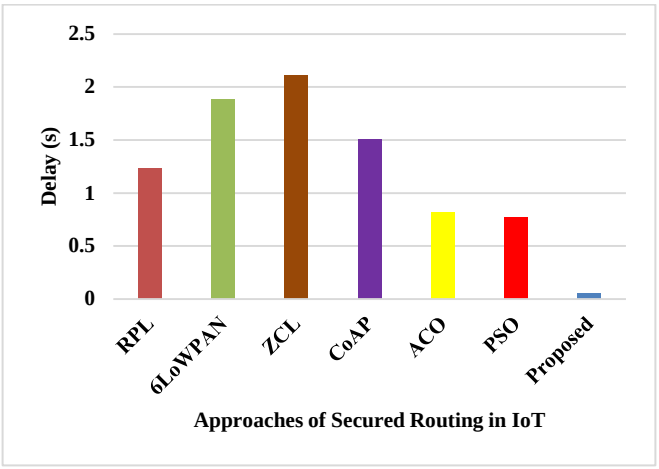


Fig. 4. Comparative analysis of delay.

Examining Fig. 4 reveals that the proposed scheme considerably reduces delay compared to existing systems. ZCL displays a higher delay due to its conventional architecture's slower transmission rate tailored to meet Zigbee transmission requirements. 6LoWPAN outperforms ZCL in delay, but it sacrifices some security; it possesses a weaker immunity to interference. A significant cause of delay in 6LoWPAN is its repeated retransmissions due to packet loss from IPv6 data chunks. CoAP, being optimized for peer-to-peer communication, has an inevitable delay. However, RPL shines the best in reducing delay compared to other secure transmission methods. This is attributed to its auto-configuration capabilities and dynamic path selection, vital for large-scale IoT devices with resource-constrained sensors. Yet, RPL's delay becomes significant in IoT's mobile environments. Conventional ACO and PSO also underperform due to i) increasing iteration counts as they seek optimal solutions in constrained problem areas and ii) premature convergence resulting in sub-optimal routing paths, leading to increased delay. In contrast, our proposed scheme excels in delay performance, being more progressive, less iterative, and encompassing a broader problem area, thereby reducing the effort needed for optimal solutions and achieving a superior delay score.

The third performance metric is energy consumption. The rationale behind selecting this metric hinges on two factors: i) our mathematical model prioritizes residual energy as a primary constraint, with manifold criterion modeling also considering residual energy apart from the trust attribute as a heuristic. This necessitates evaluating the model's impact on energy consumption, and ii) our scheme primarily focuses on sensors as essential IoT devices. Given their limited energy resources, it becomes crucial to ascertain the energy expended during secure data transmission in IoT. The objective is to balance trust-based security and energy efficiency, ensuring a prolonged network lifespan. Here, "energy" denotes the power expended on cumulative sensor operations, which encompasses data transmission and reception, data processing, and internal circuit functions. We rely on the first-order radio energy model [53] to assess this, which provides a comprehensive formula for sensors' total energy consumption.

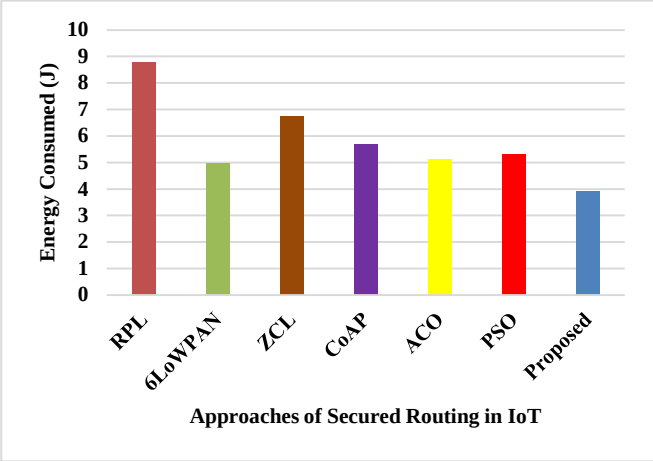


Fig. 5. Comparative analysis of energy consumption.

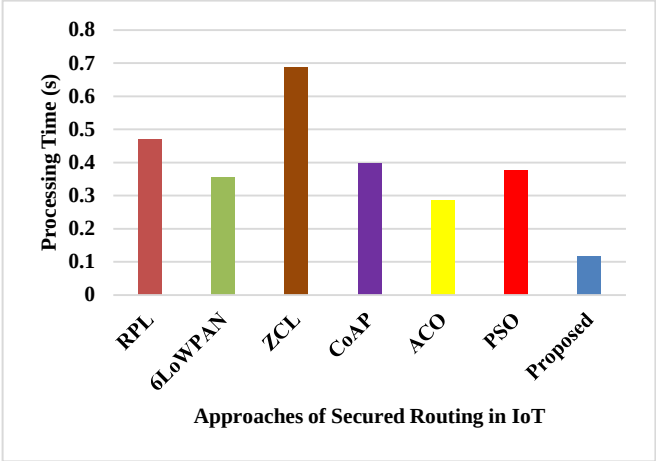


Fig. 6. Comparative analysis of algorithm processing time.

The data presented in Fig. 5 reveals that our proposed scheme significantly reduces power consumption compared to most existing systems. What stands out is that both conventional ACO and PSO demonstrate slightly higher energy consumption. The main reasons for this are: i) the increased use of attributes to achieve higher convergence performance and ii) a stronger focus on the local search optimization problem, often at the expense of global search space. The distinction between the proposed ACO and the conventional ACO lies in the former's formulation of pheromone management based on multiple criteria intertwined with an adaptive operational principle. A detailed examination of our ACO scheme further shows that their functionalities are mostly preserved despite the utilization of numerous parameters. The only variation arises from their use cases to enhance trust and energy retention. Due to the calculated global update formulation, local operatives' reliance on extensive operations diminishes as simulation time increases, leading to energy conservation. IoT secure routing schemes, such as RPL, ZCL, CoAP, and 6LoWPAN, incorporate encryption in their security variants. This inclusion demands a significant energy allocation for the ciphering and deciphering processes, which our proposed secure routing scheme avoids.

The final performance metric assessed for the proposed scheme is the processing time, an essential measure that reflects the time complexity inherent in algorithmic processing during secure routing. This metric is determined by evaluating a sensor's time to complete its operations. An efficient system model, especially one with lightweight characteristics, should display reduced processing times. If the processing time is extensive, the security scheme in use might benefit from some refinement to boost its overall efficiency. It is also crucial to note that, given sensors' limited computational capabilities, their processing time can increase when more complex operations are introduced. As a result, gauging the system's efficacy in terms of computational complexity by assessing its processing time becomes vital.

Fig. 6 illustrates that the proposed scheme significantly reduces processing time compared to other secured routing schemes. The processing time for ZCL is notably higher, as this scheme necessitates performing extensive iterative operations for massive data transmission in IoT. On closer examination, 6LoWPAN, CoAP, and conventional PSO performance reveal similar processing times. This similarity arises because these approaches segment and chunk data into smaller portions, leading to packet loss in a heterogeneous IoT network that necessitates retransmission. Additionally, the conventional PSO requires iterative computation of particles and velocities to identify optimal results. While this is effective for homogeneous systems, it is less so for heterogeneous ones, causing them to amalgamate all data packets and conduct routing. Such operations demand significant processing time and deliver sub-optimal data quality.

Furthermore, RPL displays a longer processing time, slightly more than CoAP. This is attributed to RPL's formation of a directed acyclic graph, eventually resulting in a singular link from the leaf node to the route. While this might be suitable for smaller IoT networks, the RPL graph operation must be repeated to achieve data transmission within the context of more extensive IoT networks. This repetition expends excessive resources and consumes considerable processing time for data transmission. Therefore, the proposed scheme promises minimal processing time, primarily due to diminished resource dependencies and fewer iterative operations, as emphasized in the proposed mathematical modeling.

VI. Conclusion

The landscape of the IoT is both exciting and challenging, marked by tremendous opportunities and, in parallel, considerable security vulnerabilities. In the modern age, when digital interconnectedness is both a boon and a bane, the urgency to fortify IoT against burgeoning threats cannot be overstated. The paper has addressed this urgency, providing a novel and streamlined approach that leverages a modified version of ACO toward achieving optimal security in the vast and dynamic IoT ecosystem. This bioinspired approach symbolizes our attempt to mimic nature's intuitive problem-solving methodologies. Through our process, not only is data

transmission secured, but the dual challenges of optimizing sensor energy and ensuring high trust scores are simultaneously addressed. A particular innovation in our work is the unique pheromone management system, which holistically considers residual energy and trust scores. Coupled with our manifold criterion and the dual-tiered trust evaluation system, the methodology provides an unparalleled framework for IoT security. Our research has showcased its merits, delivering impressive performance metrics compared to prevailing secure routing protocols. With a 35% improvement in throughput, 89% reduction in delay, 54% decrease in energy consumption, and 73% surge in processing speed, the approach is theoretically sound and practically efficacious. The current work, however, has its limitations. While it has achieved a balance between energy conservation and security, nuances to this balance need exploration. The inherently dynamic nature of IoT means that newer devices with diverse capacities are continually entering the ecosystem, presenting evolving challenges for security protocols. The future work of this study model will be to formulate a hybrid modeling of a bioinspired approach to optimize the security and resource management performance in a large IoT environment. Furthermore, integrating Artificial Intelligence and Machine Learning algorithms could further enhance our initial results. These can help the system to adaptively learn from emerging threats and respond proactively, ensuring a more agile and dynamic security mechanism.

ACKNOWLEDGMENT

The authors express their appreciation for the effort of Ms. Gousia Nissar in proofreading and editing the paper.

REFERENCES

[1] R. F. Olanrewaju *et al.*, "The internet of things vision: A comprehensive review of architecture, enabling technologies, adoption challenges, research open issues and contemporary applications," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 26, no. 1, pp. 51–77, 2022. doi:10.37934/araset.26.1.5177

[2] R. K. Gangawar, S. Kumari, A. K. Pathak, S. D. Gutlapalli, and M. C. Meena, "Optical fiber based temperature sensors: A Review," *Optics*, vol. 4, no. 1, pp. 171–197, 2023. doi:10.20944/preprints202302.0180.v1

[3] S. K. Ghosh *et al.*, "Temperaturepressure hybrid sensing all-organic stretchable energy harvester," *ACS Applied Electronic Materials*, vol. 3, no. 1, pp. 248–259, 2020. doi:10.1021/acsaelm.0c00816.s002

[4] N. V. Krishna Prasad *et al.*, "Ceramic sensors: A mini-review of their applications," *Frontiers in Materials*, vol. 7, p. 593342, 2020. doi:10.3389/fmats.2020.593342

[5] A. M. Rahmani, S. Bayramov, and B. Kiani Kalejahi, "Internet of things applications: Opportunities and threats," *Wireless Personal Communications*, vol. 122, no. 1, pp. 451–476, 2021. doi:10.1007/s11277-021-08907-0

[6] S. A. Siddiqui, A. Ahmad, and N. Fatima, "IoT-based disease prediction using machine learning," *Computers and Electrical Engineering*, vol. 108, p. 108675, 2023. doi:10.1016/j.compeleceng.2023.108675

[7] H. S. Kim, Y. J. Park, and S. J. Kang, "Secured and deterministic closed-loop IoT system architecture for sensor and Actuator Networks," *Sensors*, vol. 22, no. 10, p. 3843, 2022. doi:10.3390/s22103843

[8] J. Yun, I. Y. Ahn, J. Song, and J. Kim, "Implementation of sensing and actuation capabilities for IoT devices using onem2M platforms," *Sensors*, vol. 19, no. 20, p. 4567, 2019. doi:10.3390/s19204567

[9] G. M. Kapitsaki, A. P. Achilleos, P. Aziz, and A. C. Paphitou, "SensoMan: Social Management of context sensors and actuators for IoT," *Journal of Sensor and Actuator Networks*, vol. 10, no. 4, p. 68, 2021. doi:10.3390/jsan10040068

[10] C. Stolojescu-Crisan, C. Crisan, and B.-P. Butunoi, "An IoT-based Smart Home Automation System," *Sensors*, vol. 21, no. 11, p. 3784, 2021. doi:10.3390/s21113784

[11] F. Pereira, R. Correia, P. Pinho, S. I. Lopes, and N. B. Carvalho, "Challenges in resource-constrained IoT devices: Energy and communication as critical success factors for future IoT deployment," *Sensors*, vol. 20, no. 22, p. 6420, 2020. doi:10.3390/s20226420

[12] B. U. I. Khan, R. F. Olanrewaju, F. Anwar, R. N. Mir, A. Oussama, and A. Z. B. Jusoh, "Internet of Things—The Concept, Inherent Security Challenges and Recommended Solutions," in *Smart Network Inspired Paradigm and Approaches in IoT Applications*, Springer, Singapore, 2019, pp. 63-86. doi: 10.1007/978-981-13-8614-5_5

[13] B. U. I. Khan, R. F. Olanrewaju, F. Anwar, R. N. Mir, and A. R. Najeeb, "A critical insight into the effectiveness of research methods evolved to secure IoT ecosystem," *International Journal of Information and Computer Security*, vol. 11, no. 4/5, pp. 332–354, 2019. doi:10.1504/ijics.2019.101908

[14] P. Ferrara, A. K. Mandal, A. Cortesi, and F. Spoto, "Static analysis for discovering IoT vulnerabilities," *International Journal on Software Tools for Technology Transfer*, vol. 23, no. 1, pp. 71–88, 2020. doi:10.1007/s10009-020-00592-x

[15] A. H. Celdrán *et al.*, "Intelligent and behavioral-based detection of malware in IoT spectrum sensors," *International Journal of Information Security*, vol. 22, no. 3, pp. 541–561, 2022. doi:10.1007/s10207-022-00602-w

[16] A. Attkan and V. Ranga, "Cyber-physical security for IoT Networks: A comprehensive review on traditional, blockchain and artificial intelligence based key-security," *Complex & Intelligent Systems*, vol. 8, no. 4, pp. 3559–3591, 2022. doi:10.1007/s40747-022-00667-z

[17] U. Kumar, S. Navaneet, N. Kumar, and S. C. Pandey, "Isolation of DDoS attack in IoT: A new perspective," *Wireless Personal Communications*, vol. 114, no. 3, pp. 2493–2510, 2020. doi:10.1007/s11277-020-07486-w

[18] T. B. Josey and D. S. Misbha, Man-in-the-Middle attack mitigation in IoT sensors with hash-based multidimensional Lamport digital signature," in *Lecture Notes in Electrical Engineering*, Springer Nature Singapore, 2023, pp. 47–56

[19] B. U. I. Khan, F. Anwar, R. F. Olanrewaju, B. R. Pampori, and R. N. Mir, "A novel multi-agent and multilayered game formulation for intrusion detection in internet of things (IoT)," *IEEE Access*, vol. 8, pp. 98481–98490, 2020. doi:10.1109/access.2020.2997711

[20] R. Sharma and R. Arya, "Secure transmission technique for data in IoT Edge Computing Infrastructure," *Complex & Intelligent Systems*, vol. 8, no. 5, pp. 3817–3832, 2021. doi:10.1007/s40747-021-00576-7

[21] U. Panahi and C. Bayılmış, "Enabling secure data transmission for wireless sensor networks based IoT Applications," *Ain Shams Engineering Journal*, vol. 14, no. 2, p. 101866, 2023. doi:10.1016/j.asej.2022.101866

[22] E. Refaee *et al.*, "Secure and scalable healthcare data transmission in IoT based on optimized routing protocols for mobile computing applications," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–12, 2022. doi:10.1155/2022/5665408

[23] B. U. I. Khan *et al.*, "SGM: Strategic game model for resisting node misbehaviour in IoT-Cloud Ecosystem," *Information*, vol. 13, no. 11, p. 544, 2022. doi:10.3390/info13110544

[24] V. K. Quy, V. H. Nam, D. M. Linh, and L. A. Ngoc, "Routing algorithms for Manet-IoT Networks: A comprehensive survey," *Wireless Personal Communications*, vol. 125, no. 4, pp. 3501–3525, 2022. doi:10.1007/s11277-022-09722-x

[25] M. Majid *et al.*, "Applications of wireless sensor networks and internet of things frameworks in the industry revolution 4.0: A systematic literature review," *Sensors*, vol. 22, no. 6, p. 2087, 2022. doi:10.3390/s22062087

[26] B. Aslan, F. Yavuzer Aslan, and M. T. Sakallı, "Energy consumption analysis of lightweight cryptographic algorithms that can be used in the security of internet of things applications," *Security and Communication Networks*, vol. 2020, pp. 1–15, 2020. doi:10.1155/2020/8837671

[27] M. Aaqib, A. Ali, L. Chen, and O. Nibouche, "IoT trust and reputation: A survey and taxonomy," *Journal of Cloud Computing*, vol. 12, no. 1, pp. 1–20, 2023. doi:10.1186/s13677-023-00416-8

[28] Y. Liu, J. Wang, Z. Yan, Z. Wan, and R. Jäntti, "A survey on Blockchain based trust management for internet of things," *IEEE Internet of Things Journal*, vol. 10, no. 7, pp. 5898–5922, 2023. doi:10.1109/jiot.2023.3237893

[29] X. Liu, J. Yu, K. Yu, G. Wang, and X. Feng, "Trust secure data aggregation in WSN-based IIoT with Single Mobile Sink," *Ad Hoc Networks*, vol. 136, p. 102956, 2022. doi:10.1016/j.adhoc.2022.102956

[30] S. Mangalampalli, G. R. Karri, and U. Kose, "Multi Objective Trust aware task scheduling algorithm in cloud computing using whale optimization," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 791–809, 2023. doi:10.1016/j.jksuci.2023.01.016

[31] S. M. Muzammal, R. K. Murugesan, N. Jhanjhi, M. S. Hossain, and A. Yassine, "Trust and mobility-based protocol for secure routing in internet of things," *Sensors*, vol. 22, no. 16, p. 6215, 2022. doi:10.3390/s22166215

[32] S. M. Muzammal *et al.*, "A trust-based model for secure routing against RPL attacks in internet of things," *Sensors*, vol. 22, no. 18, p. 7052, 2022. doi:10.3390/s22187052

[33] S. Awan *et al.*, "Blockchain based Secure Routing and Trust Management in Wireless Sensor Networks," *Sensors*, vol. 22, no. 2, p. 411, 2022. doi:10.3390/s22020411

[34] R. Nagaraju *et al.*, "Secure routing-based energy optimization for IoT application with heterogeneous wireless sensor networks," *Energies*, vol. 15, no. 13, p. 4777, 2022. doi:10.3390/en15134777

[35] N. B. Bakhtiari, M. Rafighi, and R. Ahsan, "TTLA: Two-way trust between clients and fog servers using Bayesian Learning Automata," *The Journal of Supercomputing*, vol. 79, pp. 16152–16180, 2022. doi:10.21203/rs.3.rs-1744138/v1

[36] B. Rakesh and P. S. H, "Novel authentication and Secure Trust based RPL routing in Mobile Sink supported internet of things," *Cyber-Physical Systems*, vol. 9, no. 1, pp. 43–76, 2021. doi:10.1080/23335777.2021.1933194

[37] A. Gladkov *et al.*, "DT-RRNS: Routing protocol design for secure and reliable distributed smart sensors communication systems," *Sensors*, vol. 23, no. 7, p. 3738, 2023. doi:10.3390/s23073738

[38] S. Ramaswamy and J. Norman, "Social and QoS based trust model for secure clustering for Wireless Body Area Network," *The International Journal of Electrical Engineering & Education*, 2020. doi:10.1177/0020720920953133

[39] X. Wang, "Low-energy secure routing protocol for WSNS based on multiobjective ant colony optimization algorithm," *Journal of Sensors*, vol. 2021, pp. 1–9, 2021. doi:10.1155/2021/7633054

[40] K. Saleem and I. Ahmad, "Ant colony optimization ACO based Autonomous Secure Routing Protocol for Mobile Surveillance Systems," *Drones*, vol. 6, no. 11, p. 351, 2022. doi:10.3390/drones6110351

[41] T.-H. Nguyen and J. J. Jung, "ACO-based traffic routing method with automated negotiation for connected vehicles," *Complex & Intelligent Systems*, vol. 9, no. 1, pp. 625–636, 2022. doi:10.1007/s40747-022-00833-3

[42] G. Anjinappa and D. Bangalore Prabhakar, "A secure IoT and Edge Computing based EV selection model in V2G systems using ant colony optimization algorithm," *International Journal of Pervasive Computing and Communications*, 2022. doi:10.1108/ijpcc-06-2022-0245

[43] A. El-Ghamry, T. Gaber, K. K. Mohammed, and A. E. Hassanien, "Optimized and efficient image-based IoT malware detection method," *Electronics*, vol. 12, no. 3, p. 708, 2023. doi:10.3390/electronics12030708

[44] H. A. Alterazi *et al.*, "Prevention of cyber security with the internet of things using particle swarm optimization," *Sensors*, vol. 22, no. 16, p. 6117, 2022. doi:10.3390/s22166117

[45] H. C. Lin, P. Wang, and W. H. Lin, "Implementation of a PSO-based security defense mechanism for tracing the sources of DDoS attacks," *Computers*, vol. 8, no. 4, p. 88, 2019. doi:10.3390/computers8040088

[46] R. R. K and M. Ramakrishnan, "Internet of trust things using particle-swarm optimisation (PSO-IoT)," *SSRN Electronic Journal*, 2021. doi:10.2139/ssrn.3769174

[47] K. Janani and S. Ramamoorthy, "Threat analysis model to control IoT network routing attacks through deep learning approach," *Connection Science*, vol. 34, no. 1, pp. 2714–2754, 2022. doi:10.1080/09540091.2022.2149698

[48] M. Hosseinzadeh *et al.*, "A cluster-tree-based secure routing protocol using Dragonfly Algorithm (DA) in the internet of things (IoT) for Smart Agriculture," *Mathematics*, vol. 11, no. 1, p. 80, 2022. doi:10.3390/math11010080

[49] P. Selvaraj *et al.*, "An enhanced and Secure Trust-aware improved GSO for encrypted data sharing in the internet of things," *Applied Sciences*, vol. 13, no. 2, p. 831, 2023. doi:10.3390/app13020831

[50] M. Tanveer *et al.*, "S6AE: Securing 6lowpan using authenticated encryption scheme," *Sensors*, vol. 20, no. 9, p. 2707, 2020. doi:10.3390/s20092707

[51] K. Nichols, V. Jacobson, and R. King, "Defined-Trust Transport (DEFTT) protocol for limited domains," IETF Datatracker, https://datatracker.ietf.org/doc/draft-nichols-tsv-defined-trust-transport/ (accessed Jul. 30, 2023).

[52] J. Granjal, J. Silva, and N. Lourenço, "Intrusion detection and prevention in CoAP wireless sensor networks using anomaly detection," *Sensors*, vol. 18, no. 8, p. 2445, 2018. doi:10.3390/s18082445

[53] F. Liu, "Majority decision aggregation with binarized data in wireless sensor networks," *Symmetry*, vol. 13, no. 9, p. 1671, 2021. doi:10.3390/sym13091671.

Analyzing Sentiment in Terms of Online Feedback on Top of Users' Experiences

Mohammed Alonazi
Department of Information Systems, College of Computer Engineering and Sciences,
Prince Sattam bin Abdulaziz University, Al-Kharj, 16273, Saudi Arabia

Abstract*—Since most businesses today are conducted online, it is crucial that each customer provide feedback on the various items offered. Evaluating online product sentiment and making suggestions using state-of-the-art machine learning and deep learning algorithms requires a comprehensive pipeline. Thus, this paper addresses the need for a comprehensive pipeline to analyze online product sentiment and recommend products using advanced machine learning and deep learning algorithms. The methodology of the research is divided into two parts: the Sentiment Analysis Approach and the Product Recommendation Approach. The study applies several state-of-the-art algorithms, including Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, Bidirectional Long-Short-Term-Memory (BI-LSTM), Convolutional Neural Network (CNN), Long-Short-Term-Memory (LSTM), and Stacked LSTM, with proper hyperparameter optimization techniques. The study also uses the collaborative filtering approach with the k-Nearest Neighbours (KNN) model to recommend products. Among these models, Random Forest achieved the highest accuracy of 95%, while the LSTM model scored 79%. The proposed model is evaluated using Receiver Operating Characteristic (ROC) - Area under the ROC Curve (AUC). Additionally, the study conducted exploratory data analysis, including Bundle or Bought-Together analysis, point of interest-based analysis, and sentiment analysis on reviews (1996-2018). Overall, the study achieves its objectives and proposes an adaptable solution for real-life scenarios.

Keywords*—Sentiment analysis; product review; machine learning; recommendation system; collaborative filtering; exploratory data analysis

I. Introduction

In this era of modern computational technology, technological advancement can be seen everywhere; even the business sector is taking the initiative to enhance its revenue in computing and technology [1]. The term sentiment analysis has extensively been utilized to track out social media, allowing businesses to extract hidden information from the recorded data or identify critical information before coming into the limelight [2]. Thinking about giant tech companies such as Facebook, Google, Apple, and Microsoft, they have a huge amount of datasets. Every day, lots of data is being recorded to their central database. Besides, manually analyzing these data is time-consuming [3]. With a massive amount of dataset, it is quite difficult to manually extract meaningful insights that can help them make a business oriental decision. Turning into a product-based company, it can be stated that the product-based company tends to develop its product and launch it into the market [4]. If the thing is grocery or jewelry items so in this case, users will be giving their opinion based on the items whether the product is caught their attention or not. Suppose the clients explore a large e-commerce platform like Amazon.com. In this case, it is noticeable that the end-users threw their comments or reviews related to a specific item. Other individuals take themselves towards the advertisement phenomenon [5]. In turn, this brings us to a big question, whether it or not possible to handle such massive amounts of ratings manually and extract the business insights. So, the automated system can be the possible solutions to overcome these issues [6]. Thus, the impact of information on user sentiment and physical environments is not limited to modern technology.

The computational approach can be taken into consideration. Nowadays, machine learning algorithms have widely been utilized in biomedical imaging, forecasting things or even critical disease prognosis [7]. For the case of product analysis, it has shown their promising performance beforehand. Researchers are now using computing power to take their analysis to a satisfactory level from where meaningful insights can be extracted easily by analyzing a large number of datasets [8]. The priority of this research is to analyze the product sentiment using machine learning algorithms and propose a recommendation system for the stakeholder to make a better decision while doing online business. In this research, conventional Machine Learning (ML) algorithms were adopted to analyze the online product, and our study will significantly contribute to the research community. This is the motivation of this proposed study.

On the other hand, there are three contributions have been addressed in this study, and these are following:

- Three types of data analysis have been completed and through which business owners can make a variety of decisions towards their product.
- Various machine learning algorithms were applied to check their credibility to analyze the Sentiment, and satisfactory accuracy was turned out to be successful. This research also ensures the robust machine learning pipeline that achieved a good accuracy, and the concentrated model can be deployed to a webserver to achieve sustainable goals.
- The different assessment pointers bend assessed the proposed show, and at long last, a proposal framework has been submitted by coordination overall sifting strategy. By taking after the proposal framework, the

This study is supported via funding from Prince Satam bin Abdulaziz University project number (PSAU/2023/R/1444).

partner will be able to supply important data to their enlisted client, which can improve the request for specific things.

The manuscript is classified into six interconnected sections. Section II presents the exciting works with research gap analysis. Section III depicts the overall methodology of the proposed system with proper discussion. Section IV shows the results associated with our proposed solutions, including data-driven analysis and approaches. Section V represents the observations and discussion on the results. Finally, the conclusion of the research with future work will be discussed in Section VI.

II. LITERATURE REVIEW

This section illustrates the background study of the previous works related to the proposed model. In this section, the research gaps have been extracted with proper discussion. Many great contributors have traced fruitful online product contributions or sentiment analysis contributions.

The authors of the paper in [9] worked on an efficient way to optimize the accuracy of the sentiment analysis in Egyptian Arabic. The proposed work was identically based on the conventional semantic orientation and machine learning techniques, and the authors had achieved the highest accuracy of 92.98% while working with Support Vector Machine (SVM). The purpose of the article in [10] was to examine the attitudes of buyers regarding electrical devices by analyzing various sale tweets. The experimental results of the proposed research will be valuable to a variety of business organizations in making business decisions that will ultimately increase the sales of the products they offer. The author of the paper also claimed that they had achieved the highest accuracy of 86%, 91%, and 91% with the Logistic Regression (LR) in the phone, laptop, and television, respectively.

The paper in [11] aimed to extract the text features into the semantics of words. The authors adopted a Word Sense Disambiguation (WSD) technique to extract the features from the reviewing sentences. A supervised learning approach has been adopted to analyze the product reviews and utilized 10-fold cross-validation to validate the results. The authors had significantly optimized the performance by 10.6%, while precision was 10.9% higher and recall was 9.2% higher than baseline approaches. The author of the paper in [12] had presented a significant comparison among several conventional deep learning-based models for word embedding in product sentiment analysis. Thus, they adopted data augmentation techniques to enrich the dataset and classify it into identical classes. The research also claimed they found the highest accuracy of 96% while working with CNN-RNN based BI-LSTM algorithms.

In paper [13] proposed an Adaptive Neuro-Fuzzy Inferences System (IANFIS) model to produce a way of analyzing the sentiment of online products. The method is identically based on natural language processing to track the user's opinion. The authors classified the dataset into three interconnected parts: contents, grades, and collaborations. Then, they applied deep learning algorithms to make a prediction on the negative and positive comments from the users. The research also performs a comparison among the existing solutions. The paper in [14] aimed to implement a model for analyzing the sentiment of the users in movie reviews. The authors extracted the feeling and feedback from existing text patterns. The models had the ability to detect several types of feeling like negative, positive, and even neutral. To accomplish this goal, they utilized different machine learning algorithms and classifiers, and mechanisms of natural language processing.

The authors in [15] have presented a machine learning-based online product sentiment analysis. In this work, they showed the labeled product reviews in several websites with the help of supervised and unsupervised (lexicon-based) based algorithms. The models were then applied to the iPhone 5s reviews collected from the existing popular online shops. The authors further extracted the combination of unigram and bigram features, which placed the best results while dealing with machine learning-based classifiers.

Paper in [16] identified three subtasks that must be addressed: the definition of the target; the separation of good and bad news content from good and bad sentiment expressed on the target; and the analysis of clearly marked opinion that is defined explicitly, without the need for interpretation or the use of world knowledge. The authors in [17] created a new strategy that combines previous approaches to provide the best coverage results and competitive agreement. They had also proposed iFeel, a free Web service that provides an open API for retrieving and comparing findings from several sentiment methods for a given text. In paper [18], researchers categorized movie reviews using features based on these taxonomies paired with traditional "bag-of-words" features, and reported 90.2 percent accuracy. Furthermore, they discover that some types of assessment appear to be more critical for sentiment classification than others.

The contributions of the papers in [19] had only focused on the development of a notable features selection on online product sentiment analysis. But the researchers didn’t focus on the correct terms of algorithms and algorithm tuning to optimize sentiment analysis accuracy level. In sharp contrast, the manuscript presented three forms of data analysis. These analyses will allow business owners to make several judgments regarding their specific products. Various machine learning methods were used to check their reliability and determine adequate accuracy. This research also assures a robust machine learning pipeline that the condensed model can be deployed to a webserver to fulfill long-term objectives for product sentiment analysis.

Based on the literature review, several studies have been conducted to improve sentiment analysis and product recommendation using machine learning techniques. The authors of one study achieved the highest accuracy of 92.98% in sentiment analysis of Egyptian Arabic using Support Vector Machine. Another study analyzed buyer attitudes towards electronic devices through sale tweets and achieved high accuracy of 86%, 91%, and 91% for phone, laptop, and television respectively using Logistic Regression. One study aimed to extract text features through Word Sense Disambiguation and achieved a significant performance

improvement compared to baseline approaches. Lastly, a study compared several deep learning-based models for word embedding in product sentiment analysis and achieved an accuracy of 96% using a CNN-RNN based BI-LSTM algorithm.

Overall, the studies show the effectiveness of machine learning techniques in sentiment analysis and product recommendation. These approaches can help businesses make data-driven decisions to improve sales and customer satisfaction. However, there is still room for further research to optimize the accuracy and efficiency of these techniques.

III. METHODOLOGY

This section presents the overall design of the proposed model, including the illustration. Fig. 1 illustrates a block diagram of the proposed research through which this study was conducted. By looking at Fig. 1, it can be observed that this research was carried out through three interconnected stages. The experimental dataset was collected and preprocessed to model the data in the first stage. In the second phase, product sentiment analysis was accomplished. Finally, in the third stage, a recommendation system is proposed as the priority of this research is to provide a model that can analyze the product sentiment and transform the traditional business into a data-driven approach.

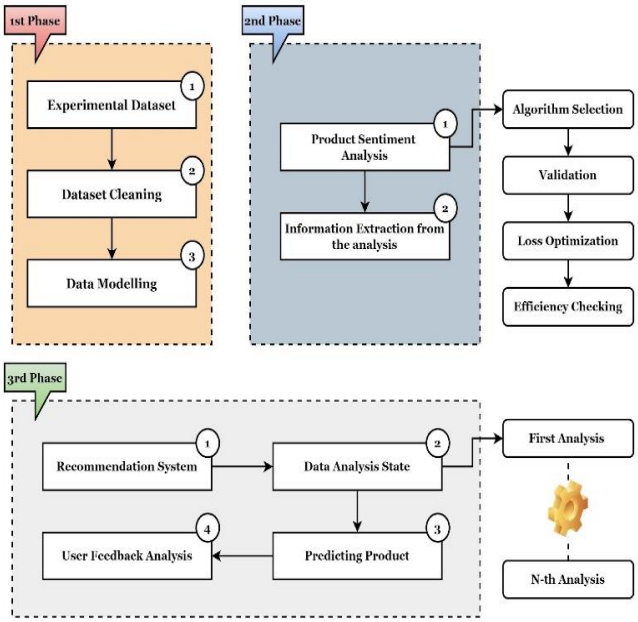


Fig. 1. Block diagram of the proposed research, including the experimental analysis to product recommendation approach.

A. Sentiment Analysis Approach (SAP)

This section explains the steps of approach, including experimental data, data preparation and concentrated research algorithms in more details.

1) Experimental dataset: From May 1996 through July 2018, Amazon's "Clothing, Shoes, and Jewelry" category received 2.5 million product ratings and information from 2.5 million customers. This collection includes reviews (scores, description, and sentiment comments), product metadata (descriptions, controls and monitors, pricing, branding, and image attributes), and links [20].

2) Data preparation: Perusing different JSON records from a single JSON record, 'ProductSample.json,' and including them in the list in such a way that each list of the list has the substance of a single JSON record [21]. Following that, iterate over the list, loading each index as JSON, extracting the data from each index, and creating a list of Tuples containing all the data from the JSON files. Again, each cycle begins with a clean JSON file converted to the right JSON format using some substitutions. Finally, a data frame is created using the list of Tuples obtained in the previous step.

3) Concentrated research algorithms (CRA): The Concentrated Research Algorithms (CRA) indicates the suggested model that has been adopted in this study to analyze the research data [22]. It is to be specified that various conventional techniques were applied in this investigation; among them, the Naïve Bayes and Decision tree algorithms were found to be satisfactory [23]. So, the mathematical interpretation and model optimization procedure were highlighted in this section. These models have a benchmark performance that appeared in the previous research. In addition, model performance is depending on the data distribution, furthermore, the Naïve based and Decision tree are capable enough to handle the product sentiment analysis data.

- Naïve Bayes Algorithm: Nave Base Classifier is a classification-type machine learning algorithm [24]. This algorithm is based on the Base Theorem. Simply put, the base theorem is a method of determining the probability of one event (X) occurring and another event (Y) occurring. If clouds are seen in the sky, there is a possibility of rain. The base theorem can be mathematically written as,

$$\frac{P(y)P(y)}{P(X)} \tag{1}$$

The above equation is a simple base equation used to determine probability in the case of a conditional event only. In practice, most datasets are multivariate, in which case the equation becomes a bit more complicated. Then we can write the equation like this:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)} \tag{2}$$

A few of the highlights of Naïve Base Classifier: It is exceptionally simple to execute and works moderately quick, works well indeed on small datasets, gives a small less exactness than other calculations, and all traits in Naïve Base are considered commonly autonomous but within the genuine world Isn't.

- Decision Tree: Both classification and regression problems can be solved with the classification and regression tree or CART algorithm [25]. In short, many people call it the Decision Tree. The decision tree looks

a lot like the branches of a tree, which is why the word 'tree' is associated with its name. The decision trio starts from the 'root node' just as the tree starts from the root. From the root node, the branches of this tree spread through different decision conditions; such nodes are called decision nodes, these nodes are called leaf nodes after making a final decision. Other Parameters of the Decision Tree: Splitting - The process of moving a dataset across a series of variables, starting from the root node, is called splitting [26]. Entropy - Entropy is the amount of chaos. When the tree is split, the amount of data of the same type/class in each node is purity. All the data in a pure node are of the same class. The lower the purity, the higher the entropy. Again, the lower the entropy, the higher the purity. Information Gain - The measure of righteousness is information gain [27]. The higher the information gain, the purer nodes the tree can create. Guinea Index - Guinea is the probability of all node members being in the same class. This value ranges from 0 to 1. Guinea value 0 means all the members of that node belong to the same category, and Guinea value 1 means the members of that node are randomly distributed or of different classes, i.e., entropy is much higher. If the value of Guinea is 0.5, then the members of the two classes are equal (if the number of classes is 2) [28].

$$Entropy = p(A)log(p(A) - p(B)log(p(B \quad (3)$$

Information Gain= Entropy Before Split - Entropy After Split

$$Gini\ Index\ = 1 - \sum(P(x = k))\text{^}2 \quad (4)$$

B. Product Recommendation Approach (PRA)

This section highlighted the product recommendation procedure. It can be said that the PRA is vital towards business transformation because user behavior and pattern cannot easily be identified if the stakeholder did not design any recommendation system. By looking at Fig. 2, it is noticeable that a flow chart has been proposed in terms of the recommendation system. The collaborative filtering approach is selected that will filter out the user ID based on age, gender, location and rating score etc. After having all of that information, the system will make a comparison set for the specific users. However, the detailed sequence and consequences are shown in Fig. 2. The diagram in Fig. 2 suggests that the proposed pipeline aims to improve e-commerce and enhance customer experience by using a collaborative filtering method for item recommendation. It implies that the pipeline could help businesses identify the most relevant products for their customers, thus improving their overall shopping experience. The use of collaborative filtering suggests that the pipeline may leverage the behavior of similar users to provide personalized recommendations, ultimately leading to increased customer satisfaction and potentially higher sales. Overall, the diagram title hints at a promising approach to improving the online shopping experience and driving business growth.

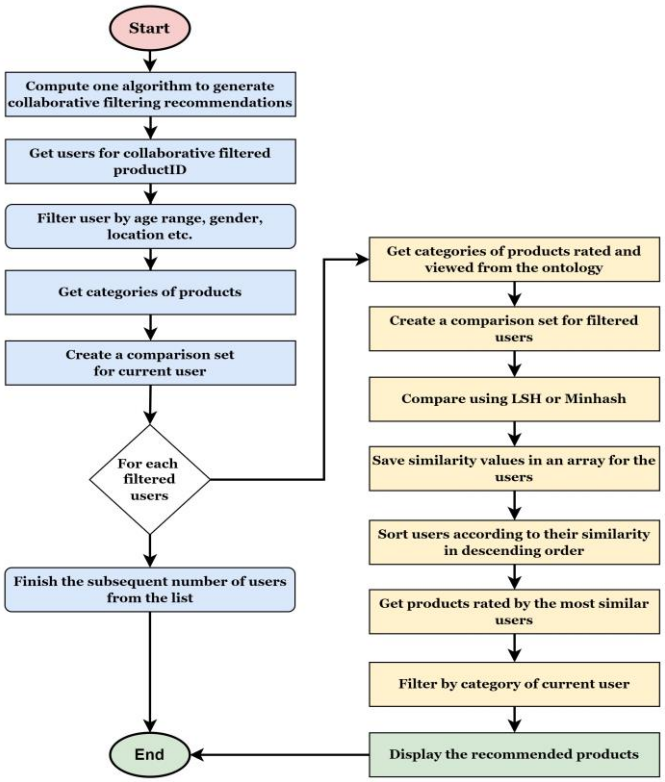


Fig. 2. Efficient pipeline for the item proposal approach towards commerce change and client design acknowledgment through the collaborative sifting method.

The research system utilizes several methods like K-Nearest Algorithm (KNN) [29], the Jaccard's coefficient, the Dijkstra algorithm, and the cosine similarity. The aim is to suggest based on users' behavior patterns. In recommendation systems, the most common types are the Collaborative Filtering Method (CFM), the Content-Based Filtering Approach (CFA), and Hybrid Recommendation System (HRS) [30]. This filtering approach generally focuses on collecting and analyzing user experience information, behaviors, or interests and predicting what they would like based on similarity with other users. The collaborative sifting approach's imperative advantage is that it does not depend on machine analyzable substances and can accurately prescribe complex things without requiring an "understanding" of the thing itself.

A typical recommendation engine processes data over the following four steps: selection, storage, analysis, and filtering. We have applied the K-Nearest Algorithm (KNN), the Jaccard's coefficient, the Dijkstra algorithm, and the cosine similarity to forecast the shortest path from the user's current location to the user's desired destination and as well as to suggest places to the users based on the rating. In Figure 3 a) shows the corresponding KNN's cluster filtering working procedure and b) shows the Dijkstra algorithm 's workflow.

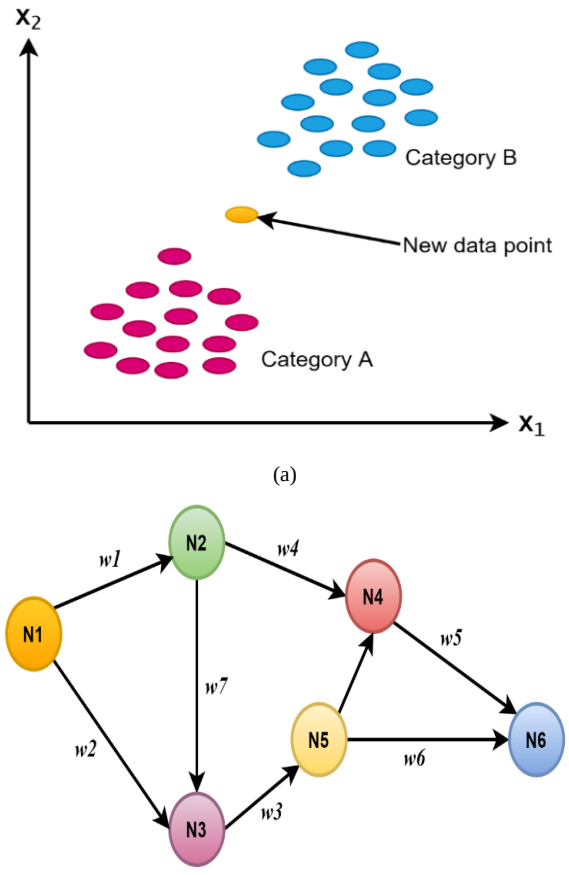


Fig. 3. a) Perform K –nearest algorithm's cluster filtering, b) Dijkstra algorithm.

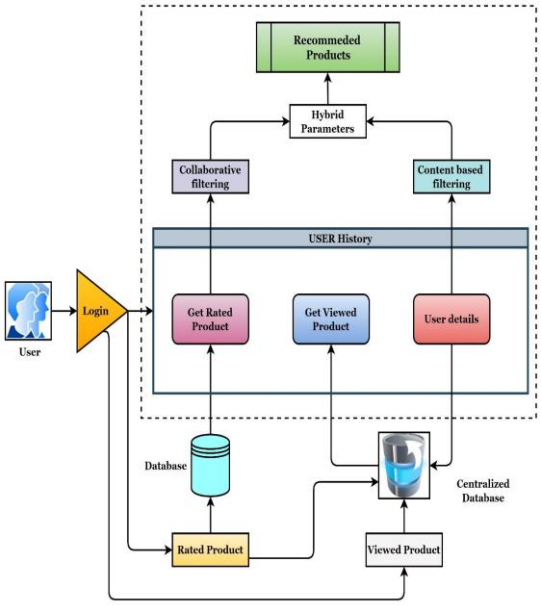


Fig. 4. Architecture diagram of the product recommendation system using various filtering methods.

In Fig. 4, it can be stated that the suggested recommendation engine is the result of a number of interconnected methodologies being used. In order to utilize the system, the user must first complete their registration before being allowed to proceed. After logging into a system, a log file will be created automatically to keep track of the user's patterns of behavior. At this point, we have implemented a cooperating filtering method as well as a content-based filtering approach. Users' comments and product ratings will be taken into consideration by the recommendation system, which will take action based on both forms of data. After obtaining all of the necessary parameters, the system will proceed to the next stage of proposing a specific product to a user who has expressed interest in it. The architecture also assures that collaborative and content-based filtering are the middle layer of this architecture that are the responsible for the product sentiment analysis. However, the hybrid parameters will then send to the model for identifying the sentiment.

IV. Result Analysis

The result analysis section is categorized into several parts: Classification Metrics Interpretation (CME), Measuring the Efficiency, Interpretation of Sentiment Analysis, Observation & Discussion. The precision of expectations from the classification calculations is evaluated by applying a classification report. The report illustrates the exactness, review, and f1-score of the key classification measurements per lesson. These measurements are computed by utilizing genuine and untrue positives and genuine and wrong negatives. The measurements comprise of four components: genuine positive, untrue positive, genuine negative, wrong negative, and wrong negative. The taking after Condition (1), (2), (3), and (4) was considered for finding the exactness, review, and f1-score. In Table I, the classification report of the machine learning calculation, is depicted. In this table, we can clearly observe that the machine learning algorithms like Random Forest (RF) and Logistic Regression provide better results, such as 95% and 94%, compared to the other algorithms like LSTM and CNN_LSTM. This is because for this dataset, we have found that low-cost classifiers work far better at deep computation because of the small size of the dataset. Thus, we have achieved the highest accuracy from conventional classifiers.

Precision: It is the relationship between the true positive estimate of the model and the overall positive estimate (both accurate and wrong) .It is articulated as:

$$Precision\ (P) = \frac{TP}{TP+FP} \tag{5}$$

Recall / Sensitivity: The probability of being capable of predicting is a positive ratio. It is given in mathematical form as:

$$Recall\ (R) = \frac{TP}{TP+FN} \tag{6}$$

F1-score: As a general rule, the harmonic mean for Accuracy and Review provides a much better; a significantly better; higher; a stronger and more enhanced gauge than the Precision Metric of the incorrectly categorized occurrences. It is given, mathematically, as:

$$F1-\text{score} = 2 \cdot \frac{Precision.Recall}{Precision + Recall} \tag{7}$$

Accuracy: It is the sum of all the cases in which the predictions were right. It is given as:

$$Accuracy\ (A) = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

TABLE I. CLASSIFICATION REPORT OF THE CONCENTRATED ALGORITHMS

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Naïve Bayes (NB)	93	92	93	93
Logistic Regression (LR)	94	93	94	94
SVM	93	93	93	93
Decision Tree (DT)	91	89	89	90
Random Forest (RF)	95	94	95	95
BI-LSTM	76	70	71	70
CNN_LSTM	77	77	75	76
Stacked LSTM	76	76	76	76
LSTM	79	78	78	77

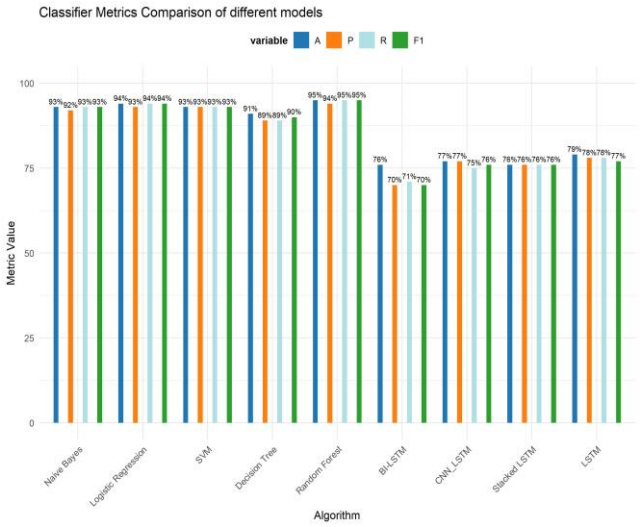
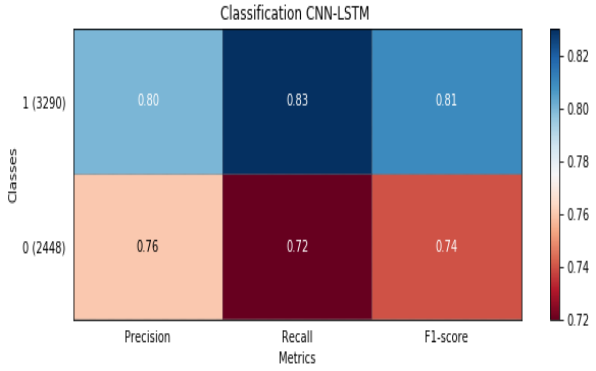


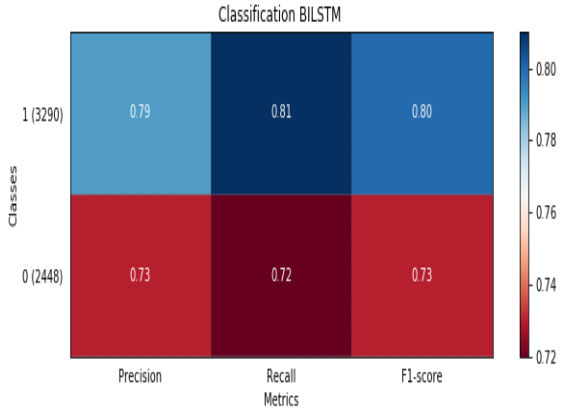
Fig. 5. Performance analysis from the different models.

On the other hand, Fig. 5 highlights the confusion matrix on top of the Random Forest and LSTM model. The confusion matrix consists of four values: TP, TN, FP, and FN. We can identify the total sensitivity and specificity ratio by following the Confusion matrix. This is another model evaluation indicator, and in the field of data science, this matrix has been utilized extensively to measure a specific model. The confusion matrix consists of four values, True positive, True negative, False positive, and False negative are the four values in the confusion matrix. Fig. 6 (a) and (5) illustrates the significant proportion of true positive and false negative values.

ROC-AUC curve is used to determine how good a model is. This evaluation indicator distinguishes the positive and negative data points from the dataset. If the ROC curve goes to 1.0, the model can accurately differentiate the positive and negative data points. Fig. 7 (a) and (b) are almost close to 1.0 or the area under the curve. It can be stated that this model is applicable to use in real life.

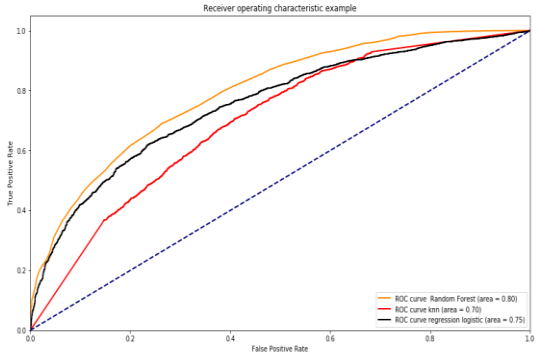


(a)

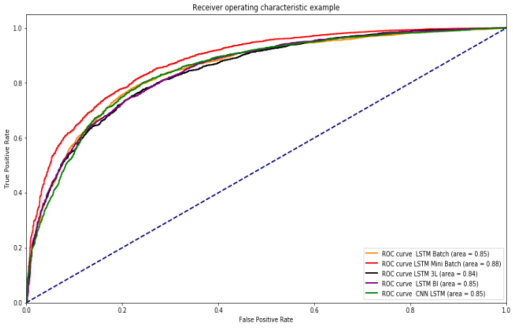


(b)

Fig. 6. (a) Confusion matrix on CNN+LSTM (b) Visualizing the confusion matrix on top of the BI-LSTM model towards product sentiment analysis.



(a)



(b)

Fig. 7. (a) Evaluating curve on top of the Random Forest model (b) Measuring the model and visualizing the ROC-AUC curve on top of the LSTM model.

A. Interpretation of Sentiment Analysis

This analysis is divided into three stages: Sentiment analysis on reviews (1996-2018), exploratory analysis on product reviews (1996-2018). This is of particular note since it is the one with the greatest number of prior sales on Amazon, as determined by an examination of "Bundles" or "Bought-Together."

B. Analysis 1: Sentiment Analysis on Reviews (1996-2018)

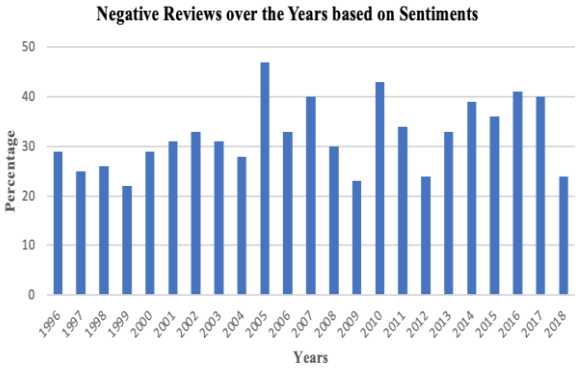
Fig. 8 indicates the exploratory product data analysis in terms of user feedback. In Fig. 8 (a) and Fig. 8 (b), the sentiment analysis on user reviews have been demonstrated for the year of between 1996 to 2018. It is noticeable that, the year 2000 was the least number and 2001 reach out at the highest number of negative product reviews based on the sentiment. At the same time, the year 2000 was the top year where significant number of positive reviews based on the users sentiment was recorded. After that, the percentage was decreased and remained same for the next consecutive years. Besides, Fig. 7 (c) and (d) shows the word cloud visualization based on the positive and negative reviews where it the keywords are tagged and, in this way, it would be convenient to find out the negative and positive keywords from the dataset.

For Positive Word Cloud, some well-known terms such as adored, idealized, decent, wonderful, best, outstanding, and so on were used to describe the goods. A large number of people who were polled were pleased with the prices of things sold on Amazon. Bra, coat, bag, and outfit are some of the most commonly discussed goods. Disappointment, terrible fit, horrible deformity, return and etc. are some of the well-known adjectives used to describe the things. Some of the most talked-about goods were shoes, binoculars, bras, batteries, and so on. Predominant item in terms of how others feel about it. There are 953 positive reviews for the Speak Unisex Chuck Taylor Classic Colors Sneaker. There are 672 positive reviews for the Talk Unisex Chuck Taylor All-Star Howdy Best Dark Monochrome Sneaker. There are 65 negative reviews for the Yaktrax Walker Footing Cleats for Snow and Ice. There are 44 negative reviews for the Speak Unisex Chuck Taylor Classic Colors Sneaker. Welcome to the Best Dark Monochrome Sneaker, with a total of 247 honest reviews. [31], which state achieved that 72.7 % was positive, 5 % was negative, and 22.3 % was neutral. Overall, Sentiment for reviews on Amazon is on the positive side as it has very few negative sentiments.

The drift for Rate of Audit over a long time, positive surveys rate has been lovely reliable between 70-80 all through a long time. Negative surveys have been diminishing recently since final three a long time; possibly they worked on the administrations and issues.

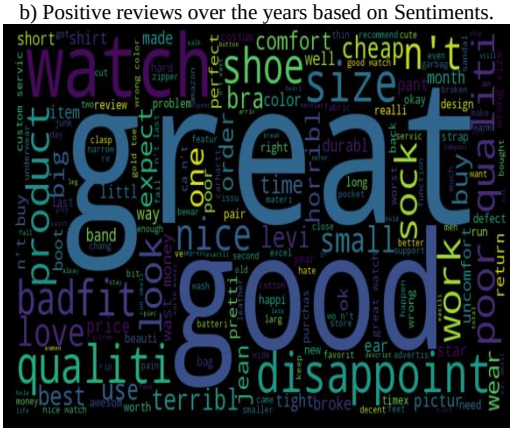
C. Analysis 2: Point of Interest-based Analysis

The lexical density of a language is a concept in computational linguistics that measures the structure and complexity of human conversation in [31]. Functional and content words are used to calculate a written or spoken composition's lexical density.

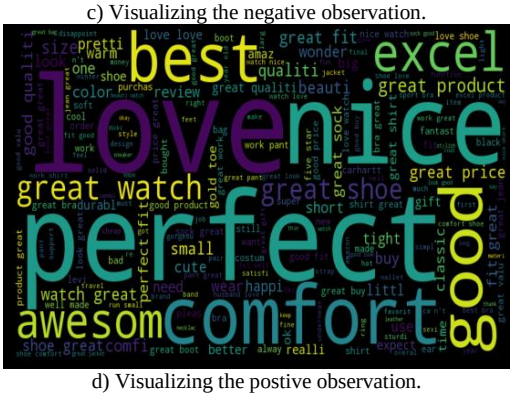


a) Negative reviews over the years based on sentiments.

b) Positive reviews over the years based on Sentiments.



c) Visualizing the negative observation.



d) Visualizing the postive observation.

Fig. 8. a) Negative reviews over the years based on the Sentiment (b) Positive reviews based on the Sentiment (c) Visualizing the negative observation through the word cloud visualization approach (d) Illustrating the positive observation with the help of word cloud visualization.

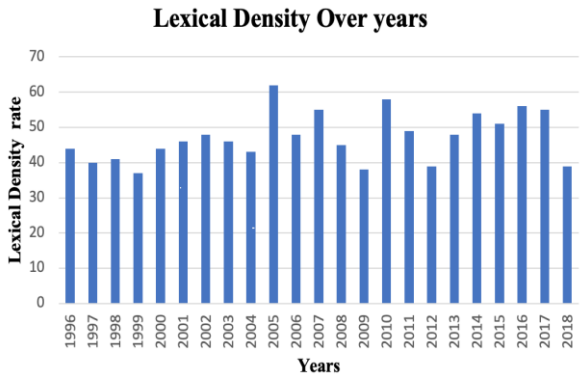


Fig. 9. Lexical density over years.

By looking at the Fig. 9, it can be observed that lexical density over years have been displayed. In 2018, the significant amount just over the 40 but in 1996, there is a downward trend at nearly 36 and remained steady for the next subsequent years.

D. Analysis 3: Bundle' or 'Bought- Together' based Analysis

In the Table II, Bundle or Bought together based analysis has been interpreted in terms of up vote, helpful rating, total votes, and percentage. Based on the reviewer ID, it is clearly demonstrated the helpful rating, and these has been illustrated owing to the fact that for the case of analysis, product sentiment analysis and review records are considerably required so that essential information can be extracted.

Taking a close look at the Fig. 10, helpfulness and average length have been displayed. The findings show that the effectiveness of review length is influenced by product category; longer evaluations are more helpful for think products. Furthermore, review helpfulness is linked to the degree of consistency between individual review ratings and total product ratings. In contrast the Fig. 11 shows the data analysis with correlation between the ASIN.

E. Analysis 3: Exploratory Data Analysis

1a) The common survey rating for the foremost commonly checked items is between 4.5 and 4.8, with little variation. 1b) Whereas there's a little converse affiliation between the recurrence level of ASINs and regular audit appraisals for the primary four ASINs, this relationship isn't critical since the normal survey for the prior four ASINs is evaluated between 4.5 and 4.8, which is respected greatly by and large audits. 2a) As illustrated within the bar chart (beat), ASINs with lower frequencies have much more change in their routine survey evaluations on the point-plot chart (foot), as demonstrated by the length of the vertical lines. As a result of the tall fluctuation, we accept that our research's normal audit evaluations for ASINs with lower frequencies are not imperative. 2b) On the other hand, we assume that the lower frequencies for ASINs are inferable to lesser quality items. 2c) Moreover, the final four ASINs have no change due to their significantly lower frequencies. Whereas the survey appraisals are a culminate 5.0, we ought not to consider these audit appraisals' significance due to the lower recurrence as shown in 2a). Based on our information examination between ASINs and audits. Rating, we have taken note that numerous ASINs with the common event had huge fluctuations; in this way, we decided that these moo event ASINs are not imperative in our think about due to the little test measure. Additionally, we found nearly no interface between ASINs and surveys. Rating in our relationship considers which is reliable with our findings.

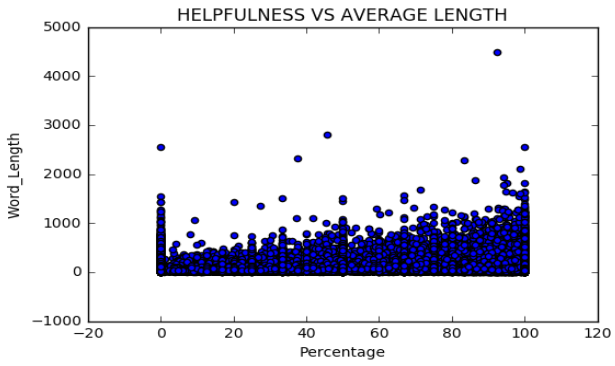
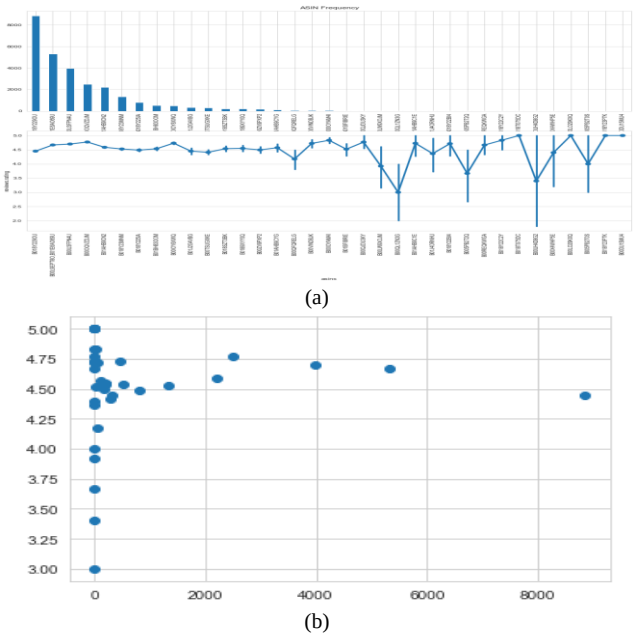


Fig. 10. Helpfulness and average length.

TABLE II. EXPLORARYTORY DATA ANALYSIS WITH THE DATASET AND THEIR FEATURES

Reviewer ID	Rating helpful	Upvote	Total Votes	Percentage	Rating
A2XVJBSRI3SWDI	5.0	0.0	N/A	0.0	N/A
A2G0LNLN79Q6HR	4.0	1.0	0.0	N/A	N/A
A2R3K1KX09QBYP	2.0	1.0	100.0	N/A	N/A
A19PBP93OF896	1.0	1.0	100.0	N/A	N/A
A19PBP93OF896	0.0	0.0	0.0	N/A	N/A
A0000188NWOSI5X2PMS	0.0	N/A	0.0	1.0	N/A
A000063614T1OE0BUSKUT	N/A	N/A	N/A	0.0	5.0
A00031045Q68JAQ1UYT	N/A	N/A	N/A	4.0	N/A
A00028781NF0U7YEN9U19	N/A	N/A	N/A	0.0	5.0
A00031045Q68JAQ1UYT	N/A	N/A	N/A	100.0	1.0



(a)

(b)

Fig. 11. (a) Exploratory data analysis on top of the ASIN (b) Correlation analysis between ASIN.

V. Discussions

The foremost regularly looked into products have routine audits within the 4.5 to 4.8 extend, with small fluctuation. Even though there's a slight converse relationship between the ASINs recurrence level and normal audit evaluations for the primary 4 ASINs, this relationship is immaterial. The normal survey for the prior 4 ASINs is evaluated between 4.5 to 4.8, which large surveys consider great. For ASINs with lower frequencies, we see that they're comparing normal audit evaluations on the point-plot chart (foot) have an essentially bigger change, as appeared by the length of the vertical lines. As a result, we recommend that the normal audit appraisals for ASINs with lower frequencies are not critical for our examination due to high variance.

On the other hand, due to their lower frequencies for ASINs with lower frequencies, we recommend that this result from more second-rate quality items 2c). Moreover, the final 4 ASINs have no fluctuation due to their altogether lower frequencies. Even though the audit evaluations are a culminate 5.0, we ought not to consider the centrality of these review evaluations due to lower recurrence. I am able to see that certain items have much more reviews than others based on the ASIN investigation, proposing a greater deal for those items. Ready to see that the ASINs have "right-tailed" dissemination, demonstrating that specific things have bigger deals, which can be connected to the higher recurrence of ASINs within the audits. Moreover, we took the log of the ASINs to normalize the information so that we seem to get a more nitty-gritty to see each ASIN and see that the dissemination is still "right-tailed." In our study, product sentiment analysis has been carried out towards in the year between 1996 to 2018. Three types of data analysis have been completed and through which business owners can make a variety of decisions towards their particular product. The different assessment pointers bend assessed the proposed show, and at long last, a proposal framework has been submitted by coordination overall sifting strategy. By taking after the proposal framework, the partner will be able to supply important data to their enlisted client, which can improve the request for specific things.

VI. Conclusion and Future Work

Nowadays, product sentiment is very important because, when a business is run online, it is important for every user to recommend their various products through pattern recognition. In order to use cutting-edge machine learning and deep learning algorithms to evaluate online product sentiment and make recommendations, a thorough pipeline is needed. This research proposes a pipeline for analyzing the online product. Also, a recommendation system has been presented through which a similar product can be filtered out for users. The study approach comprises two distinct components, namely the sentiment analysis approach and the product recommendation approach. The study uses appropriate hyperparameter optimization techniques to apply a number of cutting-edge algorithms, such as Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, Bidirectional Long-Short-Term Memory (BI-LSTM), Convolutional Neural Network (CNN), Long-Short-Term Memory (LSTM), and Stacked LSTM. The k-Nearest Neighbors (KNN) model is combined with the collaborative filtering approach in the study to make product recommendations. Of these models, the Random Forest model had the highest accuracy (95%), followed by the LSTM model (79%). The Area under the ROC Curve (AUC), also known as the Receiver Operating Characteristic (ROC) curve, is used to evaluate the proposed model. In addition, the study carried out exploratory data analysis on reviews (1996–2018) using point-of-interest-based analysis, sentiment analysis, and bundle or bought-together analysis. Overall, the study meets its goals and suggests a flexible fix for practical situations.

Different machine learning and deep learning algorithms were applied to analyze the sentiment in this research. The Random Forest was found to be satisfactory through the investigation and can recommend any product effectively. In addition, three types of analysis have been carried out in this study. This research has several limitations, such as experimenting with just one dataset, but experimenting on multiple datasets was required, which we will complete in the future. In addition, a software system will be developed where a recommendation system will be integrated. Also, various loss optimization formulas will be applied to ensure model efficiency. Evaluation indicator approaches will later justify the type of model followed for recommendation in this phase.

References

[1] G. S. Blair, K. Beven, R. Lamb, R. Bassett, K. Cauwenberghs et al., “Models of everywhere revisited: a technological perspective,” Environmental Modelling & Software, vol. 122, pp. 104521, 2019.

[2] N. Mantel and W. Haenszel, “Statistical aspects of the analysis of data from retrospective studies of disease,” National Cancer Institute, vol. 22, no. 4, pp. 719–748, 1959.

[3] B. Jiang, “Investigating the market strategy of smart food’s latest product based on business analysis models," in 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021), Guangzhou, China, vol. 43, pp. 2218-2223. 2021.

[4] A. Schwarz, S. Isaksson, U. Källman and M. Rusner, "Enabling patient safety awareness using the Green Cross method: A qualitative description of users' experience," Journal of clinical nursing, vol. 30, no. 5, pp.830-839, 2021.

[5] D. Gavilan and M. Gema, "Exploring user's experience of push notifications: a grounded theory approach," Qualitative Market Research: An International Journal, vol. 3, pp. 35-57, 2022.

[6] S. Abuelenin, S. Elmougy and E. Naguib, "Twitter sentiment analysis for arabic tweets," in International Conference on Advanced Intelligent Systems and Informatics, London UK, pp. 467-476 .2017.

[7] S. Kumar, V. Koolwal and K. K. Mohbey, "Sentiment analysis of electronic product tweets using big data framework," Jordanian Journal of Computers and Information Technology, vol. 5, pp. 43-59, 2019.

[8] B. Setya Rintyarna, R. Sarno and C. Fatichah, "Semantic features for optimizing supervised approach of sentiment analysis on product reviews," Computers, vol. 8, p. 55, 2019.

[9] V. Balakrishnan, Z. Shi, C. L. Law, R. Lim and Y. Fan, "A deep learning approach in predicting products' sentiment ratings: a comparative analysis," The Journal of Supercomputing, vol. 4, pp. 1-21, 2021.

[10] P. Sasikala and L. Mary Immaculate Sheela, "Sentiment analysis of online product reviews using dlmnn and future prediction of online product using IANFIS," Journal of Big Data, vol. 7, pp. 1-20, 2020.

[11] Y. Zhu, J. Jiang, W. Han, Y. Ding and Q. Tian, "Interpretation of users' feedback via swarmed particles for content-based image retrieval," Information Sciences, vol. 375, pp. 246-257, 2017.

[12] W. Muhammad, M. Mushtaq, K. N. Junejo and M. Y. Khan, "Sentiment analysis of product reviews in the absence of labelled data using supervised learning approaches," Malaysian Journal of Computer Science, vol. 33, pp. 118-132, 2020.

[13] L. Monsurrò and L. Dezi, "Elderly experience of smart objects: how technology and family support can make senior users overcome their limits," In 2021 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD), Marrakech, Morocco, pp. 1-5, 2021.

[14] K. Vembandasamy, R. Sasipriya and E. Deepa, "Heart diseases detection using naive bayes algorithm," International Journal of Innovative Science, Engineering & Technology, vol. 2, pp. 441-444, 2015.

[15] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," Journal of Applied Science and Technology Trends, vol. 2, pp. 20-28, 2021.

[16] Z. Zhang, "Introduction to machine learning: k-nearest neighbours," Annals of translational medicine, vol. 4, pp. 345- 356, 2016.

[17] M. Almaliki, C. Ncube and R. Ali, "The Design of adaptive acquisition of users feedback: an empirical study," 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS), Marrakech, Morocco, pp. 1-12, 2014.

[18] N. Sherief, W. Abdelmoez, K. Phalp and R. Ali, "Modelling users' feedback in crowd-based requirements engineering: an empirical study," In IFIP Working Conference on The Practice of Enterprise Modelling, pp. 174-190. Springer, Cham, 2015.

[19] S. Francesco, Lorenzo, B. Cristian, L. Danza, M. Ghellere et al., "Integrated method for personal thermal comfort assessment and optimization through users' feedback, IoT and machine learning: A case study," Sensors, vol. 18, no. 5, pp. 1602, 2018.

[20] S. Manikandan, A. Delphincarolinarani, C. Rajeswari, T. Suma and D. Sivabalaselvamani, "Recognition of font and tamil letter in images using deep learning," Applied Computer Science, vol. 17, no. 2, pp. 90–99, 2021.

[21] P. N. Tan, M. Steinbach and V. Kumar, "Introduction to data mining," in The Pearson. Upper Saddle River, NJ, USA: Pearson,vol. 54, pp. 37–69, 2006.

[22] A. Saurav, "Amazon Review Data (2018)," Amazon review, 06 March, 2022, Available: https://nijianmo.github.io/amazon/index.html.

[23] K. McCartan, A. Danielle, F. Harris and S. David, "Seen and not heard: The service user's experience through the justice system of individuals convicted of sexual offenses," International journal of offender therapy and comparative criminology, vol. 65, no. 12, pp. 1299-1315, 2021.

[24] D. Gavilan and G Martinez-Navarro, "Exploring user's experience of push notifications: a grounded theory approach. Qualitative Market Research," An International Journal, vol. 4, 2022.

[25] P. Sajjadi, L. Hoffmann, P Cimiano, and S. Kopp, "A personality-based emotional model for embodied conversational agents: Effects on perceived social presence and game experience of users," Entertainment Computing, vol. 32, pp.100313, 2019.

[26] B. Sovacool, J. Osborn, M. Martiskainen and M. Lipson, "Testing smarter control and feedback with users: Time, temperature and space in household heating preferences and practices in a Living Laboratory," Global Environmental Change, vol. 65, pp.102185, 2021.

[27] W. Wanyuan, Y. Jiang and W. Wu, "Multiagent-based resource allocation for energy minimization in cloud computing systems," IEEE Transactions on Systems Man and Cybernetics: Systems, vol. 47, no. 2, pp. 205–220, 2016.

[28] S. Kumar and M. Singh, "Big data analytics for healthcare industry: Impact, applications, and tools," Big Data Mining and Analytics, vol. 2, no. 1, pp. 1–20, 2019.

[29] K. Sahu, F. A. Alzahrani, R. K. Srivastava and R. Kumar, "Evaluating the impact of prediction techniques: Software reliability perspective," computers," Materials & Continua, vol. 67, no. 2, pp. 1471–1488, 2021.

[30] J. Athinarayanan, V. S. Periasamy, M. Alhazmi, K. A. Alatiah and A. A. Alshatwi, "Synthesis of biogenic silica nanoparticles from rice husks for biomedical applications," Ceramics International, vol. 41, no. 1, pp. 275–281, 2015.

[31] S. O. Proksch, W. Lowe, J. Wäckerle and S. N. Soroka, "Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches," Legislative Studies Quarterly, vol. 44, no. 1, pp. 97–131, 2019.

Automatic Model for Postpartum Depression Identification using Deep Reinforcement Learning and Differential Evolution Algorithm

Sunyuan Shen[1*], Sheng Qi[2], Hongfei Luo[3]
Department of Applied Engineering, Zhejiang Business College Hangzhou 310053, People's Republic of China[1, 3]
Institute of Electronic Commerce, Zhejiang Business College Hangzhou, 310053, People's Republic of China[2]

***Abstract*—Postpartum depression (PPD) affects approximately 12% of new mothers, posing a significant health concern for both the mother and child. However, many women with PPD do not receive proper care. Preventative interventions are more cost-effective for high-risk women, but identifying those at risk can be challenging. To address this problem, we present an automatic model for PPD using a deep reinforcement learning approach and a differential evolution (DE) algorithm for weight initialization. DE is known for its ability to search for global optima in high-dimensional spaces, making it a promising approach for weight initialization. The policy of the model is based on an artificial neural network (ANN), treating the categorization issue as a policymaking stage-by-stage process. The DE algorithm is used to acquire initial weight values, with the agent obtaining samples and performing classifications in each step. The habitat provides an award for every categorization activity, considering a greater award for identification of the minor category to encourage precise detection. By using a particular compensatory technique and an encouraging learning system, the operator eventually decides the most excellent method for achieving its goals. The model's efficiency is evaluated by analyzing a set of data acquired from the population-based BASIC study carried out in Uppsala, Sweden, which covers the period from 2009 to 2018 and consists of 4313 samples. The experiential results, identified by known analysis criteria, indicate that the sample achieved better precision and correctness, making it suitable for identifying PPD. The proposed model could have significant implications for identifying at-risk women and providing timely interventions to improve maternal and child health outcomes.**

Keywords*—Postpartum depression; deep reinforcement learning; differential evolution algorithm; weight initialization; artificial neural network

I. Introduction

PPD is a common condition in Sweden, affecting 8% to 15% of new mothers annually [1]. It manifests as mild to severe depressive episodes either during pregnancy or within the first year after giving birth [2, 3]. The exact cause of PPD remains unknown but is thought to result from a combination of psychosocial, psychological, and biological factors. Biologically, inflammation, the withdrawal of allopregnanolone, and genetic factors play roles. Psychosocially, factors like ongoing stress, prior depression, relationship difficulties, and significant life changes contribute to PPD risk. The consequences of PPD can be severe, impacting both mother and child. Mothers may struggle with forming emotional bonds with their child, doubt their caregiving abilities, and even have harmful thoughts towards the child [4]. Efforts have been made to predict PPD during the prenatal period. Still, currently, there is no reliable method to accurately identify women at risk of experiencing depressive symptoms after giving birth [5].

Conventional statistical methods typically analyze the relationship between two variables while factoring in other variables [6, 7]. In contrast, machine learning (ML) techniques allow for the simultaneous analysis of many interconnected variable relationships, leading to the creation of data-driven predictive models [8]. These models can then be assessed to find the most effective predictor. ML can handle complex nonlinear relationships and integrate various data types from different sources. Over the past ten years, the application of ML has expanded across medical fields including oncology, cardiology, hematology, critical care, and psychiatry. In PPD, which poses a moderate risk of a serious psychiatric condition with reasonably accurate prediction of symptom onset, ML can be highly valuable given the societal impact of PPD. Despite its potential benefits, it is impractical to monitor every individual for early PPD symptoms. A more efficient strategy is to target high-risk groups during postpartum checks by healthcare professionals like midwives or nurses, rather than the broader population. In Sweden, with its 120,000 annual births and the myriad of post-childbirth adjustments women undergo, and a typical PPD prevalence of around 12%, this targeted approach proves especially advantageous for personalized, cost-effective maternal and perinatal mental care.

Machine learning can face issues with feature extraction, affecting generalization, processing time, and precision [9]. The rise of deep learning, particularly Multi-Layer Perceptron (MLP), offers improved classification capabilities [10]. MLP, designed for nonlinear XOR problems, is versatile for various tasks, from image processing to optimization [11]. It functions like human neurons, where each node in ANN processes inputs and uses an activation function to produce an output. In MLP, nodes are interconnected across different layers, without intra-layer connections.

Medical classification poses significant challenges due to imbalanced data, where negative instances far outnumber positive ones, leading to decreased performance [9, 12, 13].

Measures can be employed at both the algorithm and data levels to address this issue. At the level of the data, downsampling, upsampling, or a mixture of both techniques can be utilized to alleviate the negative impact of imbalanced classification [14, 15]. On the other hand, algorithmic approaches involve assigning greater weight to the minority class [16, 17]. Moreover, deep learning methods offer potential solutions for tackling imbalanced classification [12, 18]. Huang et al. [19] have proposed a process to identify distinctive features in imbalanced data while maintaining inter-cluster and inter-class margins. Similarly, Yan et al. [16] have suggested a technique using the bootstrapping method to balance data in convolutional networks across mini-batches.

Population-based training can be utilized to select the most optimal solution from a population of generated models in order to optimize neural networks [20-22]. This approach mitigates the risk of being trapped in local optima, a common challenge in traditional training methods [23]. Surprisingly, a straightforward evolutionary algorithm has proven comparable to stochastic gradient descent in terms of the effectiveness of neural network training [24, 25]. Jaderberg et al. [26] successfully applied population-based training to cutting-edge models in deep reinforcement learning, machine translation, and generative adversarial networks, yielding consistent accuracy, training time, and stability enhancements. In related studies [27] and [28], effective weight training for neural networks was achieved through the adoption of differential evolution-based strategies [29] and the employment of the ABC (Artificial Bee Colony) method [30, 31], respectively.

This paper introduces a novel approach for identifying PPD by combining deep Q-learning and the DE method to initialize the load. The categorization task is formulated as an estimating challenge within an RL framework, treating it as a Markov decision process. The environment state is represented by a sample, and the agent is an ANN. To initiate the game, we explore the application of the DE algorithm to find an optimal weight initialization for the ANN. The agent classifies each sample, and its classification is awarded accordingly by giving the right choices positive awards and wrong decisions getting negative awards. For tackling dataset imbalance, the minority class is given a higher absolute value of the reward. The operator aims to amplify the accumulative awards by accurately classifying the samples throughout the policymaking procedure. Significantly, our research is pioneering in utilizing a population-based methodology that leverages an extensive and varied dataset, incorporating a wide range of clinical, psychometric self-report, and medical journal-derived variables. The performance of our proposed model on this dataset demonstrates its supremacy over alternative methods depending on initializing the arbitrary load. The primary contributions of the paper can be outlined in the following manner:

- Formulating the classification task as a guessing game within an RL framework, treating it as a Markov decision process.
- Using DE to find an optimal weight initialization for the ANN, initiating the guessing game.
- Rewarding correct and incorrect decisions positively, addressing dataset imbalance by giving higher rewards to the minority class.
- Demonstrating the superiority of the proposed model over alternative approaches that rely on random weight initialization through its performance on the dataset.

The organization of this paper is outlined below: Section II reviews relevant literature, Section III delves into the DE algorithm, and Section IV describes the proposed model. Results and their analysis are discussed in Section V. The paper concludes with a summary in Section VI, along with recommendations for future investigations.

II. Related Work

In recent years, the field of medical science has witnessed an unprecedented surge in the application of machine learning techniques to forecast and categorize a plethora of health concerns, with PPD standing out as a significant area of interest [32]. To understand the evolution and progression of these methodologies, a series of pioneering studies have been meticulously evaluated to shed light on the practices adopted and the degree of precision achieved in their PPD classification endeavors [33].

Zhang et al. [34] placed their bet on SVM and FFS-RF, emphasizing these as the most promising tools for PPD prediction. They embarked on a comprehensive longitudinal survey, engaging 508 women as respondents. The Edinburgh Postnatal Depression Scale (EPDS) served as their choice of instrument to gauge PPD risk. Delving further into their work, Zhang et al. [35] opted for EHR datasets, focusing on the detection of PPD in perinatal women. Their findings were intriguing; while logistic regression fortified with L2 regularization emerged as the top contender for data leading up to childbirth, the post-childbirth data saw MLP taking the lead. Jasiya et al. [36] proposed a machine learning system to identify risk factors and prevalence of postpartum depression in Bangladesh. Utilizing modified questions from EPDS and PHQ-2 scales and socio-demographic queries, data from 150 women was analyzed. The most effective model was identified as Random Forest. Amit et al. [37] presented a Gradient Boosting Machine (GBM)-based approach to PPD depression risk using electronic health records from 266,544 UK women between 2000 and 2017. The model assessed socio-demographic and medical variables and was evaluated alongside the standard EPDS questionnaire for improved screening accuracy. Park et al. [38] suggested an evaluation of methods to reduce bias in clinical machine learning models. Health data from the IBM MarketScan Medicaid Database, focusing on females aged 12 to 55 years with a live birth record from 2014 to 2018, was analyzed. The study examined logistic regression, random forest, and extreme gradient boosting models for postpartum depression and mental health service utilization, assessing racial disparities. Bias reduction methods like reweighing and Prejudice Remover were also explored.

Diversifying the landscape, Shin et al. [39] ventured to harness the PRAMS 2012-2013 dataset and the PHQ-2 questionnaire. Their objective was clear: to tap into various

machine learning algorithms and decipher the prevalence of PPD. Their rigorous analysis crowned Random Forest as the algorithm par excellence for PPD prediction. In a similar vein, Andersson et al. [40] crafted multiple machine learning prototypes, drawing data from Swedish hospitals. Their study, vast in its scope, found the Extremely Randomized Trees model to be unmatched in performance.

Dipping into another significant contribution, Tortajada et al. [41] embarked on a study, leveraging data accumulated from hospital settings. Their research, hinging on MLP, showcased an impressive accuracy rate of 81% in PPD prediction. On the other hand, De Choudhury [42] ventured into the realm of digital platforms, conducting a longitudinal online survey. This study examined an array of regression models, paving the way for new insights. In another noteworthy study, Nataranjan et al. [43] showcased a comparative analysis of algorithms like Functional-gradient boosting, Decision-trees, Naive Bayes, and SVM. Their work, rooted in a longitudinally curated dataset, championed Functional-gradient boosting as the superior method. Lastly, Wang et al. [44] married EHR data with machine learning techniques, with their research revealing SVM as the most fitting algorithm for their dataset.

However, while these advancements are commendable, it is essential to recognize the challenges that come with them. The sheer diversity of algorithms means that selecting the optimal one requires rigorous testing, often demanding substantial resources. Moreover, discrepancies in datasets across different studies might lead to varying conclusions, underscoring the need for standardized and universally accepted data collection methods. Additionally, the robustness of these models in real-world scenarios remains a topic of debate, necessitating further in-depth research and validation.

III. Differential Evolution

Differential Evolution (DE) [29] is an optimization algorithm that stems from populations and finds frequent applications in addressing optimization problems. It falls under the umbrella of evolutionary algorithms, drawing inspiration from the natural progression of evolution. DE is widely acknowledged for its straightforwardness and effectiveness in handling optimization problems involving continuous variables. In addition to its broad range of applications, DE has proven valuable in the realm of machine learning, particularly within the domain of training artificial neural networks. An essential aspect of neural network training revolves around weight initialization, a pivotal factor influencing convergence, generalization capabilities, and the capacity to learn intricate patterns. Traditional weight initialization methods, such as random initialization or fixed values, often grapple with the challenge of striking an optimal balance between avoiding vanishing or exploding gradients and achieving efficient learning. DE can be employed to initialize neural network weights by treating weight values as variables to be optimized. The objective is to identify an optimal set of weight values that minimize the objective function, representing network performance or error on a training dataset. Through the strategic reimagining of weight initialization as an optimization quandary, DE can adeptly navigate and investigate weight configurations that serve as a catalyst for bolstering network performance. DE offers several benefits for weight initialization in machine learning [10]:

- Exploration of Solution Space: The DE algorithm facilitates the exploration of the solution space by generating diverse candidate solutions. This is particularly advantageous for weight initialization as it helps to avoid getting stuck in local optima and enables the algorithm to search for better-performing weight configurations.
- Efficient Optimization: The DE algorithm optimizes the weights by iteratively updating them based on the difference between the target and current solutions. This efficient optimization process aids in finding suitable initial weights that can contribute to faster convergence and improved learning algorithm performance.
- Robustness to Noise: The DE algorithm is known for its robustness to noisy fitness evaluations. In weight initialization, this robustness helps to handle uncertainties and variations in the data, leading to more reliable and stable initial weight configurations.
- Flexibility and Adaptability: The DE algorithm allows for flexibility and adaptability in weight initialization. It can be customized to handle specific problem domains or constraints, such as imposing bounds on weight values or incorporating prior knowledge. This adaptability enhances the algorithm's ability to initialize weights suitable for the given learning task.

The primary procedures of DE are as follows:

- Initialization: The method begins with creating a beginning populace of chosen resolutions called "individuals." Each individual represents a potential solution to the optimization problem and is usually represented as a vector of real numbers.
- Mutation: In each iteration of the algorithm, the individuals in the population are subjected to mutation. Mutation is the process of generating new candidate solutions by perturbing existing ones. In DE, the mutation is performed by creating a trial vector for each individual using the difference between two randomly selected individuals from the population.
- Crossover: After mutation, a crossover operation is applied to combine the trial vector with the original individual. Crossover is a process that blends the information from the trial vector and the original individual to create a new candidate solution. The crossover operation in DE is typically performed using a binomial crossover scheme, where each component of the new solution is selected either from the trial vector or the original individual with a certain probability.
- Selection: The new candidate solution produced by crossover is compared with the original individual, and the better one is selected to proceed to the next iteration. The selection process ensures that only the fitter individuals survive and propagate their traits to the next

generation. This step helps in driving the search towards better solutions over time.

- Termination: The algorithm continues to iterate through mutation, crossover, and selection until a termination condition is met. The termination condition can be a maximum number of iterations, reaching a desired level of solution quality, or any other criteria defined by the problem.

IV. Model Architecture

To address our research challenge, we turned to the capabilities of DE for weight initialization and RL for imbalanced classification, particularly because existing models fall short in several aspects. Traditional models often rely on random weight initialization, which can lead to prolonged training times and the possibility of converging to suboptimal solutions. Many existing algorithms struggle with imbalanced datasets, leading to biased predictions that often overlook the minority class.

By using DE for weight initialization, we can ensure a diverse and potentially more optimal starting point for our learning algorithm. This can lead to faster convergence and potentially superior solutions compared to traditional methods.

RL is a type of machine learning wherein an agent learns to decide by taking actions in an environment to maximize a cumulative reward. It has especially apt for imbalanced classification tasks because it can be tailored to place greater emphasis on the minority class by suitably adjusting the reward mechanism. In scenarios where traditional supervised learning faces challenges because of insufficient representative data for all classes, RL can more effectively explore the decision space and devise strategies that prioritize the accurate classification of underrepresented classes. This addresses another critical limitation of many existing models: their inability to adapt to and accurately classify instances from underrepresented categories.

A. Pretraining

Weight initialization is crucial in neural network training, influencing convergence, generalization, and pattern learning. In this article, the DE algorithm treats weights as variables and minimizes the objective function, representing performance or error. DE effectively explores the weight space by iteratively evaluating and updating weight configurations through evolutionary operators. The goal is to refine weights for better convergence, reduced error, and improved generalization. Incorporating DE in pretraining enhances weight initialization, improving overall network performance.

In this article, the power of the DE algorithm is leveraged to initiate the weights of the MLP. The weights are encoded by meticulously arranging them into a vector, representing them within the DE algorithm. It should be noted that finding the most appropriate layout can be intricate, requiring persistent efforts and a multitude of experiments. Undeterred by the complexity, an optimal encoding strategy was devised through extensive trials and refinements. To provide a visual depiction of this process, Fig. 1 is presented, offering a clear illustration of how all the weights and bias terms are meticulously gathered and assembled into a comprehensive vector. This vector, acting as a candidate solution within the DE algorithm, encapsulates the essential components necessary for weight initialization. By organizing these elements thoughtfully and strategically, the stage is set for the DE algorithm to unleash its optimization prowess, guiding the MLP toward enhanced performance and increased learning capability.

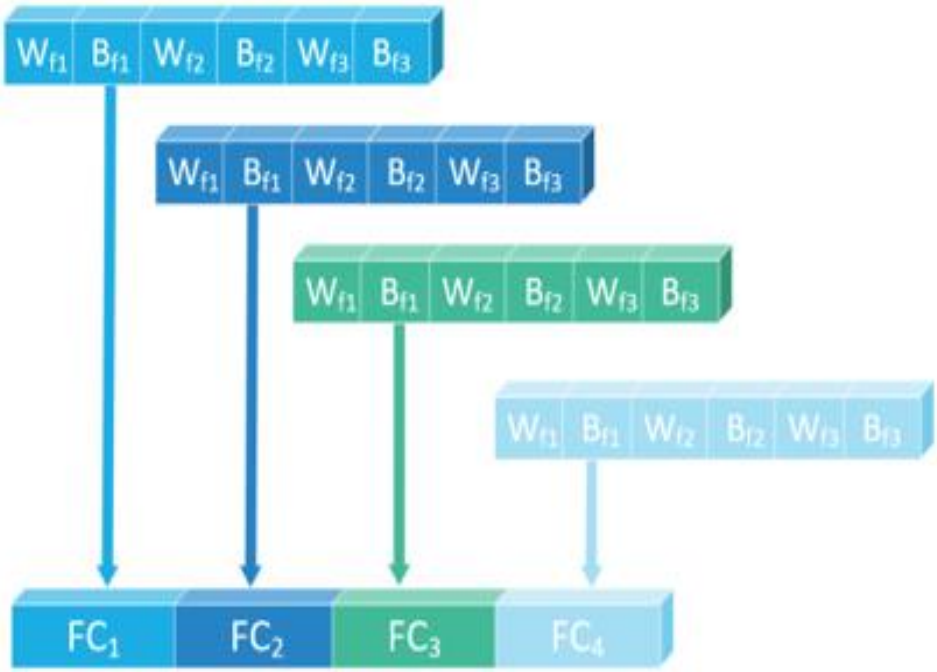


Fig. 1. Encoding strategy used in the proposed algorithm.

The efficacy of a chosen resolution is comprehensively assessed by describing and establishing athleticism performance as a crucial metric. Within the context of the specific problem domain, the quality and performance of the solution are quantified through this function, which serves as a vital tool. The fitness function is meticulously crafted to capture the essential aspects and criteria that govern the success and effectiveness of the candidate solution. The fitness or suitability of the solution can be objectively measured through the careful consideration of relevant factors and parameters, allowing informed decisions to be made and the optimization process to be guided towards optimal outcomes. The fitness function is defined as:

$$Fitness = \frac{1}{\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2} \quad (1)$$

Here, N shows the whole count of train demos, where y_i represents the goal value of the i-th sample and $\tilde{y}_i$ shows the corresponding output predicted by the model.

B. Prediction

To further improve our method of calling the problem of unbalanced classification caused by unequal data volumes in our two classes, we implemented a consecutive policymaking procedure using an RL method. This involved training an ANN model to act as an agent, making informed classifications for each instance, and effectively handling the challenges associated with imbalanced datasets. In the sequential decision-making process, each instance in the train dataset represented a distinct habitat state. The ANN model, acting as the operator, made a categorization sequence for every instance. Simultaneously, as the operator predicted the category name, for instance, it took an act denoted as a_t. At each time-step t, the agent observed an instance representing the current state of the environment, labeled as s_t. The environment provided a reward, r_t, in response to the agent's actions, aiming to guide its behavior. To address the class imbalance issue, the reward values were carefully crafted. Samples from the majority class received lower absolute reward values, while relatively higher absolute reward values were assigned to samples from the minority class. This reward design aimed to encourage the agent to prioritize the correct classification of minority class samples, contributing to mitigating the impact of imbalanced data. In this article, the reward function is defined as:

$$r_t(s_t, a_t, y_t) = \begin{cases} +1\,, a_t = y_t \; and \; s_t \in D_S \\ -1\,, a_t \neq y_t \; and \; s_t \in D_S \\ \lambda\,, a_t = y_t \; and \; s_t \in D_H \\ -\lambda\,, a_t \neq y_t \; and \; s_t \in D_H \end{cases} \quad (2)$$

where D_S, and D_H represent the minority and majority categories in order. Incorrectly/correctly categorizing a demo of the major category gains an award of $+\lambda/-\lambda$, where $0 < \lambda < 1$. We aimed to incentivize the agent to give greater attention to the minority class and mitigate the bias caused by imbalanced data by providing differential rewards based on class distribution. Through this reinforcement learning approach, the agent learned an optimal classification strategy that considered both the inherent difficulty of classifying the minority class and the importance of accurate predictions overall. During training, the agent continuously refined its decision-making capabilities and updated its policies and strategies based on the rewards received. By leveraging reinforcement learning techniques, we aimed to achieve a more balanced and effective classification performance, particularly for the underrepresented class. This novel sequential decision-making process enabled us to overcome the limitations imposed by imbalanced datasets and successfully address the challenges of imbalanced classification. As a result, we achieved improved accuracy and fairness in predictions by combining the power of artificial neural networks and reinforcement learning.

V. Experimental Results

A. Data Sources

The data used for developing the prediction models were acquired from the "Biology, Affect, Stress, Imaging and Cognition during Pregnancy and the Puerperium" (BASIC) study [45]. BASIC is a prospective cohort study conducted at the Department of Obstetrics and Gynaecology in Uppsala University Hospital, Uppsala, Sweden, and it involves a population-based approach. Between September 2009 and November 2018, pregnant women who fulfilled specific eligibility criteria were invited to take part in the study. The criteria included being 18 years of age or older, not having concealed identities, possessing sufficient proficiency in reading and comprehending Swedish, and not having been diagnosed with bloodborne infections or non-viable pregnancies based on routine ultrasound examinations. In the BASIC study, data collection primarily relied on online surveys and questionnaires administered to women at various stages: during pregnancy at the 17th and 32nd week of gestation, as well as at 6 weeks, 6 months, and 12 months after giving birth. These surveys and questionnaires were designed to gather information from participants during these specific time points. The surveys consisted of inquiries regarding various background characteristics, encompassing sociodemographic variables, psychological assessments, medical details, reproductive history, lifestyle factors, and sleep patterns. All questionnaires were completed by the participants themselves and were conducted online. Information was additionally sourced from medical journals. The study had a participation rate of 20%, but the cohort experienced a comparatively low dropout rate, as 71% of the participants remained in the study during the 12-month follow-up period. The study obtained approval from the Research Ethics Board in Uppsala (Dnr 2009/171, with amendments). Prior to their inclusion in the study, all participating women provided written informed consent. The research methods adhered to applicable guidelines and regulations.

B. Model Performance

Our project utilizes a 64-bit Windows operating system, complemented with 64 GB of RAM and a 64 GB GPU. Table I presents the hyperparameters applied to the proposed model.

TABLE I. HYPERPARAMETER SETTING FOR THE PROPOSED MODEL

Hyperparameter	Value
Epoch	256
Batch size	64
Learning rate	0.02
Dropout rate	0.5
Discount factor	0.3

TABLE II. HYPERPARAMETER SETTING FOR MACHINE LEARNING METHODS

Algorithm	Parameter	Value
Naïve Bayes	α (Lidstone smoothing parameter)	0.5
KNN	*k* (Number of neighbors)	5
	Distance Metric	Euclidean (p=2), Manhattan (p=1)
SVM	Kernel	polynomial
	γ (Kernel coefficient)	0.5
Random Forests	Number of trees	20
	Max depth of tree	10
Logistic Regression	C (Inverse regularization strength)	0.3
	Solver	liblinear
Decision Tree	Criterion	entropy
	Max depth	10

In the evaluation process, the suggested model underwent rigorous comparison with six distinct machine learning models: Naïve Bayes [46], KNN [47], SVM [48], Random forests [49], Logistic Regression [50], and Decision tree [51]. In the evaluation process, the suggested model underwent rigorous comparison with six distinct machine learning models: Naïve Bayes [46], KNN [47], SVM [48], Random forests [49], Logistic Regression [50], and Decision tree [51]. Table II shows the parameters applied to these models.

Additionally, two modified versions were included in the analysis to explore different variations of the proposed model. The first modified version, proposed+random weights, adopted a similar foundational architecture to our model but employed random weights for initialization. This alternative initialization method allowed for a comparative investigation of the impact of weight initialization on the performance of the model. The second modified version, Proposed+random weights+RL, incorporated RL techniques for classification. This integration of RL aimed to enhance the ability of the model to make accurate predictions and improve its overall performance. Standard metrics were employed to assess these models' performance, with particular emphasis on the geometry average and F-measure due to their suitability for unbalanced info [52]. The results, which can be found in Table III, clearly demonstrate the superiority of the proposed model over all other models, including the previously recognized top performer, Decision Tree. The evaluation results are shown schematically in Fig. 2 to understand the results better. Across all evaluation criteria, the proposed model consistently outperformed its counterparts. Notably, the proposed model achieved remarkable error reductions over 65%; plus, 29% in the G-averages and F-measure metrics, orderly. These substantial improvements illustrate the effectiveness of the proposed model in tackling the challenges posed by imbalanced data and its ability to generate more accurate predictions. Contrasting the offered sample with the modified versions, Offered+arbitrary loads+RL and Offered+arbitrary loads, the significance of the integration of DE and RL approaches becomes apparent. Our model demonstrated an impressive decrease in the error rate of approximately 62% when compared to these modified versions. This finding underscores the critical role played by DE and RL in enhancing the model's performance and highlights their importance in developing state-of-the-art machine learning models.

TABLE III. OUTCOMES OF SEVERAL CATEGORIZATION METHODS

	accuracy	recall	precision	F-measure	G-means
Naïve Bayes	0.6804±0.0501	0.5909±0.0653	0.5353±0.1206	0.5601±0.0652	0.6562±0.0052
KNN	0.8101±0.1553	0.7351±0.0043	0.7653±0.1002	0.7471±0.1054	0.8002±0.0402
SVM	0.7803±0.1005	0.6603±0.0054	0.6903±0.0452	0.6702±0.2103	0.7503±0.0051
Random forests	0.6901±0.1006	0.5602±0.2202	0.5453±0.1001	0.5502±0.0953	0.6505±0.2502
Logistic Regression	0.8105±0.1404	0.7706±0.0456	0.7005±0.1003	0.7302±0.1404	0.8006±0.1201
Decision tree	0.8304±0.1405	0.8103±0.1006	0.7502±0.2704	0.7902±0.2001	0.8202±0.0104
Proposed+random weights	0.8104 ± 0.0627	0.8202 ± 0.1108	0.8013 ± 0.0109	0.7918 ± 0.1623	0.8303 ± 0.2622
Proposed+random weights+RL	0.8615 ± 0.0243	0.8704 ± 0.1256	0.8509 ± 0.2691	0.8506 ± 0.0517	0.8609 ± 0.0921
Proposed	0.8907 ± 0.0384	0.9053 ± 0.1132	0.8847 ± 0.0315	0.8844 ± 0.0297	5.0423

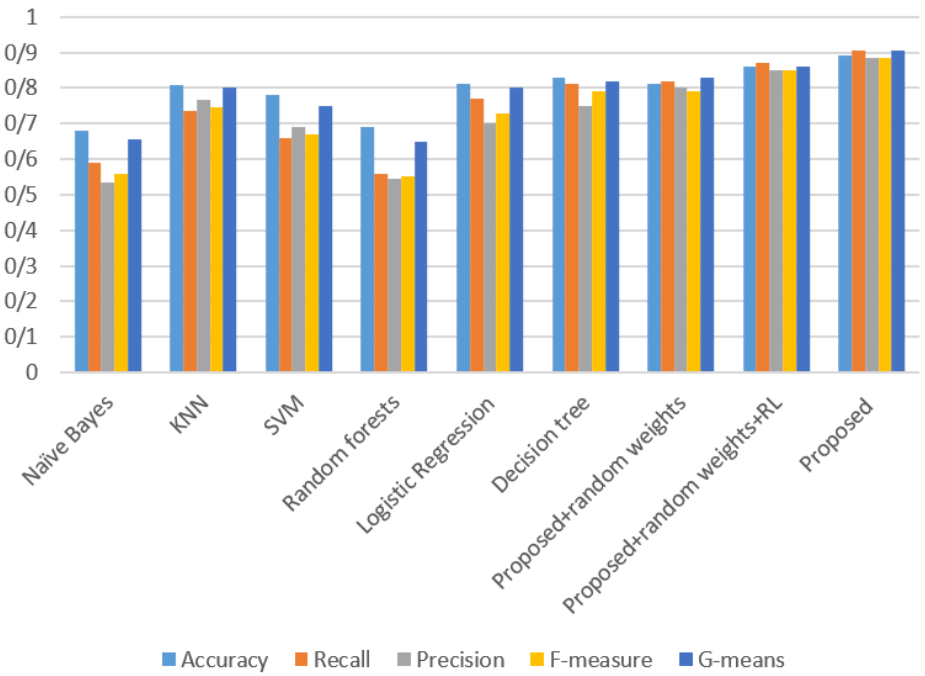


Fig. 2. Graphical comparison of various classification algorithms.

TABLE IV. HYPERPARAMETER SETTING FOR METAHEURISTIC ALGORITHMS

Algorithm	Parameter	Value
DE	scaling factor	0.5
	crossover probability	0.7
ABC	limit	n_e × dimensionality
	n_o	50% of the colony
	n_e	50% of the colony
	n_s	1
FA	light absorption coefficient	1
	attractiveness at r = 0	0.2
	scaling factor	0.25
BA	constant for loudness update	0.5
	constant for an emission rate update	0.5
	initial pulse emission rate	0.001
COA	discovery rate of alien solutions	0.25

TABLE V. OUTCOMES OF SEVERAL METAHEURISTIC METHODS

	accuracy	recall	precision	F-measure	G-means
Proposed + ABC + RL	0.8510 ± 0.1470	0.8390 ± 0.1220	0.8570 ± 0.0500	0.8400 ± 0.0080	0.8190 ± 0.4110
Proposed + GWO+ RL	0.8400 ± 0.1560	0.8290 ± 0.1010	0.8400 ± 0.2490	0.8280 ± 0.0140	0.7980 ± 0.0230
Proposed + FA+ RL	0.8250 ± 0.0020	0.8130 ± 0.1210	0.8090 ± 0.2600	0.8170 ± 0.0630	0.7780 ± 0.0000
Proposed + BA+ RL	0.8080 ± 0.0120	0.7980 ± 0.0040	0.8000 ± 0.0590	0.8030 ± 0.1430	0.7600 ± 0.1180
Proposed + COA+ RL	0.7900 ± 0.1570	0.7700 ± 0.0120	0.7840 ± 0.2590	0.7700 ± 0.1640	0.7390 ± 0.1480

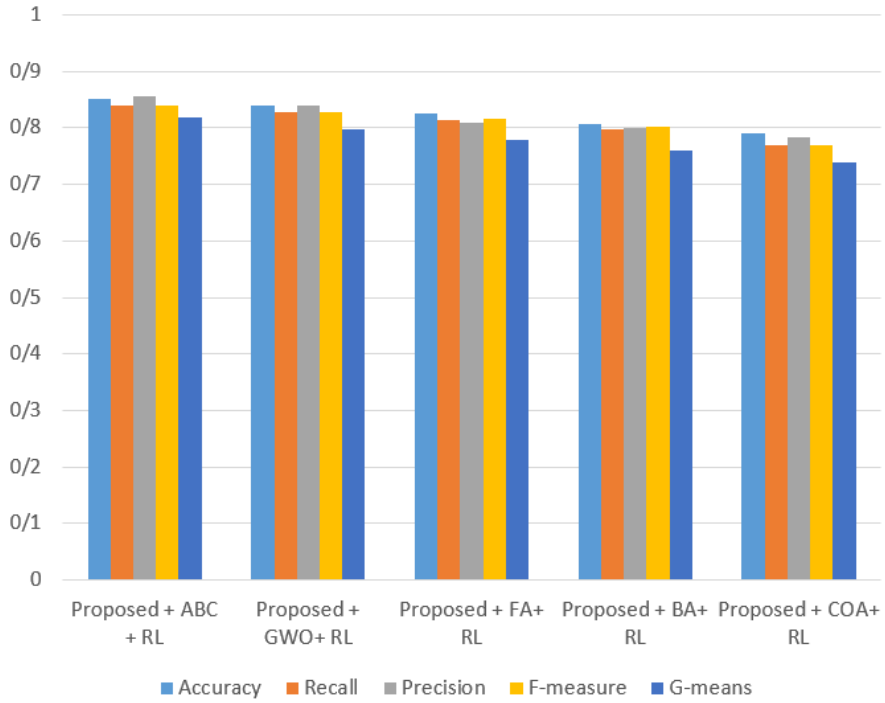


Fig. 3. Graphical comparison of various classification algorithms.

A detailed analysis was conducted in the subsequent experiment to compare the DE algorithm with various well-established metaheuristic optimization algorithms. To ensure a fair comparison, different metaheuristics were employed to derive the initial weights while keeping the remaining components of the model consistent. The evaluation encompassed six distinct algorithms, namely ABC [53], GWO [54], FA [55], BA [56], and COA [57]. For every algorithm, both the population size and the count of function evaluations are configured to 200 and 3,000, respectively. The default configurations are detailed in Table IV. The results of this comprehensive experiment were systematically presented in Table V and Fig. 3, providing valuable insights into the performance of each algorithm. Notably, the findings highlighted the remarkable achievement of DE, which demonstrated a significant reduction in error of approximately 52% when compared to the ABC algorithm. Furthermore, the DE algorithm outperformed other well-known algorithms, including GWO and BA. This outcome solidified the position of the DE algorithm as a leading contender among the considered metaheuristic optimization approaches.

C. Award Operation Effect

The rewards given to the majority and minority classes for correct and incorrect classifications are +1 and ±λ, in order. The λ value is determined by the scale of major to minor demos, and by increasing this scale, the ideal value of λ is expected to decrease. For researching the λ effect, we assessed the offered demo's efficiency through various λ on a scale of 0 to 1 (in increments of 0.1) as saving the bonus for the major category. The outcomes are represented in Fig. 4. By the time the λ is 0, the major category's effect gets insignificant, while at λ = 1, both categories have the same effects. The findings reveal the model performs optimally when λ is set to 0.4 for every measured criterion, recommending that the optimum λ value is on the scale of 0 to 1. We should notice that when it is crucial to diminish the major category's effect by adjusting λ, adjusting it to a lower level might have a detrimental effect on the overall model performance. The results indicate that the choice of λ has a substantial impact on the performance of the model. The optimal value of λ depends on the relative proportions of the majority and minority samples, underscoring the significance of careful selection to achieve the best possible outcomes.

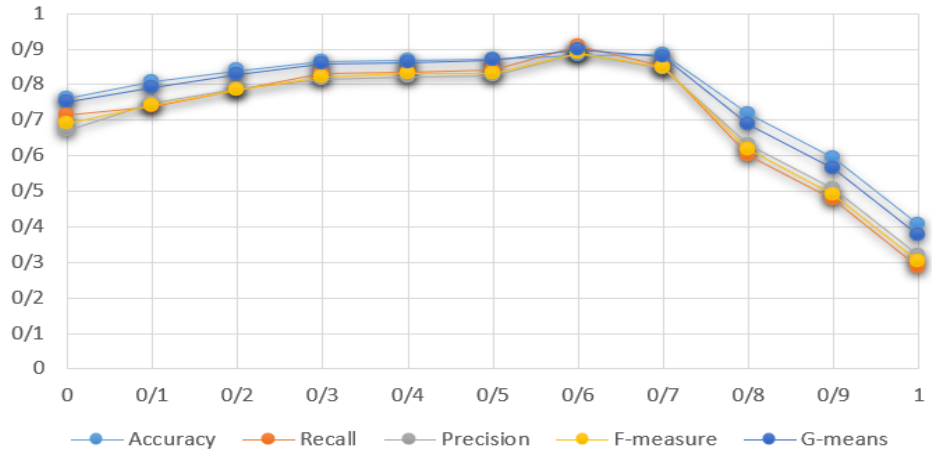


Fig. 4. Visual depiction showcasing the alteration in performance parameters caused by fluctuations in the value of λ.

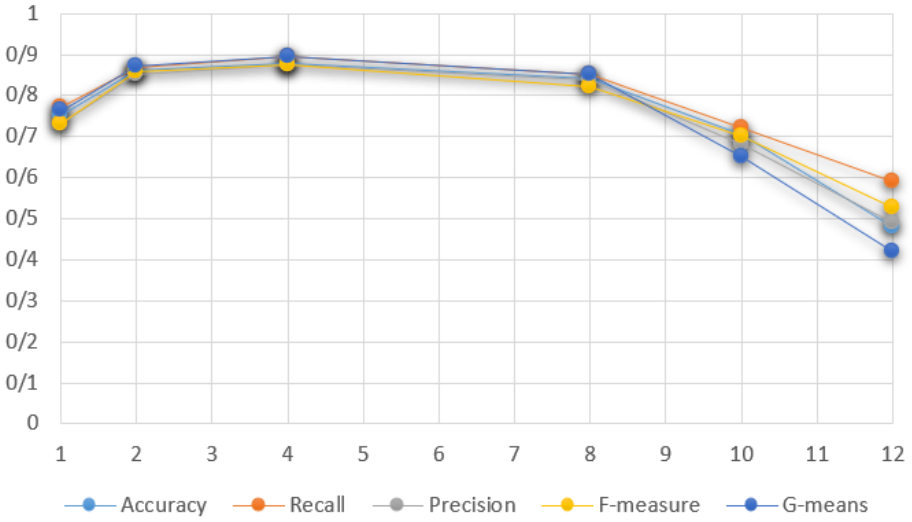


Fig. 5. The plotted performance metrics as a function of the MLP layers.

D. Impact of the MLP Layers

The article emphasizes that increasing the number of layers in an MLP leads to a higher model complexity, which in turn increases the risk of overfitting. On the other hand, having too few layers may limit the ability of the model to capture important features in the training data. In our proposed approach, we conducted experiments with six different values (1, 2, 4, 8, 10, 12) for the number of layers in the MLP to examine its impact on model performance. The results, presented in Fig. 5, demonstrate a decreasing trend in performance as the number of layers is in the 1 to 4 range, next to a rising trend for 4 to 12 values. This suggests that having four layers in the MLP yields optimal performance and achieves the best results.

E. Impact of the Loss Function

Various techniques are available to tackle data imbalances in machine learning models, including adjusting data augmentation methods and selecting an appropriate loss function. Among these techniques, the choice of loss function plays a crucial role in enabling the model to learn from the minority class effectively. To assess the efficacy of different loss functions, we examined five specific functions: WCE [58], BCE [59], DL [60], TL [61], and CL [62]. BCE and WCE commonly use loss functions that equally treat positive and negative examples. However, in the case of imbalanced datasets where the emphasis needs to be placed on the minority class, these loss functions may not be suitable.

On the other hand, DL and TL loss functions are better suited for imbalanced datasets as they yield improved performance ojn the minority class. As a promising loss function, CL is particularly beneficial for applications involving unbalanced data. By adjusting the weights of the loss function, CL can assign lower importance to simple examples and focus more on learning complex samples. To evaluate the effectiveness of these loss functions, we conducted experiments and presented the results in Table 6 and Fig. 6. The findings demonstrate that the CL function surpasses the TL function, resulting in a 25% reduction in the error rate for the accuracy metric and a 39% reduction for the F-measure metric. However, it is worth noting that the CL function performs 60% worse than the FL function, which is a specialized loss function specifically designed for binary classification tasks. It is important to consider these results in

the context of the specific problem at hand and the nature of the dataset. While the CL function outperforms the TL function, it falls short when compared to the FL function. Further investigation is required to understand the factors contributing to these differences in performance and to explore the potential of customized loss functions specifically tailored to address the challenges posed by imbalanced datasets. Additionally, research can focus on developing novel loss functions or adapting existing ones to strike a balance between emphasizing the minority class and maintaining overall classification accuracy across various classification tasks and datasets.

TABLE VI. RESULTS FROM VARIOUS LFS

	accuracy	recall	precision	F-measure	G-means
WCE	0.7432± 0.1132	0.7200± 0.0100	0.7174± 0.0155	0.7215± 0.0130	0.7510± 0.1255
BCE	0.8001± 0.1021	0.7830± 0.1145	0.7520± 0.1051	0.7622± 0.0102	0.8025± 0.0101
DL	0.8042± 0.2033	0.8151± 0.0025	0.7936± 0.0106	0.7992± 0.0106	0.826± 0.1000
TL	0.8320± 0.1203	0.8252± 0.0006	0.8014± 0.0216	0.8148± 0.0436	0.8450± 0.2062
CL	0.8630± 0.0247	0.8545± 0.2045	0.8415± 0.0148	0.8440± 0.0152	0.8639± 0.0152

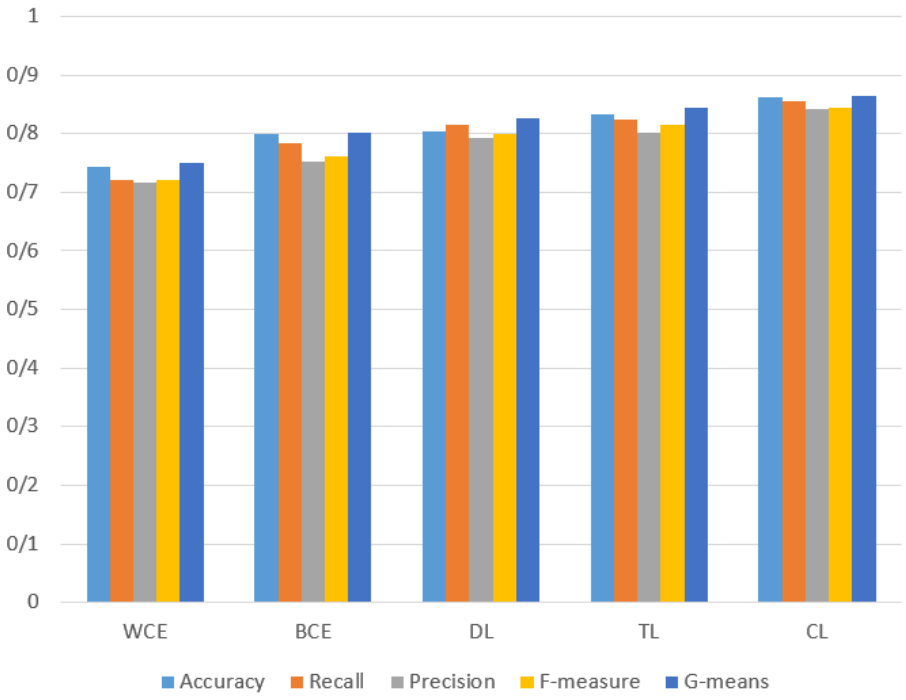


Fig. 6. Graphical comparison of various loss functions.

F. Discussion

The findings presented in this study have important implications for PPD identification and intervention. Developing an automated model using a deep reinforcement learning approach and a DE algorithm for weight initialization shows the potency of advanced ML approaches for addressing the challenges associated with PPD identification. One of the key advantages of using the DE algorithm for weight initialization is its ability to explore high-dimensional spaces and find optimal weight values effectively. This ensures that the model is initialized in a manner that enables it to make accurate predictions and classify PPD effectively. By leveraging an ANN and treating the categorization issue as a policymaking stage-by-stage process, the sample considers the complexity and nuances of PPD, enhancing its predictive capabilities. Using an encouraging learning system and a particular compensatory technique further enhances the model's performance. By assigning a higher reward for identifying the minority class, the model is incentivized to focus on precise detection, addressing the challenge of identifying at-risk individuals. This approach acknowledges the importance of early identification to provide timely interventions and support to those most in need. The evaluation of the model's performance using a comprehensive dataset acquired from the population-based BASIC study in Uppsala, Sweden, strengthens the validity of the findings. With a large sample size of 4313 samples spanning a significant period, the study provides robust evidence of the high accuracy of the model in identifying PPD. This accuracy underscores the potential effectiveness of the model in real-world applications for identifying at-risk women. The implications of this research are far-reaching. By accurately

identifying women at risk for PPD, healthcare professionals can provide timely interventions and support, thus improving maternal and child health outcomes. Preventative interventions targeted at high-risk individuals have been shown to be more cost-effective, making the automated model a valuable tool in resource allocation and optimizing healthcare services.

To offer a more in-depth assessment of our model's capabilities, we reached out for external expert opinions. By teaming up with seasoned professionals, possessing extensive experience in the field, we embarked on a comprehensive qualitative review of the model's performance. These experts, hailing from diverse backgrounds and having a rich tapestry of experiences in similar research areas, thoroughly scrutinized the model's underpinnings, methodologies, and outcomes. Their rigorous evaluations and constructive feedback painted a clear picture. Their collective insights resoundingly echoed our preliminary findings, particularly highlighting the model's unparalleled precision, steadfast reliability, and robust adaptability. When our model was placed side by side with pre-existing algorithms for a comparative analysis, it distinctly stood out, showcasing its superior design and performance. The external validation from such esteemed professionals not only fortified our confidence in the model but also underscored its potential for real-world applications and future research endeavors.

However, it is important to acknowledge the limitations of this study and consider avenues for future research. Firstly, the dataset used in this study was acquired from a specific population-based study conducted in Uppsala, Sweden. While this provides valuable insights into the model's performance within that particular context, it raises questions about the generalizability of the findings to diverse populations and settings [63]. Variations in cultural, socioeconomic, and healthcare factors may influence the prevalence and presentation of PPD, potentially impacting the model's performance [64]. Therefore, future studies should aim to validate the model using datasets from different regions and populations to ensure its applicability across various contexts [65].

Additionally, while the model demonstrates high accuracy in identifying PPD, assessing its performance in real-world clinical arrangements is crucial [66]. The controlled environment of the study may not fully reflect the complexities and challenges faced by healthcare professionals in their daily practice.

Evaluating the model's effectiveness in a clinical setting, where multiple factors can influence the identification and treatment of PPD, would provide valuable insights into its practical utility. Longitudinal studies tracking patient outcomes and the impact of the model's predictions on treatment decisions and health outcomes would further enhance our understanding of its clinical relevance. Furthermore, expanding the scope of research beyond the model accuracy is essential. While accuracy is a crucial metric, evaluating other performance measures such as sensitivity, specificity, positive predictive value, and negative predictive value is equally important [67]. These metrics provide a more comprehensive assessment of the model's diagnostic capabilities and ability to identify individuals at risk and those not at risk for PPD. Understanding the model performance across these measures can guide healthcare professionals in effectively utilizing its predictions and making informed interventions and resource allocation decisions.

Moreover, assessing the impact of the model on patient outcomes is a critical aspect that requires further investigation [68]. While timely identification of at-risk women is essential, evaluating whether the interventions based on the model predictions lead to improved maternal and child health outcomes is equally vital [69]. Conducting studies that measure the effectiveness of interventions guided by the model, such as targeted support programs or personalized treatment plans, would provide valuable evidence of the model's potential to impact patient outcomes positively.

Finally, it is important to consider the ethical implications and potential challenges associated with the implementation of an automated model for PPD identification [70]. Issues such as privacy, data security, and the potential for biases in the model's predictions need to be thoroughly examined and addressed. Ensuring transparency, fairness, and accountability in developing and deploying such models is essential to maintaining trust among healthcare professionals and the wider public [71].

VI. CONCLUSION

In this study, we have developed an automated model to identify PPD using a deep reinforcement learning approach combined with a DE algorithm for weight initialization. The DE algorithm is renowned for its ability to effectively explore high-dimensional spaces and find optimal weight values, making it well-suited for weight initialization in our model. Our approach utilizes an ANN and treats the PPD classification problem as a policymaking stage-by-stage process. At every stage, the operator acquires samples and employs classifications, while the habitat maintains rewards for every classifying activity. For inspiring precise detection, a greater award is determined for recognizing the minor category. Through a particular compensatory technique and an encouraging learning system, the operator learns and chooses the most effective method for achieving the goals. To evaluate our sample's efficiency, we analyzed a comprehensive set of data obtained from the population-based BASIC study conducted in Uppsala, Sweden, spanning from 2009 to 2018, and comprising 4313 samples. The experiential results were assessed by known analysis criteria, revealing our sample achieved greater precision and correctness, demonstrating its suitability for identifying PPD. These findings carry significant implications for identifying at-risk women and providing timely interventions to improve maternal and child health outcomes.

Reinforcement learning algorithms often face the challenge of striking a balance between exploration and exploitation. Future research can delve deeper into exploring effective strategies for addressing this trade-off in the context of PPD identification. Techniques such as adaptive exploration policies, multi-objective optimization, or

incorporating domain knowledge can help optimize the model's performance in identifying at-risk women while minimizing false positives and negatives. Moreover, investigating the potential of transfer learning and domain adaptation techniques can contribute to improving the generalization capabilities of the PPD identification model. By leveraging knowledge gained from related domains or pre-trained samples, the sample's efficiency can be enhanced when applied to different populations, cultures, or healthcare settings. This research direction can help address the challenges of model generalizability and make the automated PPD identification model more robust.

REFERENCES

[1] L. J. Miller, "Postpartum depression," Jama, vol. 287, no. 6, pp. 762-765, 2002.

[2] I. S. Yim, L. R. Tanner Stapleton, C. M. Guardino, J. Hahn-Holbrook, and C. Dunkel Schetter, "Biological and psychosocial predictors of postpartum depression: systematic review and call for integration," Annual review of clinical psychology, vol. 11, pp. 99-137, 2015.

[3] M. Bloch, R. C. Daly, and D. R. Rubinow, "Endocrine factors in the etiology of postpartum depression," Comprehensive psychiatry, vol. 44, no. 3, pp. 234-246, 2003.

[4] D. K. Sit and K. L. Wisner, "The identification of postpartum depression," Clinical obstetrics and gynecology, vol. 52, no. 3, p. 456, 2009.

[5] S. Asif et al., "Severe obstetric lacerations associated with postpartum depression among women with low resilience–a Swedish birth cohort study," BJOG: An International Journal of Obstetrics & Gynaecology, vol. 127, no. 11, pp. 1382-1390, 2020.

[6] E. Y. Wan, C. A. Moyer, S. D. Harlow, Z. Fan, Y. Jie, and H. Yang, "Postpartum depression and traditional postpartum care in China: role of zuoyuezi," International Journal of Gynecology & Obstetrics, vol. 104, no. 3, pp. 209-213, 2009.

[7] P. Tong, L.-P. Dong, Y. Yang, Y.-H. Shi, T. Sun, and P. Bo, "Traditional Chinese acupuncture and postpartum depression: a systematic review and meta-analysis," Journal of the Chinese Medical Association, vol. 82, no. 9, pp. 719-726, 2019.

[8] Y. Lee et al., "Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review," Journal of affective disorders, vol. 241, pp. 519-532, 2018.

[9] S. V. Moravvej et al., "RLMD-PA: A reinforcement learning-based myocarditis diagnosis combined with a population-based algorithm for pretraining weights," Contrast Media & Molecular Imaging, vol. 2022, 2022.

[10] S. V. Moravvej, S. J. Mousavirad, D. Oliva, G. Schaefer, and Z. Sobhaninia, "An improved de algorithm to optimise the learning process of a bert-based plagiarism detection model," in 2022 IEEE Congress on Evolutionary Computation (CEC), 2022: IEEE, pp. 1-7.

[11] S. Zhang, C. Tjortjis, X. Zeng, H. Qiao, I. Buchan, and J. Keane, "Comparing Data Mining Methods with Logistic Regression."

[12] S. Danaei et al., "Myocarditis Diagnosis: A Method using Mutual Learning-Based ABC and Reinforcement Learning," in 2022 IEEE 22nd International Symposium on Computational Intelligence and Informatics and 8th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics (CINTI-MACRo), 2022: IEEE, pp. 000265-000270.

[13] H. Zareiamand, A. Darroudi, I. Mohammadi, S. V. Moravvej, S. Danaei, and R. Alizadehsani, "Cardiac Magnetic Resonance Imaging (CMRI) Applications in Patients with Chest Pain in the Emergency Department: A Narrative Review," Diagnostics, vol. 13, no. 16, p. 2667, 2023.

[14] I. Mani and I. Zhang, "kNN approach to unbalanced data distributions: a case study involving information extraction," in Proceedings of workshop on learning from imbalanced datasets, 2003, vol. 126: ICML, pp. 1-7.

[15] S. V. Moravvej, M. J. M. Kahaki, M. S. Sartakhti, and A. Mirzaei, "A method based on attention mechanism using bidirectional long-short term memory (BLSTM) for question answering," in 2021 29th Iranian Conference on Electrical Engineering (ICEE), 2021: IEEE, pp. 460-464.

[16] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1, 2005: Springer, pp. 878-887.

[17] L. Hong et al., "GAN-LSTM-3D: An efficient method for lung tumour 3D reconstruction enhanced by attention-based LSTM," CAAI Transactions on Intelligence Technology, 2023.

[18] M. Soleimani, Z. Forouzanfar, M. Soltani, and M. J. Harandi, "Imbalanced Multiclass Medical Data Classification based on Learning Automata and Neural Network," EAI Endorsed Transactions on AI and Robotics, vol. 2, 2023.

[19] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5375-5384.

[20] S. Moravvej, M. Maleki Kahaki, M. Salimi Sartakhti, and M. Joodaki, "Efficient GAN-based method for extractive summarization," Journal of Electrical and Computer Engineering Innovations (JECEI), vol. 10, no. 2, pp. 287-298, 2022.

[21] S. V. Moravvej, S. J. Mousavirad, M. H. Moghadam, and M. Saadatmand, "An LSTM-based plagiarism detection via attention mechanism and a population-based approach for pre-training parameters with imbalanced classes," in Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part III 28, 2021: Springer, pp. 690-701.

[22] M. Bahadori, M. Soltani, M. Soleimani, and M. Bahadori, "Statistical Modeling in Healthcare: Shaping the Future of Medical Research and Healthcare Delivery," in AI and IoT-Based Technologies for Precision Medicine: IGI Global, 2023, pp. 431-446.

[23] M. Soltani, M. Bahadori, and M. Soleimani, "Optimal Predictive Maintenance and Spare Part Inventory Policies for a Degrading System Subjected to Imperfect Actions," Available at SSRN 4558570.

[24] S. V. Moravvej, M. Joodaki, M. J. M. Kahaki, and M. S. Sartakhti, "A method based on an attention mechanism to measure the similarity of two sentences," in 2021 7th International Conference on Web Research (ICWR), 2021: IEEE, pp. 238-242.

[25] S. V. Moravvej, A. Mirzaei, and M. Safayani, "Biomedical text summarization using conditional generative adversarial network (CGAN)," arXiv preprint arXiv:2110.11870, 2021.

[26] M. Jaderberg et al., "Population based training of neural networks," arXiv preprint arXiv:1711.09846, 2017.

[27] S. J. Mousavirad, D. Oliva, S. Hinojosa, and G. Schaefer, "Differential evolution-based neural network training incorporating a centroid-based strategy and dynamic opposition-based learning," in 2021 IEEE Congress on Evolutionary Computation (CEC), 2021: IEEE, pp. 1233-1240.

[28] D. Karaboga, B. Akay, and C. Ozturk, "Artificial bee colony (ABC) optimization algorithm for training feed-forward neural networks," in Modeling Decisions for Artificial Intelligence: 4th International Conference, MDAI 2007, Kitakyushu, Japan, August 16-18, 2007. Proceedings 4, 2007: Springer, pp. 318-329.

[29] S. V. Moravvej, S. J. Mousavirad, D. Oliva, and F. Mohammadi, "A Novel Plagiarism Detection Approach Combining BERT-based Word Embedding, Attention-based LSTMs and an Improved Differential Evolution Algorithm," arXiv preprint arXiv:2305.02374, 2023.

[30] S. Vakilian, S. V. Moravvej, and A. Fanian, "Using the artificial bee colony (ABC) algorithm in collaboration with the fog nodes in the Internet of Things three-layer architecture," in 2021 29th Iranian Conference on Electrical Engineering (ICEE), 2021: IEEE, pp. 509-513.

[31] M. S. Sartakhti, M. J. M. Kahaki, S. V. Moravvej, M. javadi Joortani, and A. Bagheri, "Persian language model based on BiLSTM model on COVID-19 corpus," in 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA), 2021: IEEE, pp. 1-5.

[32] S. R. Low, S. A. Bono, and Z. Azmi, "The effect of emotional support on postpartum depression among postpartum mothers in Asia: A systematic review," Asia-Pacific Psychiatry, p. e12528, 2023.

[33] K. M. Deligiannidis et al., "Zuranolone for the treatment of postpartum depression," American Journal of Psychiatry, vol. 180, no. 9, pp. 668-675, 2023.

[34] W. Zhang, H. Liu, V. M. B. Silenzio, P. Qiu, and W. Gong, "Machine learning models for the prediction of postpartum depression: application and comparison based on a cohort study," JMIR medical informatics, vol. 8, no. 4, p. e15516, 2020.

[35] Y. Zhang, S. Wang, A. Hermann, R. Joly, and J. Pathak, "Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women," Journal of affective disorders, vol. 279, pp. 1-8, 2021.

[36] J. F. Raisa, M. S. Kaiser, and M. Mahmud, "A machine learning approach for early detection of postpartum depression in Bangladesh," in International Conference on Brain Informatics, 2022: Springer, pp. 241-252.

[37] G. Amit et al., "Estimation of postpartum depression risk from electronic health records using machine learning," BMC Pregnancy and Childbirth, vol. 21, no. 1, pp. 1-10, 2021.

[38] Y. Park et al., "Comparison of methods to reduce bias from clinical prediction models of postpartum depression," JAMA network open, vol. 4, no. 4, pp. e213909-e213909, 2021.

[39] D. Shin, K. J. Lee, T. Adeluwa, and J. Hur, "Machine learning-based predictive modeling of postpartum depression," Journal of Clinical Medicine, vol. 9, no. 9, p. 2899, 2020.

[40] S. Andersson, D. R. Bathula, S. I. Iliadis, M. Walter, and A. Skalkidou, "Predicting women with depressive symptoms postpartum with machine learning methods," Scientific reports, vol. 11, no. 1, p. 7877, 2021.

[41] S. Tortajada et al., "Prediction of postpartum depression using multilayer perceptrons and pruning," Methods of information in medicine, vol. 48, no. 03, pp. 291-298, 2009.

[42] !!! INVALID CITATION !!! {}.

[43] S. Natarajan, A. Prabhakar, N. Ramanan, A. Bagilone, K. Siek, and K. Connelly, "Boosting for postpartum depression prediction," in 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2017: IEEE, pp. 232-240.

[44] K. L. Wisner, E. L. Moses-Kolko, and D. K. Sit, "Postpartum depression: a disorder in search of a definition," Archives of women's mental health, vol. 13, pp. 37-40, 2010.

[45] S. Andersson, D. R. Bathula, S. I. Iliadis, M. Walter, and A. Skalkidou, "Predicting women with depressive symptoms postpartum with machine learning methods," Scientific reports, vol. 11, no. 1, pp. 1-15, 2021.

[46] G. I. Webb, E. Keogh, and R. Miikkulainen, "Naïve Bayes," Encyclopedia of machine learning, vol. 15, pp. 713-714, 2010.

[47] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings, 2003: Springer, pp. 986-996.

[48] M. A. de Almeida, "DATA MINING: DETERMINAC AO DE AGRUPAMENTOS EM GRANDES BASES DE DADOS," Universidade Federal do Rio de Janeiro, 2013.

[49] L. Breiman, "Random forests," Machine learning, vol. 45, pp. 5-32, 2001.

[50] M. P. LaValley, "Logistic regression," Circulation, vol. 117, no. 18, pp. 2395-2399, 2008.

[51] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," Journal of Chemometrics: A Journal of the Chemometrics Society, vol. 18, no. 6, pp. 275-285, 2004.

[52] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," J Inf Eng Appl, vol. 3, no. 10, 2013.

[53] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," Applied soft computing, vol. 8, no. 1, pp. 687-697, 2008.

[54] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," Advances in engineering software, vol. 69, pp. 46-61, 2014.

[55] X.-S. Yang, "Firefly algorithm, stochastic test functions and design optimisation," International journal of bio-inspired computation, vol. 2, no. 2, pp. 78-84, 2010.

[56] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," Nature inspired cooperative strategies for optimization (NICSO 2010), pp. 65-74, 2010.

[57] S. Vakilian, S. V. Moravvej, and A. Fanian, "Using the cuckoo algorithm to optimizing the response time and energy consumption cost of fog nodes by considering collaboration in the fog layer," in 2021 5th International Conference on Internet of Things and Applications (IoT), 2021: IEEE, pp. 1-5.

[58] Ö. Özdemir and E. B. Sönmez, "Weighted cross-entropy for unbalanced data with application on covid x-ray images," in 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 2020: IEEE, pp. 1-6.

[59] F. Huang, J. Li, and X. Zhu, "Balanced Symmetric Cross Entropy for Large Scale Imbalanced and Noisy Data," arXiv preprint arXiv:2007.01618, 2020.

[60] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced NLP tasks," arXiv preprint arXiv:1911.02855, 2019.

[61] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in Machine Learning in Medical Imaging: 8th International Workshop, MLMI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Proceedings 8, 2017: Springer, pp. 379-387.

[62] S. A. Taghanaki et al., "Combo loss: Handling input and output imbalance in multi-organ segmentation," Computerized Medical Imaging and Graphics, vol. 75, pp. 24-33, 2019.

[63] G. J. Domek, L. Heller Szafran, A. Jimenez-Zambrano, and L. Silveira, "Impact on maternal postpartum depressive symptoms of a primary care intervention promoting early language: a pilot study," Maternal and Child Health Journal, vol. 27, no. 2, pp. 346-355, 2023.

[64] V. A. Yakupova, A. D. Suarez, and L. A. Shraibman, "Socioeconomic Risk Factors for Postpartum Depression and Postpartum Post-Traumatic Stress Disorder," CLINICAL PSYCHOLOGY, vol. 20, no. 1, pp. 182-201, 2023.

[65] K. J. Tanna and K. M. Unadkat, "Socio-economic and Psychological Correlates of Postpartum Depression at Six Months," J Indian Med Assoc, vol. 121, no. 3, pp. 21-4, 2023.

[66] K. M. Tabb et al., "Racial differences in immediate postpartum depression and suicidal ideation among women in a Midwestern delivery hospital," Journal of affective disorders reports, vol. 1, p. 100008, 2020.

[67] J. Meijer et al., "Predictive accuracy of Edinburgh postnatal depression scale assessment during pregnancy for the risk of developing postpartum depressive symptoms: a prospective cohort study," BJOG: An International Journal of Obstetrics & Gynaecology, vol. 121, no. 13, pp. 1604-1610, 2014.

[68] B. D. Thombs et al., "Depression screening and patient outcomes in pregnancy or postpartum: a systematic review," Journal of psychosomatic research, vol. 76, no. 6, pp. 433-446, 2014.

[69] M. E. Gerbasi et al., "Associations between commonly used patient-reported outcome tools in postpartum depression clinical practice and the Hamilton Rating Scale for Depression," Archives of Women's Mental Health, vol. 23, pp. 727-735, 2020.

[70] M. Conway and D. O'Connor, "Social media, big data, and mental health: current advances and ethical implications," Current opinion in psychology, vol. 9, pp. 77-82, 2016.

[71] L. Hidalgo-Padilla, M. Toyama, J. H. Zafra-Tanaka, A. Vives, and F. Diez-Canseco, "Association between maternity leave policies and postpartum depression: a systematic review," Archives of Women's Mental Health, vol. 26, no. 5, pp. 571-580, 2023.

Secure Cloud-Connected Robot Control using Private Blockchain

Muhammad Amzie Muhammad Fauzi[1], Mohamad Hanif Md Saad[2], Sallehuddin Mohamed Haris[3], Marizuana Mat Daud[4]
Faculty of Engineering & Built Environment, Universiti Kebangsaan Malaysia, Selangor, Malaysia[1, 2, 3]
Institute of IR4.0, Universiti Kebangsaan Malaysia, Selangor, Malaysia[4]

***Abstract*—With the increasing demand for remote operations and the challenges posed during the COVID-19 pandemic, industries across various sectors, including logistics, manufacturing, and education, have adopted virtual solutions. Cloud-based robot control has emerged as a viable approach for enabling safe remote operation of robots. However, along with the benefits, there are also risks associated with cloud-based robots. In this study, a secure cloud-based robot control system using a blockchain system was developed. The robot utilizes supervisory control to navigate via the internet. The communication system of the robot relies on the ThingsSentral cloud-based IoT platform; enabling communication between the user via a GUI developed using Python Tkinter and the local robot over the internet. To facilitate internet communication, the robot in this study incorporates an ESP32 microcontroller, which provides a low-cost and low-power system capable of connecting to Wi-Fi. However, cloud-based control systems are susceptible to cyberattacks, prompting the use of blockchain cybersecurity in this study to mitigate the risks. The data sent by the supervisor is stored within a private blockchain developed using Python, simultaneously being transmitted to the cloud platform. The developed security system addresses the risks associated with cloud-based robot control systems, such as data tampering and unauthorized misuse, by leveraging the Proof of Work (PoW) and hashing mechanisms.**

Keywords*—Internet of Things (IoT); robot control; cloud computing; cybersecurity; blockchain

I. Introduction

Industries from a variety of sectors have adopted virtual ways of conducting business in these challenging times due to the enormous problems posed by social distance measures. As a result, there is a huge increase in demand for robots that can be operated safely and remotely for practical research in academic settings. The ability to securely and easily control robots from a distance has become possible thanks to cloud-based interconnection, assuring the continuity of hands-on learning experiences [1].

Cloud-based robot control refers to the practice of remotely controlling robots through cloud computing infrastructure. In order to enable effective and scalable control of robotic systems, it includes utilizing the cloud's capabilities and resources [2]. The robot offloads some duties or calculations to the cloud, which can offer improved computational capabilities, storage, and network connectivity, as opposed to merely depending on onboard processing and control systems [3]. Nowadays, autonomous vehicles or cloud-based robots are mostly used in the service sector. There are many different kinds of service robots, including robots for cleaning [4], [5] and housekeeping [6], surveillance [7], entertainment [8], rehabilitation [9], and so forth [10]. In Malaysia, service robots have been widely used in restaurants to deliver food to customers' tables [11].

These days, residential robots are also substantially included in service robots. In smart homes, restaurants, and hospitals, it is crucial [12], [13]. Roombas [14], vacuum cleaners [5], [15], lawnmowers [16], [17], waiter robots [18], security robots, and sentry robots [19], [20] are a few examples of residential robots. Domestic service robots are becoming more and more common because of how convenient they are and how much time they can save homeowners for other things. To improve their usefulness and performance, they frequently make use of cutting-edge sensors, artificial intelligence, and cloud connectivity. Service robots are becoming more and more popular, which will lead to concerns regarding their security.

Because robot systems are networked and rely on software control, guaranteeing cybersecurity is essential in this field. However, because they are primarily focused on other issues, many initiatives fail to take into account how susceptible robot control systems are to cyberattacks. Frequently, projects create their own cloud connections without paying enough attention to security, potentially creating holes in coding frameworks and other parts. Robots are integrated with the Internet as cyber-physical systems, making them vulnerable to different cybersecurity threats like malware, phishing schemes, and illegal access [21]. Businesses and individuals must prioritize cybersecurity precautions, such as using safe passwords, keeping software updated, and utilizing antivirus and anti-malware software, in order to reduce these risks.

Robots that are autonomous and connected to the cloud have the risk of being misused by authorized users in addition to being exploited by hackers. Robots connected to the cloud execute user-issued orders. A security measure needs to be implemented on the robot to ensure that the commands being delivered are coming from the intended user. Blockchain technology can assist in tracking the actions taken in response to commands provided through the system's digital ledger. The technology of blockchain is now gaining a lot of interest and may help with IoT security challenges. Due to its decentralized architecture, ability to provide data immutability, and non-repudiation services, blockchain technology appears to be a

promising approach for securing IoT and protecting user/data privacy.

The advantages of blockchain technology are dependability, security, and efficiency. By limiting access to authorized users, it guarantees the timeliness and accuracy of information for users within a members-only network. A blockchain transaction cannot be changed after it has been recorded, not even by administrators. Time-consuming reconciliations are no longer necessary thanks to the distributed ledger, and smart contracts automate transaction execution [22]. Block-based transactional data is stored in decentralized databases called blockchains. Before being uploaded to the blockchain, each transaction is first checked for accuracy and encrypted by the parties involved or trusted nodes. Depending on the access and encryption methods used, blockchains can be either public or private [23]. There are lots of public blockchain used currently, mainly used in cryptocurrency transactions such as Bitcoin [24] and Ethereum [25] while Hyperledger Fabric [26] is an example of private blockchain.

In this paper, we present a secure cloud-based robot control using blockchain. The robot's control system allows it to navigate toward a designated and specified location while having a security system to prevent cyberattacks such as data breaches and misuse by authorized users. The paper is organized as follows: The next section explains related works on the types of robot control system and cybersecurity of cloud-based robot. The following section presents the steps taken to complete this study, the IoT platform used in this paper (The ThingsSentral), and the development of a supervisory control system. Then, the following section covers the implementation and testing of the cloud-based robot control system, and a performance test analysis of the supervisory control system using blockchain technology. The last section gives the conclusion and future work of the research.

II. Related Works

A. Robot Control Systems

There are numerous ways to control service robots. The control system that is chosen depends on the particular needs of the service robot application for which it is intended. For instance, on-board control systems, LAN control systems, and cloud-based control systems can all be utilized to operate service robots [27]–[29]. The types of control systems and a few instances of their use are presented in Fig. 1.

1) On-board control system: A control system called an on-board control system is one that is built into the robot's hardware. The robot has the ability to operate without external systems. This kind of control system is widely used by autonomous service robots, which are robots that can travel and do tasks on their own thanks to sensors and algorithms. Restaurant robots, cleaning robots, logistic robots, and social robots are a few examples of robots with on-board control systems.

In comparison to cloud-based service robots, on-board control systems can offer real-time reactivity, a high degree of autonomy, less latency or buffering, and better security. To execute the command without error, applications of service robots employing on-board control systems must be controlled without buffer [6], [30]. ROS has some restrictions even though it can be used as an on-board control system. Fully autonomous robots may have several advantages for the user [31], although initial setup may be laborious. The kinematics computation may be aided by the use of coding, such as MATLAB [32], to carry out a command, such as grabbing an object with a robotic arm.

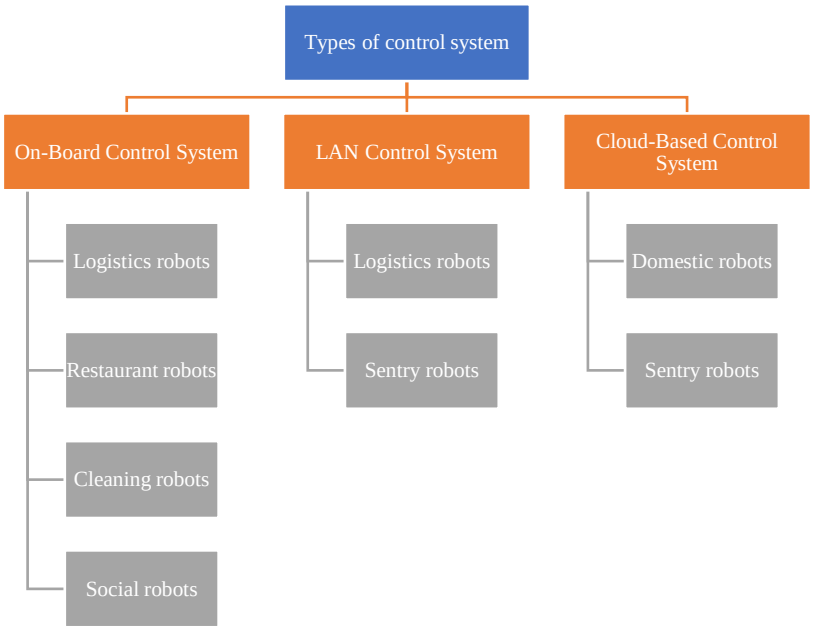


Fig. 1. Types of robot control system.

2) LAN control system: In a LAN (Local Area Network) control system, the robot is connected to a control station or a remote device over a local area network. This approach is frequently used by industrial robots and mapping sentry robots [28]. The robot receives commands from the control station via the LAN and reacts with sensor data and status updates.

LAN management mechanism service robots are machines with remote controls connected to a local network. Due to their connection to a single network, LAN control system service robots may have various benefits over cloud-based and on-board control systems, including reduced latency, high levels of autonomy, and improved security. However, the drawbacks of using LAN to control the service robot are limited external connectivity and unreliable network conditions [33]. This can be avoided by sending data to the server across multi-layered networks or by adding additional relay nodes [34], [35]. This strategy could be laborious, though, as the system's complexity and price will skyrocket.

3) Cloud-based control system: In order to remotely control and manage the behaviour of a service robot, the control system uses cloud computing technology. Domestic robots and sentry robots are a couple of the uses for cloud-based service robots [29]. By fusing the robot's control system with cloud-based servers, this approach enables the robot to quickly access and process enormous amounts of data.

The use of cloud computing in standalone and remotely networked mobile robot systems is the current focus of robotics research. Real-time robot path planning with cloud-based, computationally expensive evolutionary algorithms [36]. Recent studies show that cloud-based control system using IoT platform is increasingly popular. 46 IoT platforms have been described as open, whereas 25 systems are deemed to be open by some research but not by the platforms themselves [37]. In this study, the cloud platform used is The ThingsSentral cloud platform [38]. ThingsSentral is a cloud platform developed by the CAISER research group at the Faculty of Engineering and Built Environment (FKAB), Universiti Kebangsaan Malaysia (UKM).

B. Robot Cybersecurity

Service robot cyber security refers to the safeguarding of service robots against cyberthreats such as malware, hacking, and other cyberattacks. As service robots become more tightly connected to the internet and other devices, the potential of hacks increases [39]. Cyberattacks can take many different forms. DDoS is the most typical [40]. The cloud robot will cease operations if a DDoS attack is launched against its server, which could cause a significant loss [41]. When building service robots [42], secure coding techniques ought to be applied to minimize the risk of vulnerabilities.

1) Authentication: Strong authentication mechanisms should be used to increase the security of service robots [43]. It is essential to make sure that these robots are only accessible to those who are allowed. Passwords, biometric authentication [44], face recognition [45], and other authentication techniques can all be used. Users must supply a legitimate password or passphrase for password-based authentication in order to prove their identity. Biometric identification verifies users by using distinctive biological traits like fingerprints or iris patterns. To authenticate people, face recognition technologies compare and evaluate facial traits. Service robots can build strong security standards, prohibiting unwanted access, and guaranteeing that only authorized users can control and interact with the robots.

2) Encryption: Data encryption is another method to improve the cyber-security of service robots in addition to authentication. Communication between service robots and their control systems should be encrypted to prevent unauthorized data interception [46]. Secure communication technologies that can be used for this include SSL/TLS and VPN. The usage of service robots involves the use of numerous sensors to collect a lot of data, some of which could be sensitive. Through the cloud service, the data is sent to a communication protocol [47]. In order to lower the danger of cyberattacks, data transfer encryption must be robust. Users' sensitive data can be well-protected by using encrypted data.

3) Blockchain: There is only a few research that has been done on cloud-based service robots with blockchain technology as the complexity of the system is high. However, robotics and cloud computing relate highly to one another and are becoming increasingly important as both technologies evolve in the IR4.0 applications. Robotic system management and deployment platforms can be made available by cloud computing. Robotic systems can be centralized handled and watched, allowing for real-time changes and upgrades, by leveraging cloud-based software platforms. Some of these applications are in the domains of smart city and smart industry [48]. Therefore, we can also say that blockchain and cloud computing relate to one another with blockchain and robotics.

Blockchain is used in cloud computing to build decentralized storage solutions that are safe [49]. Blockchain-based storage, in contrast to conventional centralized cloud storage, eliminates single points of failure and potential hacker targets [50]. Blockchain improves security and attack resilience by dispersing data across a decentralized network. Blockchain makes it possible to create secure and open identity management systems in the context of robotics and cloud computing [51]. Users can securely validate their identities when using cloud services through decentralized and tamper-proof records [52], [53]. Blockchain's incorporation into cloud computing provides greater data security, less chances of data tampering, and improved identity verification processes. Fig. 2 shows the basic operation of Blockchain.

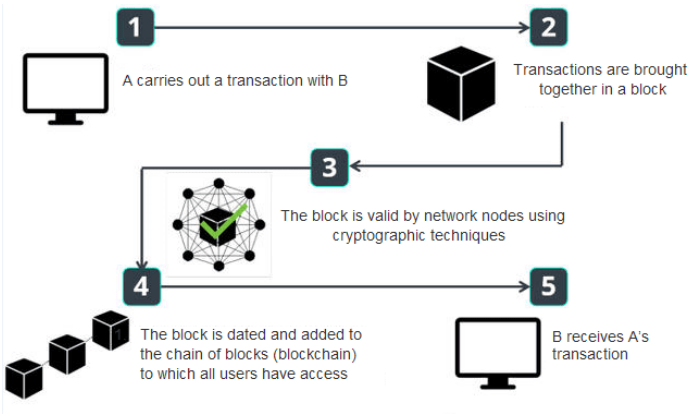


Fig. 2. Basic process of blockchain adopted from (URL:- https://www.centralcharts.com/en/gm/1-learn/1-cryptocurrency/42-trading/699-definition-of-blockchain).

III. Methodology

Fig. 3 depicts the main flowchart of this study. The study begins with the construction of robot using microcontroller, actuators, sensors and power supplies. This study continues with the development of the robot communication system with the cloud platform ThingsSentral. After achieving satisfactory result of the testing of communication system developed, this study continues to the development of GUI control system. The results achieved in this study depend on the GUI developed and also the cybersecurity implemented in this study.

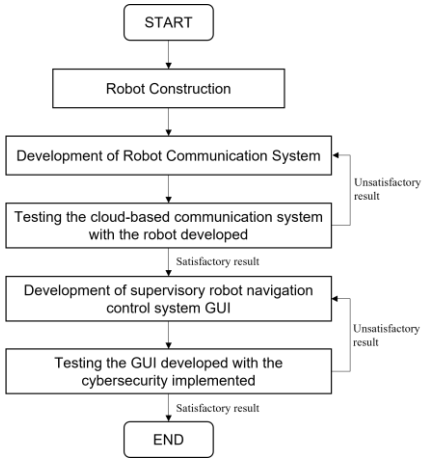


Fig. 3. Flowchart of the study.

A. Development of Robot Communication System

This study uses cloud-based control system as its main communication system. The robot developed, which is located at FKAB, UKM must have the ability to be controlled from a base station located elsewhere. This study uses ThingsSentral cloud platform to send the command from the base station to the robot. The data sent from the supervisor is stored inside the blockchain at the same time it is sent towards the cloud platform. This ensures the command sent by the supervisor as blockchain transactions. Supervisors can also read data from cloud platform such as latest data or distance. The robot reads the data from cloud and sends sensor data to ThingsSentral to be read by the supervisor. Fig. 4 shows the framework of this study.

The software used to develop the communication system between the robot and ThingsSentral cloud platform is Arduino IDE and Python. Both the Arduino IDE and Python software communicate to ThingsSentral web application using Web API. Fig. 5 shows the ThingsSentral web application that was used in this study.

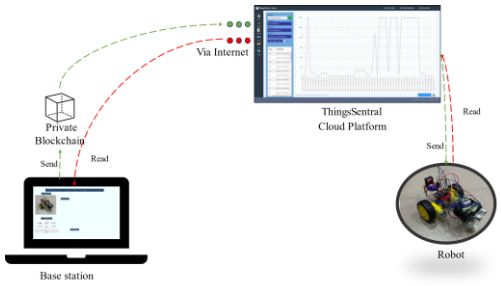


Fig. 4. Framework of the study.

Fig. 5. ThingsSentral web application.

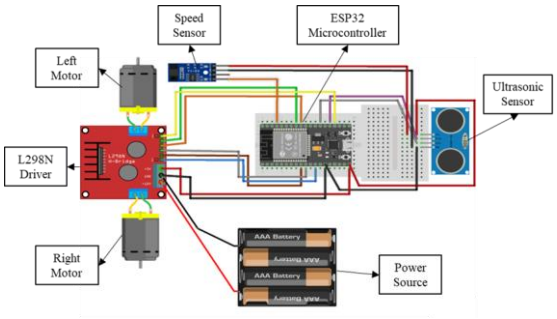


Fig. 6. Schematic diagram of robot.

The robot developed in this study relies on batteries or direct current as its primary power source. The batteries supply power to activate the motors, which are connected to the tires, microcontroller, and sensors. The microcontroller chosen for this research is the ESP32, which utilizes program memory and data memory to store commands and data. Fig. 6 illustrates the schematic diagram of the robot utilized in this study, featuring the ESP32 microcontroller, actuators, and sensors such as ultrasonic and speed sensors. For communication between the robot and the user as a supervisory control system, the ThingsSentral platform is employed. This cloud-based control system allows the robot to be moved remotely and wirelessly. To facilitate control over the robot, a web GUI application has been developed using Python software. This application enables the user to control the robot's actuators and obtain real-time sensor data from the robot.

The data exchange between the ESP32 microcontroller and the cloud platform ThingsSentral is facilitated through an intermediary platform—a GUI developed using Python. This communication system allows for the sending and receiving of data via ThingsSentral. The adoption of cloud-based control systems for service robots is dependent on the specific requirements and constraints of each application. ThingsSentral offers a security system to users, ensuring that the communication mechanism between the user and the robot can be safely controlled without the risk of unauthorized manipulation by attackers.

Moreover, ThingsSentral enables users to control the developed robots over a different network. This means that the robot can be operated from various locations, regardless of the user's distance, as long as there is a stable internet connection [54]. In this study, a cloud platform is essential for the communication system between the robot and the controller because it provides a dedicated static IP address for sending and receiving user data. This differs from a local area network (LAN) control system, as the robot requires its own IP address to ensure proper functioning of the communication system when a cloud platform is not available to assign the IP address. However, it is important to note that the availability of static IP addresses for the robot is limited, which presents some limitations to this solution.

Cloud-based control systems offer numerous advantages, but they also come with drawbacks, including latency and dependence on external networks. The focus of this study is on addressing the security and privacy concerns associated with such systems. While ThingsSentral provides a security system for users, it is still susceptible to misuse by authorized individuals. Hence, the objective of this study is to develop a cybersecurity system to prevent the misuse of the control system by authorized users. In this study, the chosen cybersecurity solution is Blockchain. The reason for selecting Blockchain is its inherent immutability and permanence. Transactions recorded in a Blockchain system cannot be altered or deleted, as they are permanently stored within the system. This ensures the accuracy of every transaction, and users are unable to assign blame to other parties for commands issued by themselves. By leveraging Blockchain, the study aims to enhance the security and integrity of the control system.

B. Secure Cloud-based Robot Control Implementation and Testing

Fig. 7 illustrates the robot prototype utilized in this study, which was developed based on the schematic diagram depicted in Fig. 6. The communication of data is facilitated through the ESP32 microcontroller, serving as the processor for communication between the robot and the ThingsSentral cloud platform. The robot itself is equipped with two tires, each connected to a DC motor, an L298N driver responsible for motor direction control, sensors including an encoder and an ultrasonic sensor, and AA batteries serving as the power source. The microcontroller handles the reading and sending of data to the ThingsSentral web application. To enable this transmission, an intermediary (API) is required to establish a connection between the gateway and the internet.

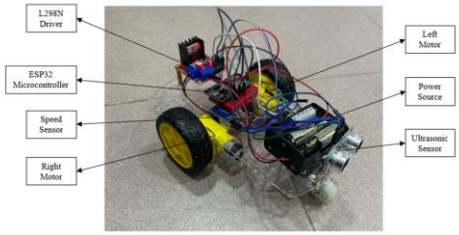


Fig. 7. Robot prototype used in this study.

In this study, HTTP protocols are utilized for both data transfer and data ingestion from ThingsSentral. The sensor nodes receive the transmitted data, which includes the distance measured by the ultrasonic sensor and the motor speed obtained from the encoder. The Arduino IDE software is employed to facilitate data transmission from the sensor nodes to the ThingsSentral cloud platform, as the data is stored within the ESP32 microcontroller. Through programming, the data collected by the sensor nodes can be transmitted to ThingsSentral using URLs and internet networks.

C. Development of Supervisory Control System GUI

Fig. 8 showcases the developed supervisory control system created using Python Tkinter in this study. The graphical user interface (GUI) includes buttons that have been programmed to transmit data values to ThingsSentral, which serve as instructions to control the robot's motors. For example, when the "Stop" button is pressed, it sends a command value of 0. Each motor will be assigned a value of 0 and 1 depending on the state of the motor after the command value is sent. This command value is then utilized by the ESP32 microcontroller to instruct the motors to halt the robot's movement. Each button on the GUI corresponds to a different command value, enabling control over the robot's navigation.

Furthermore, in addition to the buttons used to send command values for robot navigation, the developed GUI also displays the most recent sensor readings obtained from the node sensor via the ThingsSentral cloud platform. The node sensor, located at the bottom right of Fig. 8, represents the ultrasonic sensor. To obtain and display the latest sensor reading, the HTTP Get protocol is employed to retrieve data from ThingsSentral. The ultrasonic reading provides information to the user regarding the distance between the

robot and any obstacles in front of it. An example of the value sent by the GUI to the ThingsSentral cloud platform is presented in Table I.

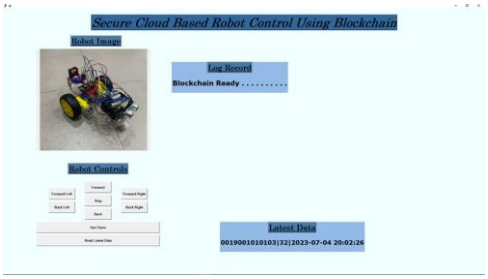


Fig. 8. Supervisory control system.

IV. Results and Discussion

A. Results

The timestamp provided in Table I depicts when the command value was received by ThingsSentral cloud platform. However, the time may differ from when the command value was sent by the supervisory GUI as buffer is present when controlling the robot through cloud-based control system. The difference in time depends on factors such as internet speed and the load sent by the GUI. Additionally, Fig. 9 presents a graphical representation of the command values for each motor plotted against the timestamp.

After successfully testing the supervisory control system's performance, the security of the robot control system underwent evaluation. Employing Python's hashlib module, this study crafted a blockchain system. The command payload from the supervisory control system is incorporated into a block generated by the module. This block includes an index indicating its position in the chain, the block's hash, the timestamp of the sent payload, the payload data specifying the robot's movement command, proof of work, and the previous block's hash. Utilizing the SHA-256 algorithm in the hashlib module, the block's validity, determined by its hash, proof of work, and the previous block's hash, is verified.

TABLE I. Command Value of each Motor

#	Desired Navigation	Command Value	Left Motor	Right Motor	Timestamp
1	Stop	0	0	0	28/6/2023 17:11:24
2	Forward	1	1	1	28/6/2023 17:11:26
3	Forward Left	5	0	1	28/6/2023 17:11:28
4	Forward	1	1	1	28/6/2023 17:11:30
5	Stop	0	0	0	28/6/2023 17:11:32
6	Backward Right	4	-1	0	28/6/2023 17:11:34
7	Backward	2	-1	-1	28/6/2023 17:11:37
8	Stop	0	0	0	28/6/2023 17:11:39
9	Forward Right	3	1	0	28/6/2023 17:11:47
10	Stop	0	0	0	28/6/2023 17:11:50

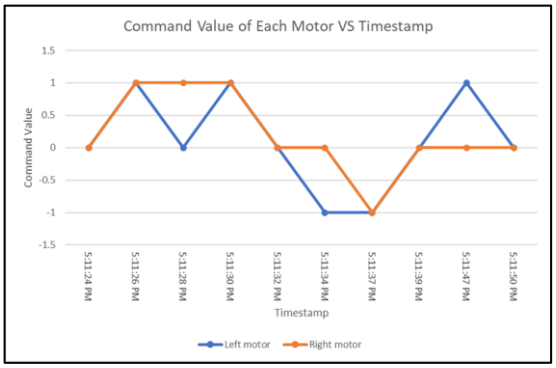


Fig. 9. Graph of command value of each motor Vs timestamp.

The GUI allows for the display of the chain by pressing the "Get Chain" button, as depicted in Fig. 10. The chain is presented on the right side of the GUI, showcasing the total length of the chain and listing the blocks. Each block includes information such as the index number, timestamp, payload (determined by the button pressed to navigate the robot), PoW, and the previous hash. The latest data on the bottom right of Fig. 10 indicates the latest ultrasonic data read from ThingsSentral cloud platform. This data depicts the distance between the robot and an obstacle in front of it.

Fig. 10. Blockchain system in this study.

B. Discussion

The supervisory control system, developed using Python software, is connected to the ESP32 microcontroller through programmed code using Arduino. This connection is facilitated by making ThingsSentral serve as an intermediary. The control system sends data to the cloud, which is then read by the robot via the microcontroller to execute the issued commands. The value of the data sent by the GUI determines the direction of the robot's navigation, instructing the motor driver to activate the robot's motors. The communication of data between the robot and the ThingsSentral cloud platform is achieved using the HTTP GET protocol. Upon receiving the command value, the ESP32 utilizes it as an instruction to engage the motor driver, which controls the robot's movement by directing the motor's speed and rotational direction.

Every command value corresponds to a specific action for the motors, determining the robot's movement or halt according to the intended navigation. For example, if the supervisor instructs the robot to move forward, each motor will be assigned a value between 0 and 1, depending on the desired direction. A positive value indicates forward movement, while a negative value signifies backward motion. Table I provides

an overview of the command values assigned to each motor based on the desired navigation.

The security system developed in this study is based on Blockchain technology. Blockchain offers data tampering and breach prevention mechanisms through its Proof of Work (PoW) and hashing mechanisms. PoW serves as a consensus mechanism in Blockchain networks, creating a robust barrier against data breaches. To modify a block within the blockchain, an attacker would need to alter the block's contents and recalculate the hash, which is computationally expensive and time-consuming due to PoW. Additionally, rewriting the entire blockchain would require an attacker to possess the majority of the network's computational power, making it highly improbable.

The blockchain system developed in this study utilizes the hashlib module in Python software. This module provides a universal interface for various secure cryptographic hash and message digest algorithms. Each hash algorithm has its designated constructor method, creating a hash object with a straightforward interface. The hashlib module supports several hash algorithms, including Secure Hash Algorithm 1 (SHA-1), SHA-224, SHA-256, SHA-384, SHA-512, and MD5. These functions generate fixed-size hash values, known as the hash value or digest, based on the input data they receive.

V. Conclusion

This research presents the development of a secure cloud-based robot control system incorporating blockchain technology. Utilizing an ESP32 microcontroller and Wi-Fi connectivity to ThingsSentral, the robot aims to navigate specific desired locations through a cloud-based control system. The implementation involves Python Tkinter for a user-friendly GUI with dedicated buttons, serving as controllers for robot navigation. The GUI facilitates data transmission to the microcontroller via ThingsSentral, serving as an intermediary cloud platform. However, inherent risks of data tampering and misuse in cloud-based control systems are acknowledged. To counter these concerns, the study introduces a blockchain security system, leveraging hashing, Proof of Work (PoW), and transaction records. Through these mechanisms, the study aims to mitigate the identified risks and enhance overall cybersecurity in the cloud-based robot control system. While this study provides valuable insights, limitations include exclusive use of private blockchain technology visible only to certain organization members, raising concerns about potential transaction payload changes that may impact user confidence in system security. Future research may explore leveraging public blockchains like Ethereum for enhanced transaction verification, offering transparency and validation through universal observation by a diverse user base.

Acknowledgment

The authors would like to thank Universiti Kebangsaan Malaysia for supporting this research. We would also like to thank the Faculty of Engineering and Built Environment and the Institute of IR4.0, UKM for providing the research facilities.

References

[1] Y. Wei, S. Hong, M. Ma, J. Xie, Z. Lu, and X. Zheng, “Raspberry Pi 4B-based cloud-based robot design and demonstration platform construction,” 2023 IEEE 3rd Int. Conf. Power, Electron. Comput. Appl. ICPECA 2023, pp. 1736–1739, 2023, doi: 10.1109/ICPECA56706.2023.10076246.

[2] Y. Wang, N. Wang, Z. Chen, and W. Chen, “A fully cloud-based modular home service robot,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10464 LNAI, no. 100, pp. 320–334, 2017, doi: 10.1007/978-3-319-65298-6_30.

[3] G. Tian, H. Chen, and F. Lu, “Cloud computing platform based on intelligent space for service robot,” 2015 IEEE Int. Conf. Inf. Autom. ICIA 2015 - conjunction with 2015 IEEE Int. Conf. Autom. Logist., no. August, pp. 1562–1566, 2015, doi: 10.1109/ICInfA.2015.7279535.

[4] R. Wang, H. Guan, and J. Lan, “Home services - Track transport robot control system design,” Appl. Mech. Mater., vol. 341–342, pp. 646–649, 2013, doi: 10.4028/www.scientific.net/AMM.341-342.646.

[5] Z. Wei, “Design of Control System for Dust-Collecting Robot Based on DSP,” Proc. - 2017 Int. Conf. Comput. Network, Electron. Autom. ICCNEA 2017, vol. 2017-Janua, pp. 437–440, 2017, doi: 10.1109/ICCNEA.2017.85.

[6] M. Kim, T. Kang, D. Song, and S. J. Yi, “Development of a small-sized intelligent home service robot,” 2021 18th Int. Conf. Ubiquitous Robot. UR 2021, pp. 565–570, 2021, doi: 10.1109/UR52253.2021.9494667.

[7] M. H. Md Saad, “Room Searching Performance Evaluation for the JagaBotTM Indoor Surveillance Robot,” KnE Eng., vol. 1, no. 2015, pp. 1–6, 2016, doi: 10.18502/keg.v1i1.486.

[8] H. S. Ahn et al., “Entertainment services of a healthcare robot system for older people in private and public spaces,” ICARA 2015 - Proc. 2015 6th Int. Conf. Autom. Robot. Appl., pp. 217–222, 2015, doi: 10.1109/ICARA.2015.7081150.

[9] B. Li, G. Li, Y. Sun, G. Jiang, J. Kong, and D. Jiang, “A review of rehabilitation robot,” Proc. - 2017 32nd Youth Acad. Annu. Conf. Chinese Assoc. Autom. YAC 2017, pp. 907–911, 2017, doi: 10.1109/YAC.2017.7967538.

[10] G. Fabregat, J. A. Belloch, J. M. Badia, and M. Cobos, “Design and Implementation of Acoustic Source Localization on a Low-Cost IoT Edge Platform,” IEEE Trans. Circuits Syst. II Express Briefs, vol. 67, no. 12, pp. 3547–3551, 2020, doi: 10.1109/tcsii.2020.2986296.

[11] “The rise of robot waiters in Malaysia, driven by labour shortage in F&B sector | The Star.” https://www.thestar.com.my/lifestyle/living/2022/08/24/the-rise-of-robot-waiters-in-malaysia (accessed Mar. 02, 2023).

[12] Y. Shi and H. Fan, “Research on structure design and kinematics equation of restaurant service robot manipulator,” Adv. Mater. Res., vol. 490–495, pp. 2700–2703, 2012, doi: 10.4028/www.scientific.net/AMR.490-495.2700.

[13] Z. Zhaohui, X. Mei, X. Bian, H. Cai, and J. Ti, “Development of an intelligent interaction service robot using ROS,” Proc. 2016 IEEE Adv. Inf. Manag. Commun. Electron. Autom. Control Conf. IMCEC 2016, pp. 1738–1742, 2017, doi: 10.1109/IMCEC.2016.7867516.

[14] E. Ruiz, R. Acuña, N. Certad, A. Terrones, and M. E. Cabrera, “Development of a control platform for the mobile robot Roomba using ROS and a Kinect sensor,” Proc. - 2013 IEEE Lat. Am. Robot. Symp. LARS 2013, pp. 55–60, 2013, doi: 10.1109/LARS.2013.57.

[15] K. M. Hasan, Abdullah-Al-Nahid, and K. J. Reza, “Path planning algorithm development for autonomous vacuum cleaner robots,” 2014 Int. Conf. Informatics, Electron. Vision, ICIEV 2014, pp. 1–6, 2014, doi: 10.1109/ICIEV.2014.6850799.

[16] A. V. Proskokov, M. V. Momot, D. N. Nesteruk, E. S. Terentyev, and A. D. Veretennikov, “Software and Hardware Control Robotic Lawnmowers,” J. Phys. Conf. Ser., vol. 1059, no. 1, 2018, doi: 10.1088/1742-6596/1059/1/012018.

[17] M. Ryalat, M. Alsherqatli, and H. Elmoaqet, “IoT-aided Smart Lawnmower,” ACM Int. Conf. Proceeding Ser., 2019, doi: 10.1145/3386164.3387298.

[18] T. M. N. U. Akhund, M. A. B. Siddik, M. R. Hossain, M. M. Rahman, N. T. Newaz, and M. Saifuzzaman, "IoT Waiter Bot: A Low Cost IoT based Multi Functioned Robot for Restaurants," ICRITO 2020 - IEEE 8th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir., pp. 1174–1178, 2020, doi: 10.1109/ICRITO48877.2020.9197920.

[19] D. Dai et al., "Detecting, locating and crossing a door for a wide indoor surveillance robot," 2013 IEEE Int. Conf. Robot. Biomimetics, ROBIO 2013, no. December, pp. 1740–1746, 2013, doi: 10.1109/ROBIO.2013.6739719.

[20] S. Roy, T. Vo, S. Hernandez, A. Lehrmann, A. Ali, and S. Kalafatis, "IoT Security and Computation Management on a Multi-Robot System for Rescue Operations Based on a Cloud Framework," Sensors, vol. 22, no. 15, 2022, doi: 10.3390/s22155569.

[21] H. Sadek, R. Bassim, S. H. El-Ghonemy, M. Soltan, and D. El-Serafi, "Clinical characteristics and cognitive functions of late-onset psychoses: A case-control study," Middle East Curr. Psychiatry, vol. 19, no. 3, pp. 149–156, 2012, doi: 10.1097/01.XME.0000415592.81690.e8.

[22] "What is Blockchain Technology? - IBM Blockchain | IBM." https://www.ibm.com/my-en/topics/what-is-blockchain (accessed Feb. 14, 2023).

[23] "What's a blockchain?" https://www.centralcharts.com/en/gm/1-learn/1-cryptocurrency/42-trading/699-definition-of-blockchain (accessed Jul. 07, 2023).

[24] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, "An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends," Proc. - 2017 IEEE 6th Int. Congr. Big Data, BigData Congr. 2017, pp. 557–564, 2017, doi: 10.1109/BigDataCongress.2017.85.

[25] S. Huh, S. Cho, and S. Kim, "Managing IoT devices using blockchain platform," Int. Conf. Adv. Commun. Technol. ICACT, pp. 464–467, 2017, doi: 10.23919/ICACT.2017.7890132.

[26] E. Androulaki et al., "Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains," Proc. 13th EuroSys Conf. EuroSys 2018, vol. 2018-Janua, 2018, doi: 10.1145/3190508.3190538.

[27] I. Karabegović, E. Karabegović, M. Mahmić, and E. Husak, "The application of service robots for logistics in manufacturing processes," Adv. Prod. Eng. Manag., vol. 10, no. 4, pp. 185–194, 2015, doi: 10.14743/apem2015.4.201.

[28] M. Azfar, C. K. Wei, and G. Leng, "Design and Development of an Indoor UAV," no. 1, pp. 89–93, 2015.

[29] G. Li, H. Wang, X. Ying, and J. Liu, "A proxy-based cloud infrastructure for home service robots," Proc. 2015 27th Chinese Control Decis. Conf. CCDC 2015, no. 61375087, pp. 5718–5723, 2015, doi: 10.1109/CCDC.2015.7161824.

[30] A. Koubâa et al., "Turtlebot at Office: A Service-Oriented Software Architecture for Personal Assistant Robots Using ROS," Proc. - 2016 Int. Conf. Auton. Robot Syst. Compet. ICARSC 2016, pp. 270–276, 2016, doi: 10.1109/ICARSC.2016.66.

[31] S. Muszynski, J. Stuckler, and S. Behnke, "Adjustable autonomy for mobile teleoperation of personal service robots," Proc. - IEEE Int. Work. Robot Hum. Interact. Commun., pp. 933–940, 2012, doi: 10.1109/ROMAN.2012.6343870.

[32] X. Wang, T. Zhao, and D. Wang, "Grab application of remote operation service robot," J. Phys. Conf. Ser., vol. 1633, no. 1, 2020, doi: 10.1088/1742-6596/1633/1/012031.

[33] S. Jeon and J. Lee, "Multi-robot control architecture for hospital delivery service in unstable network environment," ICINCO 2017 - Proc. 14th Int. Conf. Informatics Control. Autom. Robot., vol. 2, no. Icinco, pp. 270–277, 2017, doi: 10.5220/0006410502700277.

[34] R. Teng, S. Araki, S. Shimizu, K. Yano, and Y. Suzuki, "Multi-Channel Utilization for Local Data Sharing in Multi-Layered Wireless Robotic Networks," 2019 IEEE Int. Conf. Pervasive Comput. Commun. Work. PerCom Work. 2019, pp. 1009–1014, 2019, doi: 10.1109/PERCOMW.2019.8730666.

[35] K. Makino and T. Murase, "Cooperative Mobility Control with Longcut Route Selection to Form Ad Hoc Networks for Multiple Autonomous Mobile Robots," 2021 IEEE 10th Glob. Conf. Consum. Electron. GCCE 2021, pp. 930–931, 2021, doi: 10.1109/GCCE53005.2021.9621922.

[36] X. Dai, X. Ning, Z. Yao, and H. Shao, "Integrating cloud model in evolutionary algorithm for path planning of mobile robots," 2010 IEEE Int. Conf. Inf. Autom. ICIA 2010, pp. 2352–2356, 2010, doi: 10.1109/ICINFA.2010.5512189.

[37] B. Vogel, Y. Dong, B. Emruli, P. Davidsson, and R. Spalazzese, "What is an open IoT platform? Insights from a systematic mapping study," Futur. Internet, vol. 12, no. 4, pp. 1–19, 2020, doi: 10.3390/FI12040073.

[38] M. H. Md Saad, M. H. S. Akmar, A. S. S. Ahmad, K. Habib, A. Hussain, and A. Ayob, "Design, Development Evaluation of A Lightweight IoT Platform for Engineering Scientific Applications," 2021 IEEE 12th Control Syst. Grad. Res. Colloquium, ICSGRC 2021 - Proc., no. August, pp. 271–276, 2021, doi: 10.1109/ICSGRC53186.2021.9515199.

[39] H. Pu, L. He, P. Cheng, M. Sun, and J. Chen, "Security of Industrial Robots: Vulnerabilities, Attacks, and Mitigations," IEEE Netw., pp. 1–12, 2022, doi: 10.1109/MNET.116.2200034.

[40] S. Yu, Towards the Cloud Computing from, vol. 2. Springer International Publishing, 2018.

[41] W. Dudek and W. Szynkiewicz, "Cyber-security for mobile service robots - Challenges for cyber-physical system safety," J. Telecommun. Inf. Technol., no. 2, pp. 29–36, 2019, doi: 10.26636/JTIT.2019.131019.

[42] G. Lawitzky and M. Buss, "Service robots," IT - Inf. Technol., vol. 49, no. 4, pp. 211–212, 2007, doi: 10.1524/itit.2007.49.4.211.

[43] M. N. Zhukova, V. V. Zolotarev, V. G. Zhukov, and A. S. Polyakova, "Service robot security from unauthorized access by connection control," Proc. - Int. Conf. Dev. eSystems Eng. DeSE, vol. October-20, pp. 526–529, 2019, doi: 10.1109/DeSE.2019.00102.

[44] K. Chellappan and M. S. A. Razak, "Adapting Service Robot Mechanism in Designing Movable Makerspace for Knowledge Society Building," 2021 6th Int. Conf. Robot. Autom. Eng. ICRAE 2021, pp. 339–343, 2021, doi: 10.1109/ICRAE53653.2021.9657792.

[45] J. Wang, J. Zheng, S. Zhang, J. He, X. Liang, and S. Feng, "A face recognition system based on local binary patterns and support vector machine for home security service robot," Proc. - 2016 9th Int. Symp. Comput. Intell. Des. Isc. 2016, vol. 2, pp. 303–307, 2016, doi: 10.1109/ISCID.2016.2079.

[46] H. Cai, X. Liu, and A. Cangelosi, "Security of Cloud Intelligent Robot Based on RSA Algorithm and Digital Signature," 2019 IEEE Symp. Ser. Comput. Intell. SSCI 2019, pp. 1453–1456, 2019, doi: 10.1109/SSCI44817.2019.9002649.

[47] P. Kotuszewski et al., "Cyber-Security Assessment of Industry 4.0 Enabled Mechatronic System," Complexity, vol. 2021, 2021, doi: 10.1155/2021/6670625.

[48] D. C. Nguyen, P. N. Pathirana, M. Ding, and A. Seneviratne, "Integration of Blockchain and Cloud of Things: Architecture, Applications and Challenges," IEEE Commun. Surv. Tutorials, vol. 22, no. 4, pp. 2521–2549, 2020, doi: 10.1109/COMST.2020.3020092.

[49] A. K. Shrestha and J. Vassileva, "Towards decentralized data storage in general cloud platform for meta-products," ACM Int. Conf. Proceeding Ser., 2016, doi: 10.1145/3010089.3016029.

[50] S. Uthayashangar, T. Dhanya, S. Dharshini, and R. Gayathri, "Decentralized Blockchain Based System for Secure Data Storage in Cloud," 2021 Int. Conf. Syst. Comput. Autom. Networking, ICSCAN 2021, 2021, doi: 10.1109/ICSCAN53069.2021.9526408.

[51] X. Su, I. Ullah, M. Wang, and C. Choi, "Blockchain-Based System and Methods for Sensitive Data Transactions," IEEE Consum. Electron. Mag., vol. 2248, no. c, pp. 1–9, 2021, doi: 10.1109/MCE.2021.3076985.

[52] K. Yu, L. Tan, M. Aloqaily, H. Yang, and Y. Jararweh, "Blockchain-Enhanced Data Sharing with Traceable and Direct Revocation in IIoT," IEEE Trans. Ind. Informatics, vol. 17, no. 11, pp. 7669–7678, 2021, doi: 10.1109/TII.2021.3049141.

[53] B. Sun, Q. Dang, Y. Qiu, L. Yan, C. Du, and X. Liu, "Blockchain Privacy Data Access Control Method Based on Cloud Platform Data," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 6, pp. 10–18, 2022, doi: 10.14569/IJACSA.2022.0130602.

[54] S. M. Haris, M. Zamry, and A. Samah, "A CONCEPTUAL DESIGN OF CLOUD-BASED AUTONOMOUS GROUND VEHICLE ROBOT NAVIGATION CONTROL FOR IR4 . 0 APPLICATIONS," vol. 7, no. 28, pp. 164–175, 2022, doi: 10.35631/JISTM.728011.

An Edge Computing-based Handgun and Knife Detection Method in IoT Video Surveillance Systems

Haibo Liu, Zhubing HU*
Department of Electrical and Electronics Engineering, Hebei Petroleum University of Technology, Chengde 067000, China

***Abstract*—Real-time handgun and knife detection on edge devices within the Internet of Things (IoT) video surveillance systems hold paramount importance in ensuring public safety and security. Numerous methods have been explored for handgun and knife detection in video-based surveillance systems, with deep learning-based approaches demonstrating superior accuracy compared to other methods. However, the current research challenge lies in achieving high accuracy rates while managing the computational demands to meet real-time requirements. This paper proposes a solution by introducing a single-stage convolutional neural network (CNN) model tailored to address this challenge. The proposed method is developed using a custom dataset, encompassing model generation, training, validation, and testing phases. Extensive experiments and performance evaluations substantiate the efficacy of the proposed approach, which achieves remarkable accuracy results, thus showcasing its potential for enhancing real-time handgun and knife and knife detection capabilities in IoT-based video surveillance systems.**

***Keywords*—Real-time detection; handgun and knife detection; edge devices; IoT video surveillance; deep learning; convolutional neural network**

I. Introduction

Video Surveillance Systems (VSS) have gained tremendous significance in various domains due to their ability to monitor and analyze activities in real-time [1, 2]. With the advent of the Internet of Things (IoT), these surveillance systems have become even more versatile and effective [3]. The integration of IoT technology into video surveillance systems has opened up new avenues for efficient data collection, analysis, and decision-making in diverse applications ranging from security and safety to healthcare and industrial automation [4, 5].

In the realm of IoT-based video surveillance, a pivotal role is played by edge computing devices [6]. These devices, situated at the edge of the network, are responsible for processing data closer to the data source, thereby reducing latency, conserving bandwidth, and enabling real-time analytics [7]. Edge computing enhances the capabilities of IoT video surveillance systems by enabling rapid data processing and timely response to detected events.

One specific and critical application in this context is real-time detection on edge devices [8]. The ability to detect firearms in real time has significant implications for enhancing public safety and security measures [9, 10]. Achieving accurate and rapid handgun detection on edge devices requires sophisticated methods that can handle the computational constraints posed by these devices while maintaining high levels of accuracy.

Deep learning-based approaches have garnered substantial attention in the realm of real-time handgun detection due to their remarkable capabilities in handling complex visual patterns and achieving high accuracy [11, 12]. These methods utilize deep neural networks to automatically learn intricate features from images and videos, thus enabling accurate object detection tasks [13]. This has led to a surge in research efforts exploring deep learning-based methodologies for real-time handgun detection compared to traditional methods.

Despite the advancements, there exist certain limitations and research challenges in the realm of deep learning-based approaches for handgun detection [14]. Pursuing high accuracy while maintaining real-time performance demands innovative solutions [15, 16]. Addressing these challenges necessitates further investigation and exploration of novel methodologies to ensure the efficacy of real-time handgun detection systems.

In this study, we propose a deep learning method utilizing single-stage convolutional neural network (CNN) architecture to address the requirements of handgun detection. The adopted deep learning approach is justified by its ability to balance accuracy and real-time constraints, making it a promising candidate for the addressed research challenge. The proposed model is trained, validated, and tested using a custom dataset, allowing for robust evaluation of its performance.

This research contributes to the field in three key ways. Firstly, a custom dataset is generated specifically designed for the challenge of handgun detection. Secondly, an efficient deep-learning method is proposed for accurate and real-time handgun detection on edge devices. Lastly, extensive experiments and performance evaluations are conducted to validate the effectiveness of the proposed method, shedding light on its potential contributions to the domain of IoT-based video surveillance and public safety.

II. Related Work

The author in [11] presented a method for automatic handgun detection using deep learning in video surveillance images. The approach involves training a deep neural network on a labeled dataset of surveillance images containing handguns. The network utilizes convolutional layers to extract features and make predictions. The method achieves promising results in detecting handguns in real-time video streams. However, there are some limitations to consider. The accuracy of detection can be influenced by variations in lighting, object occlusions, and different camera angles. Additionally, the

model's performance might degrade when faced with new environments or different handgun types not well-represented in the training data. Further research is needed to enhance the robustness of the method and address these challenges effectively.

The paper in [17] introduced a technique for handgun detection using human pose information. The method involves utilizing pose estimation models to extract key human joint positions from images. By analyzing the spatial relationships between these joints, potential handguns can be identified. The approach demonstrates effectiveness in identifying handguns in various poses. However, limitations include potential false positives due to similar joint configurations and the reliance on accurate pose estimation, which may suffer in challenging scenarios such as low-resolution or occluded images. Further refinement of the method and addressing these limitations are essential for real-world application.

The paper in [18] presented TYOLOV5, a real-time handgun detection system for videos based on quasi-recurrent neural networks. The method integrates YOLOv5 architecture with temporal information to enhance detection accuracy. It successfully detects handguns in video streams, but it may face challenges with complex backgrounds and rapid motion, leading to false positives or missed detections. Further improvements are needed to optimize its performance in dynamic scenarios and mitigate limitations related to occlusions and varying lighting conditions.

The author in [19] focused on enhancing handgun detection by combining visual features with body pose-based data. The method involves extracting both appearance-based features and human body joint positions from images. By integrating these features, the detection algorithm achieves improved accuracy in identifying handguns. However, challenges like limited effectiveness in cases of occlusion and varying poses, as well as potential false positives from similar joint configurations, need to be addressed for robust real-world deployment.

The author in [20] presented the CCTV-Gun benchmark for handgun detection in CCTV images. The method involves curating a dataset with labeled images containing handguns to assess detection algorithms. Various state-of-the-art models are evaluated using this benchmark, demonstrating their effectiveness. However, limitations include potential biases in the dataset and a focus on handguns only, neglecting other potential threats. To address these limitations, future work should encompass a more diverse range of objects and consider broader contextual factors to ensure comprehensive video surveillance.

The paper in [13] introduced a deep-learning framework for handgun and knife detection using edge devices with indoor video surveillance cameras. The method employs a neural network model optimized for edge computing to identify handguns and knives. While achieving real-time detection, limitations arise from potential constraints of edge devices, such as limited processing power and memory. Additionally, the model's performance might be affected by variations in lighting conditions and camera angles, warranting further research to enhance robustness and adaptability to diverse scenarios.

III. Research Methodology

A. Dataset Preparation

The dataset creation process involves two distinct variations: augmented and non-augmented. Augmentation entails the application of transformations such as rotation, scaling, and flips to the original images. These alterations expand the dataset's diversity and complexity, enabling the model to comprehend a broader array of scenarios. Rotation introduces images from various angles, scaling accounts for size variations, and flips reflect different orientations. Rotation is one of the key augmentation techniques, and it entails rotating the original images at different angles. This introduces images from various perspectives, allowing the model to learn from different viewpoints and orientations. For example, in a dataset of handwritten digits, rotating the images can help the model recognize numbers written at various angles, just like how humans can read numbers whether they are upside down or sideways. This augmentation enriches the dataset by simulating real-world variability, enhancing the model's adaptability when confronted with novel situations. Scaling, another important augmentation technique, takes care of size variations. This means resizing the images to different scales, which can simulate scenarios where objects appear closer or farther away in the real world. For instance, in an image dataset for object recognition, scaling can help the model recognize objects that are either close up or in the distance.

Flips are yet another augmentation technique and involve creating mirror images or reversing the orientation of the original images. This mimics situations where an object or scene is seen from a different perspective or orientation. For instance, in image recognition for self-driving cars, flips can help the model adapt to objects that are seen in the rearview mirror or through the side mirrors

As shown in Table I, in terms of dataset composition, it adheres to a structured distribution of 70-20-10, allocated for training, validation, and testing, respectively. This distribution is strategically designed to ensure that the model learns extensively, validates its performance, and rigorously tests its capabilities. With 70% of the data devoted to training, the model grasps underlying patterns and learns to recognize handguns and knives under differing conditions. The 20% validation subset enables performance evaluation during training, allowing fine-tuning and parameter adjustment. Lastly, the 10% testing fraction evaluates the model's generalization on entirely new, unseen data, objectively assessing its practical applicability.

Incorporating these assumptions into the broader context underscores the importance of assembling a dataset that encapsulates the intricacies of real-world scenarios. The diversity of images featuring handguns and knives, captured from multiple angles, lighting settings, and backgrounds, emulates the complexity of actual situations. To empower the model for precise detection and classification, annotations encompass bounding box coordinates and class labels. This information equips the model to not only identify the presence of handguns and knives but also understand their spatial arrangement within the images. The combination of data augmentation techniques, well-structured dataset distribution,

and comprehensive annotations collectively fortifies the model's ability to generalize effectively, paving the way for robust performance across diverse real-world settings.

B. YOLO-based Model Setup

Fig. 1 shows the structure of the proposed method. As shown in Fig 1, the YOLO-based handgun and knife detection models are configured by adapting key parameters in the model's architecture and hyperparameters. This involves selecting the appropriate YOLO variant or backbone, setting the number of classes to 2 for handguns and knives, and specifying a consistent input image size of 416x416 pixels. Anchor box sizes are tailored based on object statistics, enhancing object localization accuracy. Hyperparameters like learning rate, weight decay, and loss function weights are meticulously fine-tuned through iterative experimentation to optimize accuracy while accounting for computational efficiency. Table II shows the model configuration for Yolo-based models in the proposed method.

TABLE I. NO AUGMENTED IMAGES IN A DATASET

Number of Images	Train (70%)	Valid (20%)	Test (10%)
965	676	193	96

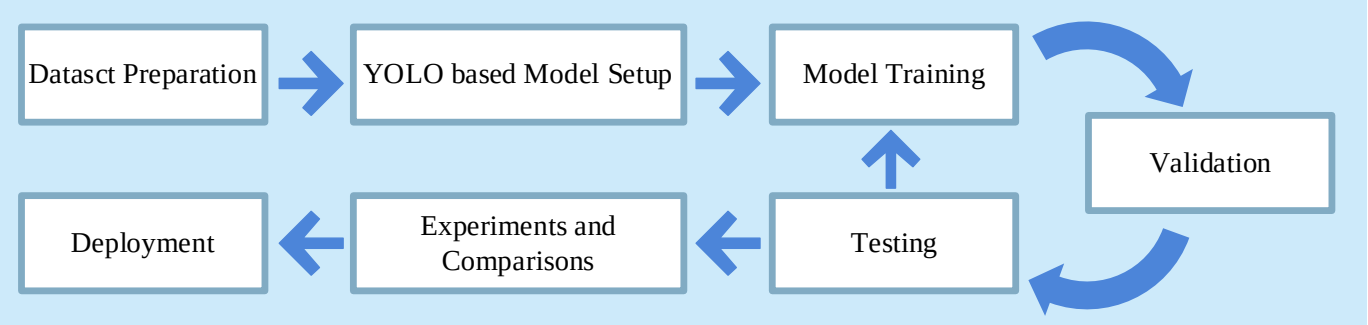


Fig. 1. The structure of the proposed method.

TABLE II. MODEL CONFIGURATION IN THE PROPOSED METHOD

```
model:
  # Choose an appropriate variant: 'n'
  type: YOLOv5  # Or CSPDarknet53 for backbone
nc: 2  # Number of classes: handguns and knives
# Input image size
img_size: 416
# Anchors - Adapt these based on your dataset statistics
anchors:
  - [10,13, 16,30, 33,23]
  - [30,61, 62,45, 59,119]
  - [116,90, 156,198, 373,326]
# Hyperparameters - Fine-tune these based on experimentation
hyp:
  lr0: 0.001  # Initial learning rate
  lrf: 0.2  # Learning rate reduction factor
  momentum: 0.937  # SGD momentum
  weight_decay: 0.0005  # Weight decay
  giou: 0.05  # GIoU loss weight
  cls: 0.58  # Classification loss weight
  cls_pw: 1.0
  obj: 1.0
  obj_pw: 1.0
  iou_t: 0.20  # IOU threshold for objectness loss
  anchor_t: 4.0  # Anchor-multiple threshold
  fl_gamma: 0.0  # Focal loss gamma
  hsv_h: 0.0138
  hsv_s: 0.678
  hsv_v: 0.36
```

As shown in Table II, in the YOLOv5 configuration, the model type is specified by the type parameter, which can be 'YOLOv5' or 'CSPDarknet53', determining the core architecture and feature extractor. The nc parameter sets the number of classes, denoting handguns and knives, while img_size standardizes input image dimensions at 416x416 pixels. The anchors parameter encompasses anchor box sets, vital for object localization, and should be adapted to object aspect ratios and scales. The hyperparameters, central to model optimization, include lr0 for the initial learning rate, lrf for learning rate reduction, momentum for optimization acceleration, and weight_decay for regularization. Parameters like giou, cls, and obj influence loss functions, while iou_t and anchor_t dictate object detection thresholds. Fine-tuning factors like cls_pw, obj_pw, and fl_gamma, along with hsv_h, hsv_s, and hsv_v for augmentation, are also pivotal. It's crucial to iteratively fine-tune these parameters based on experimentation and evaluation to strike the right balance between accuracy and computational efficiency in the context of handgun and knife detection using YOLO models.

C. Training

Training a YOLOv5 model for handgun and knife detection involves several crucial stages that collectively contribute to its accuracy and adaptability. Firstly, during the data loading and augmentation phase, the model's script meticulously processes images and annotations sourced from the training dataset. Augmentation techniques, encompassing random rotations, scaling, flips, and color adjustments, are strategically applied. This augmentation strategy enhances the model's robustness, allowing it to handle a diverse array of real-world scenarios. By exposing the model to a wider range of training examples through augmentation, it gains the capacity to discern objects across varying angles, scales, and lighting conditions.

Subsequently, in the loss calculation step, the model embarks on each training iteration by predicting bounding box coordinates and class probabilities for every object within the images. The pivotal loss function comes into play, which amalgamates crucial components, including localization loss (measured by the Generalized Intersection over Union or GIoU metric), objectness loss, and classification loss. This calculated loss acts as a gauge of the dissimilarity between the model's predictions and the factual annotations, thereby steering the optimization process toward convergence. The calculated losses provide feedback that guides the model in adjusting its internal parameters to align with ground truth annotations more accurately.

As the training unfolds, the process of backpropagation and optimization plays a central role. The computed loss is propagated backward through the model's layers, influencing the gradient updates of the model's weights and biases. The optimization method employed here is stochastic gradient descent (SGD), a foundational algorithm in machine learning. The learning rate and momentum parameters within the optimization process directly impact the extent of weight updates, influencing the model's capacity to navigate the optimization landscape. Furthermore, to finely tune the training procedure, learning rate scheduling is introduced. By incorporating the lr0 parameter and the reduction factor lrf, the learning rate gradually diminishes across training epochs. This dynamic learning rate adjustment facilitates a controlled convergence process, enhancing the accuracy and precision of the model's predictions.

D. Validation and Testing

Validation and testing are essential steps in generating accurate models for handgun and knife detection using the YOLO models. These phases ensure that the trained models not only perform well on the training data but also generalize effectively to unseen scenarios.

During the validation phase, a separate subset of the dataset is used to assess the model's performance as it undergoes training. This helps prevent overfitting, where the model becomes overly specialized to the training data and struggles to perform on new data. The validation dataset consists of images the model hasn't seen before, and the annotations for these images are used to evaluate the model's predictions. By comparing the predicted bounding box coordinates and class probabilities to the ground truth annotations, metrics such as mean average precision (mAP) are calculated. mAP quantifies the model's accuracy across different object categories and various confidence thresholds. This validation process aids in fine-tuning hyperparameters, adjusting learning rates, and making decisions on model checkpoints that offer the best trade-off between precision and recall.

The testing phase evaluates the model's performance on entirely new and unseen data, further confirming its generalization capabilities. A distinct testing dataset is used to assess how well the model can detect handguns and knives in real-world scenarios it has not encountered during training or validation. Similar to validation, the model's predictions are compared to the ground truth annotations to calculate metrics like mAP, offering insights into the model's accuracy on unfamiliar data. Testing validates the model's readiness for real-world deployment and gives an indication of how well it will perform in live environments.

IV. Results And Discussion

This section presents the visual representation of our experimental results and performance evaluation. Fig. 2 demonstrates a visual representation of our experimental results for Yolo models. Moreover, for performance evaluation, standards performance evaluation metrics, including precision, recall, and F- score, are employed inspired by [21, 22]. The details of performance evaluation are discussed in the following sections.

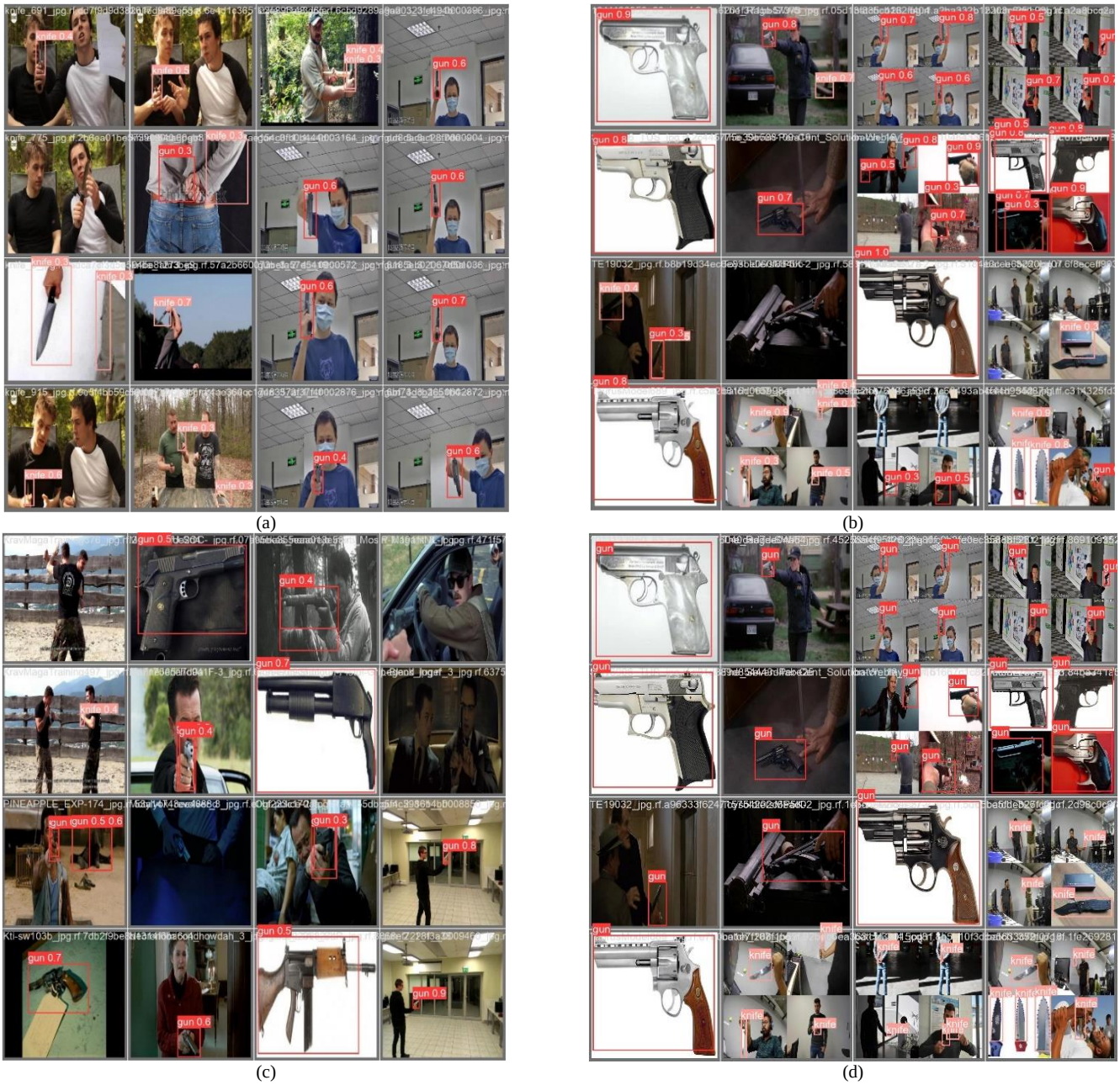


Fig. 2. Visual illustration of experimental results, (a): YOLOv5n-no-aug, (b): YOLOv5n-aug, (c): YOLOv8n-no-aug, (d): YOLOv8n-aug.

V. Performance Evaluations

A. Performance Evaluations of YOLOv5n with No Augmentation Results

In this study, we first employed the YOLOv5 model for the specific task of handgun and knife detection. Notably, we chose to conduct our experiments without incorporating any data augmentation techniques into the dataset. This decision was made to assess the inherent capability of the model without any external modifications to the training data. After training, we rigorously evaluated the model's performance using standard metrics such as precision, recall, and F1-score. These metrics provide a comprehensive view of the model's ability to correctly identify instances of handguns and knives in the test dataset. The absence of augmentation allowed us to directly gauge the model's performance on the original dataset, shedding light on its raw detection capabilities and potential strengths or weaknesses. Fig. 3 shows the results of the performance evaluation of the generated YoloV5 model with no augmentation.

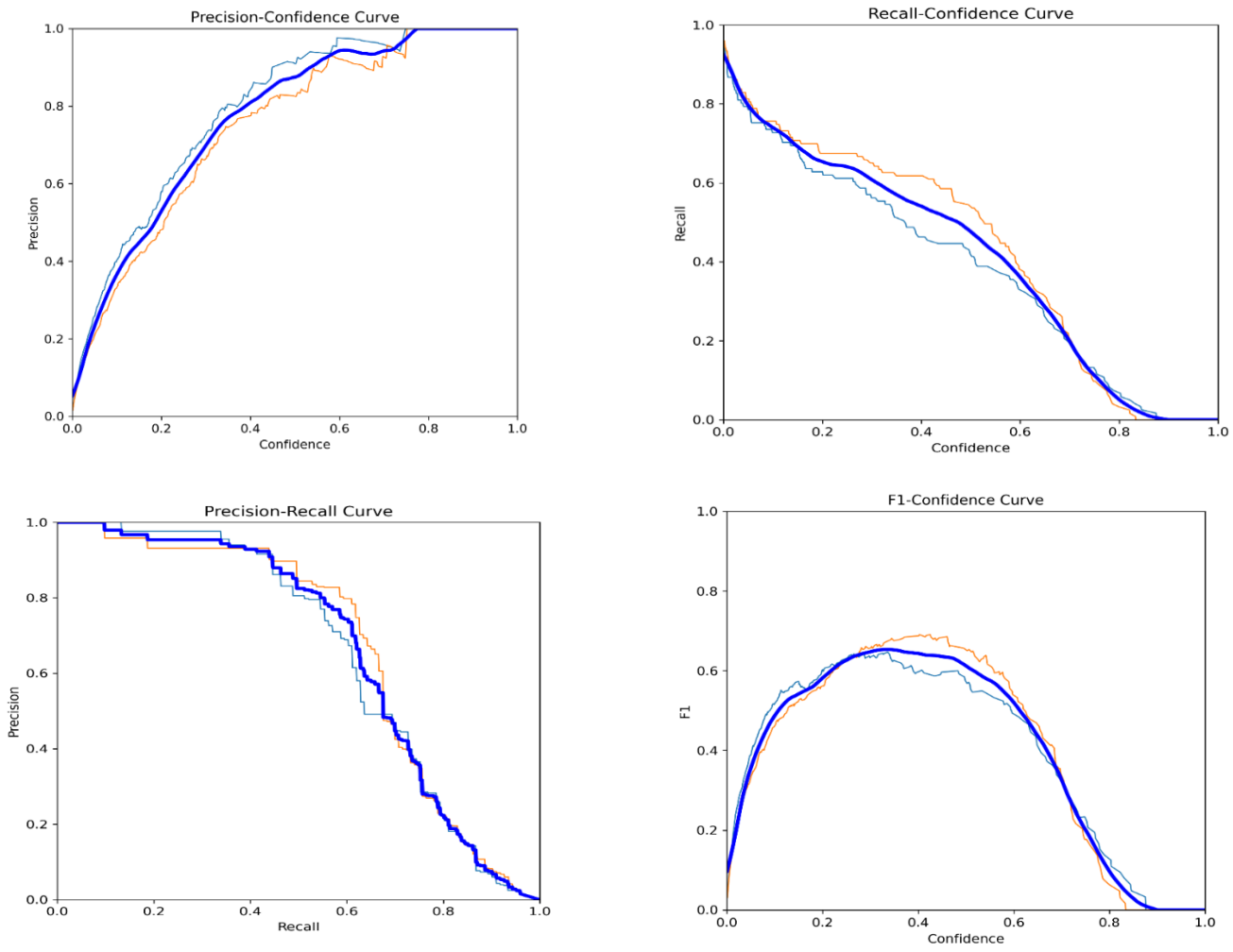


Fig. 3. The result of YOLOv5n on no-augmented dataset.

As depicted in Fig. 3, the evaluation of our YOLOv5 model using precision, recall, PR-curve, and F1-score has provided insightful results that showcase its effectiveness, even in the absence of data augmentation. The average precision (P-curve) of 0.77 for gun detection and 0.92 for knife detection suggests that the model is capable of correctly identifying a significant portion of relevant instances within these classes. Similarly, the high average recall (R-curve) of 0.65 indicates the model's proficiency in capturing a considerable proportion of actual positives within the dataset.

The PR-curve, with an average value of 0.66, illustrates a balanced trade-off between precision and recall. This implies that the model strikes a commendable equilibrium between minimizing false positives and maximizing true positives. Moreover, the F1-score of 0.65 signifies a harmonious blend of precision and recall, indicating the model's strong performance in terms of both accuracy and completeness.

Considering these metrics collectively, the YOLOv5 model demonstrates its reliability and suitability for real-time applications. Despite the absence of data augmentation, the model maintains a consistent and respectable level of performance across multiple evaluation criteria. The high recall values suggest that the model effectively captures instances of handguns and knives, essential for accurate detection in scenarios where prompt identification is critical. Furthermore, the balanced PR-curve and F1-score underscore the model's potential for reliable and precise detection, making it a promising candidate for real-time applications where accurate and swift identification of these objects is paramount.

B. Performance Evaluations YOLOv5n with Augmentation Results

Secondly, we developed a YOLOv5 model for handgun and knife detection. Through dataset augmentation, we diversified the training data with rotations, scaling, and flips. This improved the model's adaptability to real-world scenarios. We evaluated the model using precision, recall, and F1-score, highlighting its capacity to identify instances accurately. The augmentation-enhanced model showcases potential for effective real-time applications, addressing dataset limitations and fostering improved detection performance.

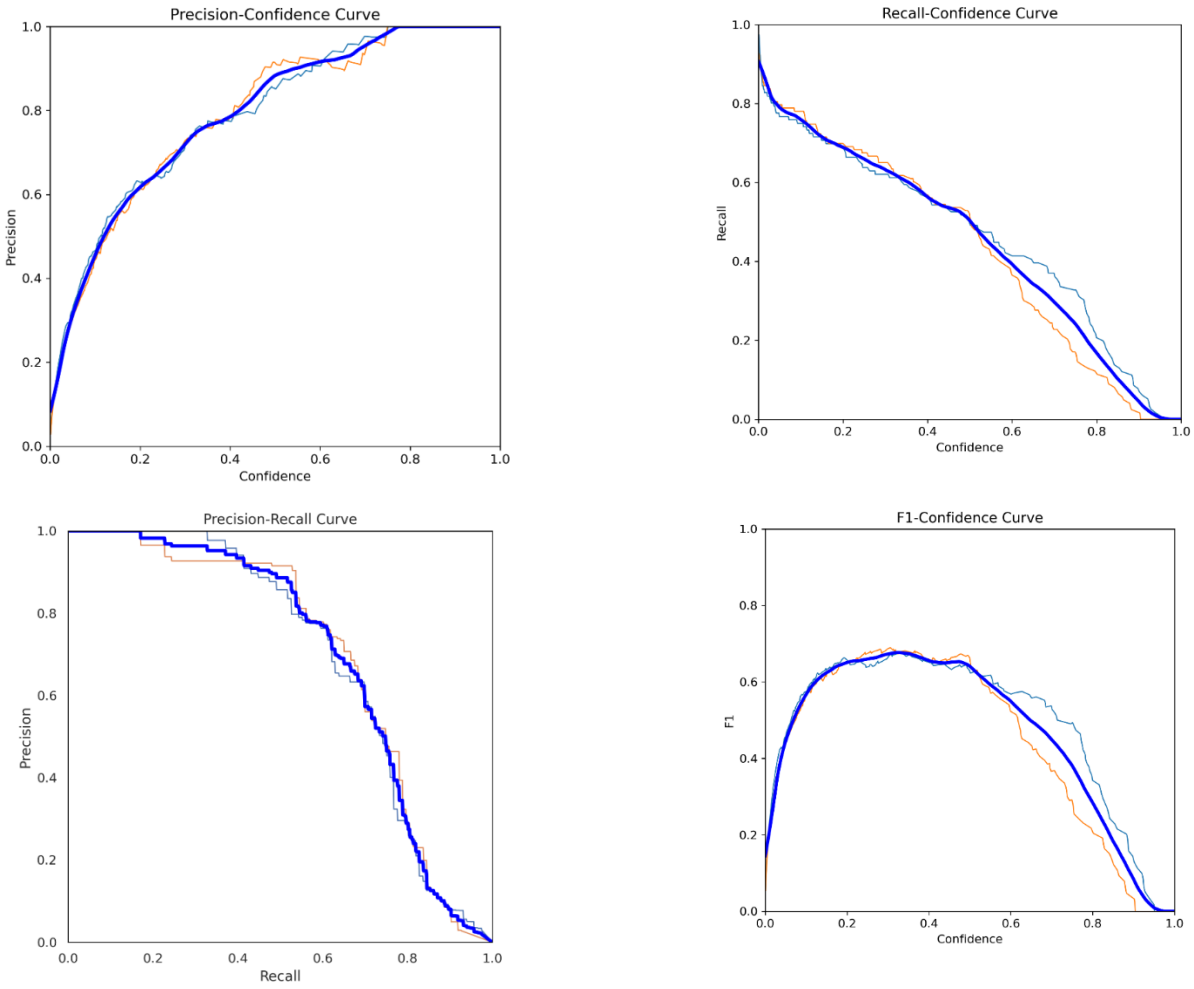


Fig. 4. The result of YOLOv5n on the augmented dataset.

As illustrated in Fig. 4, the evaluation results of the YOLOv5 model for handgun and knife detection, enriched with dataset augmentation, highlight its improved performance compared to the version without augmentation. The higher average precision values of 0.77 for gun and 0.91 for knife detection imply that the augmented model excels in correctly pinpointing instances of these objects. Moreover, the substantial average recall of 0.69 underscores its proficiency in capturing a noteworthy portion of actual positives within the dataset.

The PR-curve's average value of 0.69 signifies that the model effectively balances precision and recall, indicating its capacity to minimize false positives while maximizing true positives. This indicates the model's heightened accuracy in distinguishing relevant instances from the background. The F1-score of 0.68, combining precision and recall, reflects the model's improved overall performance and ability to harmonize between precise detection and comprehensive coverage.

These enhanced metrics collectively demonstrate that the augmented model presents a substantial advancement. Augmentation has expanded the model's understanding of different object appearances and contexts, enabling it to generalize better to unseen scenarios. This has led to heightened accuracy in identifying handguns and knives. Consequently, the augmented YOLOv5 model holds greater potential for real-time applications, where the improved precision, recall, and balanced performance make it a more reliable tool for swift and accurate object detection in dynamic environments.

C. Performance Evaluations YOLOv8n with No Augmentation Results

Thirdly, we developed a YOLOv8n model specifically designed for the detection of handguns and knives. Notably, our experimentation followed a no-augmentation approach, where the dataset remained unaltered. We aimed to evaluate the model's performance in its raw form without the influence of external data modifications. Subsequently, the model underwent a comprehensive evaluation, utilizing precision, recall, and F1-score as the primary metrics. These metrics allowed us to assess the model's precision in identifying instances accurately, its ability to capture actual positives, and the balance between these two factors. Through this evaluation, we sought to gain insights into the model's intrinsic detection capabilities when subjected to real-world scenarios without the aid of dataset augmentation techniques.

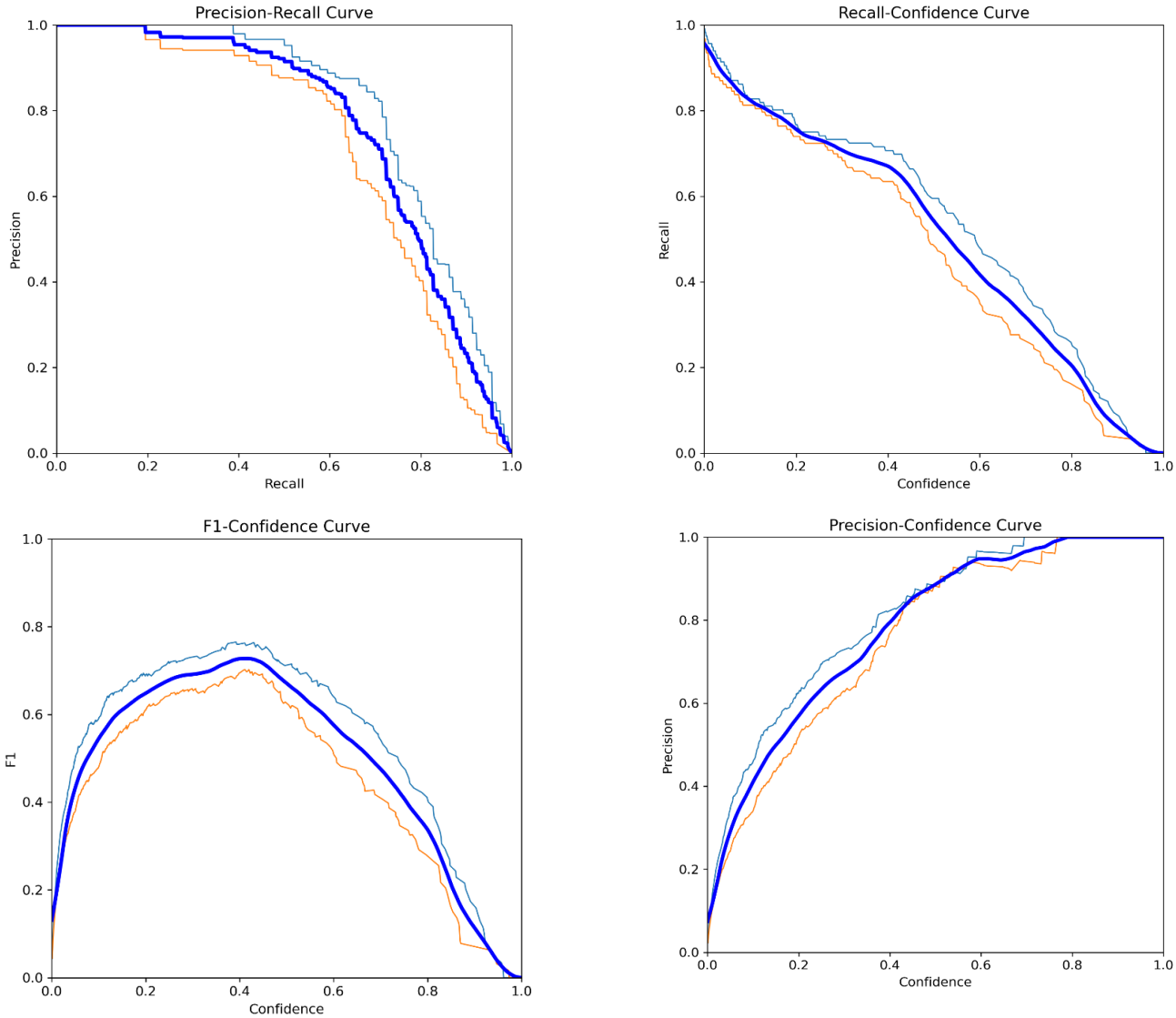


Fig. 5. The result of YOLOv8n on no-augmented dataset.

As shown in Fig. 5, the evaluation results of the YOLOv8n model in handgun and knife detection are highly promising. With an average precision of 0.79 for gun detection and an impressive 0.96 for knife detection, the model showcases its accuracy in correctly identifying instances within these specific classes. These values indicate that the model's predictions are consistently precise, minimizing the occurrence of false positives and boosting its reliability in distinguishing objects of interest.

The average recall of 0.80 is a testament to the YOLOv8n model's exceptional ability to capture a significant proportion of true positives, thereby avoiding missed detections. This indicates that the model effectively identifies and localizes instances of handguns and knives in a wide range of scenarios. The high recall value reflects its proficiency in comprehensively covering the target classes, which is vital for real-time applications where objects might appear in various orientations and scales.

The PR-curve's average value of 0.76 highlights the balanced trade-off between precision and recall achieved by the YOLOv8n model. This equilibrium suggests that the model can achieve high levels of accuracy in identifying relevant instances while maintaining a strong ability to capture true positives. A balanced PR-curve is especially advantageous in scenarios where minimizing false alarms and maximizing detections are critical, making the model suitable for real-world applications.

The F1-score of 0.73 reflects the YOLOv8n model's capacity to integrate precision and recall harmoniously. This indicates that the model is not only precise but also exhibits comprehensive coverage of relevant instances. The F1-score is particularly valuable as it provides a single metric that considers both false positives and false negatives, offering a holistic assessment of the model's performance in a real-world context.

Comparing the YOLOv8n model's performance with the YOLOv5 model, both without augmentation and the augmented YOLOv5 model, reveals distinct trends. While the YOLOv5 model without augmentation had lower precision, recall, PR-curve, and F1-score values, the augmented YOLOv5 model showcased improved performance. However, the YOLOv8n model consistently outperformed both counterparts, excelling in all metrics. This superiority can be attributed to the unique architecture and design choices of YOLOv8n, allowing it to capture object features and contexts better, ultimately resulting in higher accuracy, recall, and balanced performance.

D. Performance Evaluations YOLOv8n With Augmentation Results

Lastly, we developed a YOLOv8n model tailored specifically for handgun and knife detection. Contrasting with the no-augmentation approach, we expanded the dataset through augmentation techniques, effectively diversifying the training data. Similar to the generated YOLOv5 model with augmentation, we aimed to enhance the model's ability to generalize across a broader range of real-world scenarios by introducing variations like rotations, scaling and flips. Our experimentation involved a comparison of the augmented dataset against the original one to evaluate the model's performance under different conditions. This allowed us to gauge the impact of augmentation on the model's detection capabilities, assessing its potential for improved accuracy and robustness when faced with varying object orientations, scales, and backgrounds.

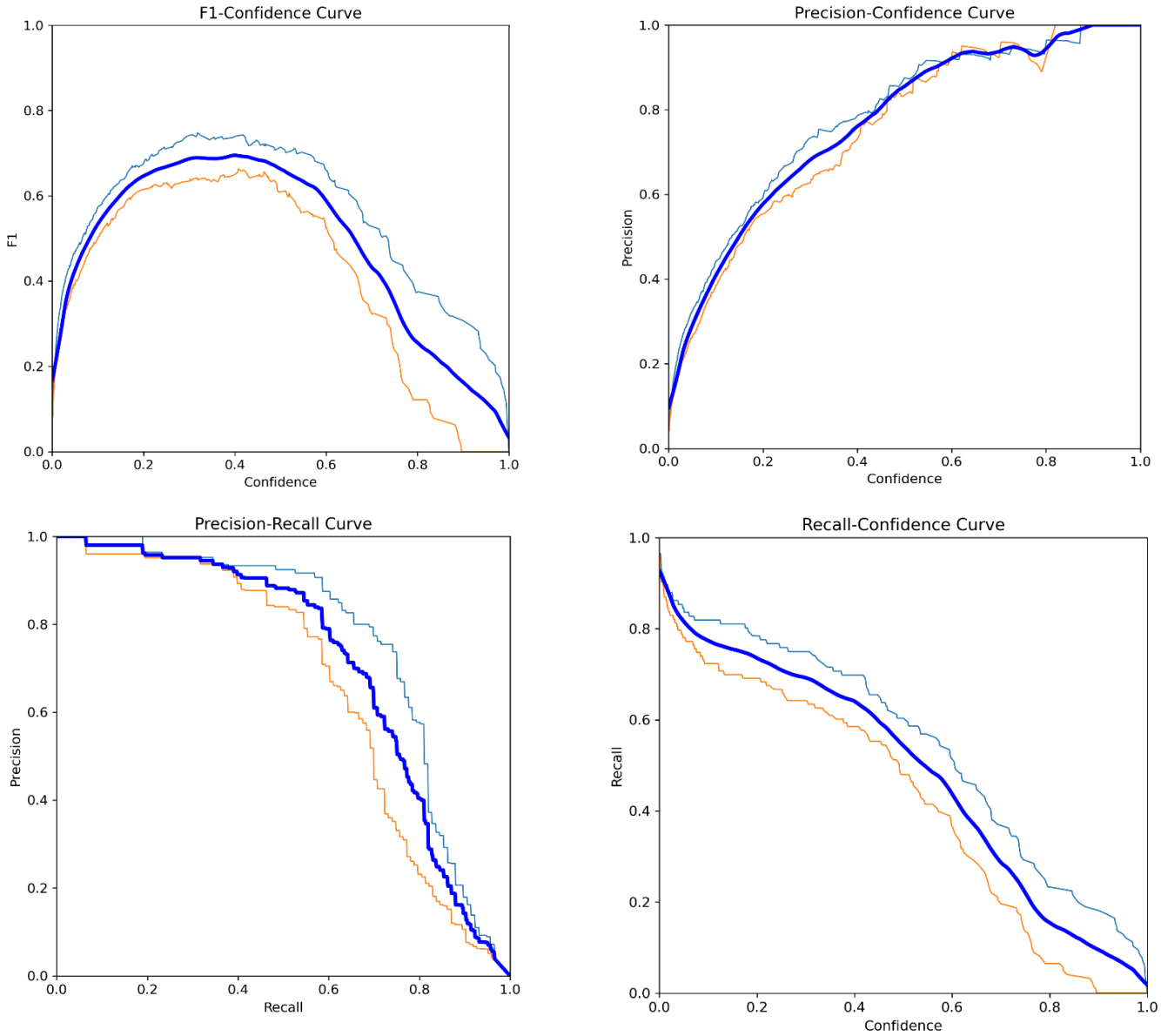


Fig. 6. The result of YOLOv8n on the augmented dataset.

As illustrated in Fig. 6, the evaluation results of the YOLOv8n model with augmentation underscore its remarkable performance, with average precision values of 0.89 for guns and 0.93 for knives. These values suggest that the model accurately identifies instances within these classes, indicating a notable improvement compared to YOLOv8n without augmentation. Additionally, the impressive average recalls of 0.76 highlights the model's proficiency in capturing a significant portion of true positives, further affirming its robustness.

The PR-curve's average value of 0.71 reflects the YOLOv8n model with augmentation's ability to achieve an effective balance between precision and recall. This balance is pivotal in real-time applications, ensuring that the model minimizes false positives while maximizing the identification of true positives. Similarly, the F1-score of 0.69 signifies the model's success in harmonizing precision and recall, which is crucial for maintaining high accuracy and comprehensive coverage.

When comparing the YOLOv8n model with augmentation against its no-augmentation counterpart, the improvements in precision, recall, PR-curve, and F1-score affirm the value of dataset augmentation. Augmentation techniques introduce diversity to the training data, enabling the model to better adapt to real-world variations in object appearance, background, and orientation. This results in enhanced detection performance and better prepares the model for challenges it might encounter in dynamic environments. To ensure fair and objective comparisons between the proposed methodology and other popular methods discussed in the manuscript, a rigorous and standardized evaluation protocol must be employed. By adhering to a transparent and reproducible evaluation framework, the manuscript can provide a clear and credible basis for comparing the proposed approach against existing methods.

In comparison to YOLOv5 without augmentation, the YOLOv8n model with augmentation consistently outperforms it across all metrics. This indicates that YOLOv8n's architecture, combined with augmentation, provides a more effective framework for handgun and knife detection tasks. The YOLOv5 model, although renowned, demonstrates limitations in terms of precision and recall in comparison to both versions of YOLOv8n, reinforcing the advantages of the latter. Fig. 7 shows the comparison of performance results of different experiments.

As depicted in Fig. 7, when contrasting with augmented YOLOv5, the YOLOv8n model maintains its superiority. This suggests that YOLOv8n's architectural enhancements, coupled with augmentation, result in a more refined and adaptable model. While augmentation does enhance YOLOv5, the performance boost offered by YOLOv8n is still apparent, showcasing its advanced capabilities in handling object detection tasks.

Ultimately, the YOLOv8n model with augmentation emerges as the optimal choice for handgun and knife detection in real-time scenarios. Its superior performance across multiple metrics attests to its accuracy, versatility, and robustness. Augmentation proves to be a crucial factor, as it empowers the model to handle diverse and challenging situations, making it more reliable and effective in real-world applications where timely and accurate detection is essential.

As a result, the combination of YOLOv8n's architecture and dataset augmentation yields a powerful model that excels in handgun and knife detection tasks. Its superior precision, recall, PR-curve, and F1-score values, when compared to both YOLOv8n without augmentation and YOLOv5 models, demonstrate its efficacy. This model is well-equipped to address the intricacies of real-time applications, offering heightened accuracy, adaptability, and efficiency in identifying handguns and knives.

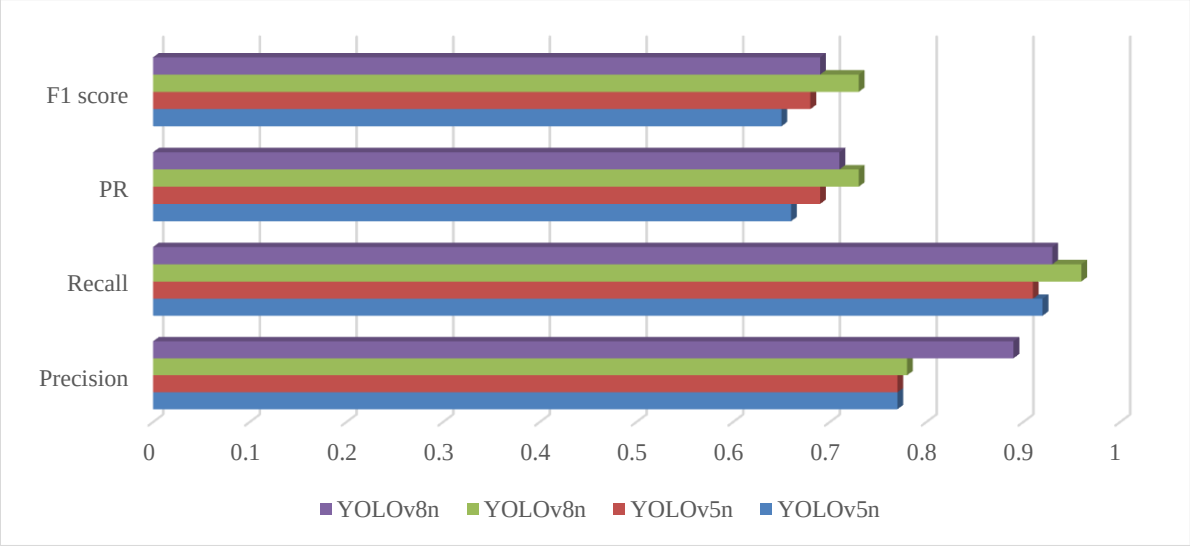


Fig. 7. Comparison of performance results of different experiments.

Moreover, to ascertain the scalability and efficacy of the proposed dataset creation methodology, it is imperative to conduct a comprehensive evaluation of the results obtained. This evaluation process should involve comparing the performance of models trained on augmented and non-augmented datasets across various real-world scenarios and challenges. By using the augmented dataset, the model's ability to adapt to different angles, sizes, and orientations can be thoroughly tested, allowing for a more robust assessment of its capabilities. Metrics such as accuracy, precision, and recall should be considered, along with real-world benchmarks and use cases. The results of this evaluation will not only validate the significance of the augmentation techniques but also demonstrate the dataset's utility in enhancing the model's generalization and adaptability, making it a crucial step in ensuring the success of the proposed work in various practical applications.

VI. CONCLUSION

Real-time handgun and knife detection on edge devices are paramount for enhancing the effectiveness of IoT video surveillance systems. This paper addresses the significance of accurate and timely firearm detection in such systems, highlighting the various methods explored in video-based surveillance contexts. Deep learning-based approaches have demonstrated superior results in handgun and knife detection due to their ability to learn intricate patterns, yet they face challenges concerning accuracy and computational efficiency for real-time operation. This study proposes a solution by introducing a single-stage convolutional neural network model tailored to address the aforementioned research challenge. The proposed method involves model generation through a custom dataset and encompasses comprehensive training, validation, and testing phases. Experimental results and performance evaluations validate the effectiveness of the proposed approach in achieving accurate firearm detection, demonstrating its potential impact on IoT video surveillance systems. Two potential avenues for future research stem from the findings of this study. Firstly, considering the evolving nature of IoT technologies and edge computing, exploring methods to optimize the computational efficiency of the proposed single-stage convolutional neural network model would be valuable. Addressing the current challenges of high computation costs while maintaining real-time capabilities could lead to more scalable and practical implementations. Secondly, delving into the integration of multi-modal sensor inputs, such as audio and environmental data, with the proposed handgun and knife detection model could enhance its robustness and accuracy in complex real-world scenarios. By incorporating additional contextual information, the proposed approach could offer more reliable and comprehensive firearm detection outcomes in diverse IoT video surveillance applications.

REFERENCES

[1] Montero, D., et al., Multi-camera BEV video-surveillance system for efficient monitoring of social distancing. Multimedia Tools and Applications, 2023: p. 1-25.

[2] Singh, R.P., et al. An Intelligent Video Surveillance System using Edge Computing based Deep Learning Model. in 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT). 2023. IEEE.

[3] Li, H., Q. Zhang, and G. Kuldeep. 3TierView-A Three-tier Privacy-preserving Live Video Surveillance IoT System. in Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services. 2023.

[4] Khan, A.A., et al., Data Security in Healthcare Industrial Internet of Things with Blockchain. IEEE Sensors Journal, 2023.

[5] Adithya, K. and R. Girimurugan, Benefits of IoT in Automated Systems. Integration of Mechanical and Manufacturing Engineering with IoT: A Digital Transformation, 2023: p. 235-270.

[6] Gulve, S.P., S.A. Khoje, and P. Pardeshi. Implementation of IoT-based smart video surveillance system. in Computational Intelligence in Data Mining: Proceedings of the International Conference on CIDM, 10-11 December 2016. 2017. Springer.

[7] Elgazzar, K., et al., Revisiting the internet of things: New trends, opportunities and grand challenges. Frontiers in the Internet of Things, 2022. 1: p. 1073780.

[8] Jan, O.R., et al., Real-Time Flood Monitoring with Computer Vision through Edge Computing-Based Internet of Things. Future Internet, 2022. 14(11): p. 308.

[9] Rohith, M. and A. Sunil. Comparative analysis of edge computing and edge devices: key technology in IoT and computer vision applications. in 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT). 2021. IEEE.

[10] Hasan, R., et al. A color frame reproduction technique for IoT-based video surveillance application. in 2017 IEEE International Symposium on Circuits and Systems (ISCAS). 2017. IEEE.

[11] Salido, J., et al., Automatic handgun detection with deep learning in video surveillance images. Applied Sciences, 2021. 11(13): p. 6085.

[12] Ghazal, M., N. Waisi, and N. Abdullah, The detection of handguns from live-video in real-time based on deep learning. TELKOMNIKA (Telecommunication Computing Electronics and Control), 2020. 18(6): p. 3026-3032.

[13] Berardini, D., et al., A deep-learning framework running on edge devices for handgun and knife detection from indoor video-surveillance cameras. Multimedia Tools and Applications, 2023: p. 1-19.

[14] Gali, M., S. Dhavale, and S. Kumar. Real-time image based weapon detection using YOLO algorithms. in International Conference on Advances in Computing and Data Sciences. 2022. Springer.

[15] Vallez, N., A. Velasco-Mata, and O. Deniz, Deep autoencoder for false positive reduction in handgun detection. Neural Computing and Applications, 2021. 33(11): p. 5885-5895.

[16] Pavithra, T., et al. Real-Time Handgun Detection in Surveillance Videos based on Deep Learning Approach. in 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC). 2022. IEEE.

[17] Velasco-Mata, A., et al., Using human pose information for handgun detection. Neural Computing and Applications, 2021. 33(24): p. 17273-17286.

[18] Duran-Vega, M.A., et al., TYOLOV5: a temporal Yolov5 detector based on quasi-recurrent neural networks for real-time handgun detection in video. arXiv preprint arXiv:2111.08867, 2021.

[19] Ruiz-Santaquiteria, J., et al., Improving handgun detection through a combination of visual features and body pose-based data. Pattern Recognition, 2023. 136: p. 109252.

[20] Yellapragada, S., et al., CCTV-Gun: Benchmarking Handgun Detection in CCTV Images. arXiv preprint arXiv:2303.10703, 2023.

[21] Garg, R. and S. Singh. Intelligent video surveillance based on YOLO: a comparative study. in 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3). 2021. IEEE.

[22] Ashraf, A.H., et al., Weapons detection for security and video surveillance using cnn and YOLO-v5s. CMC-Comput. Mater. Contin, 2022. 70: p. 2761-2775.

Advanced Seismic Magnitude Classification Through Convolutional and Reinforcement Learning Techniques

Qiuyi Lin, Jin Li*
School of Architectural Engineering, Chuzhou Polytechnic, Chuzhou 239000, Anhui, China

***Abstract*—Earthquake Early Warning (EEW) systems are crucial in reducing the dangers associated with earthquakes. This paper delves into the realm of EEWs, focusing on rapidly determining earthquake magnitudes (EMs). Traditional methods for swift magnitude categorization often grapple with challenges such as data disparity and cumbersome processes. Our research introduces an innovative EEW model, employing a 7-second seismic waveform record from three different components provided by the China Earthquake Network Center (CENC). This empirical, quantitative study pioneers a method combining dilated convolutional techniques with a novel mutual learning-based artificial bee colony (ML-ABC) algorithm and reinforcement learning (RL) for EM classification. The proposed model utilizes an ensemble of convolutional neural networks (CNNs) to simultaneously extract feature vectors from input images, which are then amalgamated for classification. To address the imbalances in the dataset, we implement an RL-based algorithm, conceptualizing the training process as a series of decisions with individual samples representing distinct states. Within this framework, the network operates as an agent, receiving rewards or penalties based on its precision in distinguishing between the minority and majority classes. A key innovation in our approach is the initial weight pre-training using the ML-ABC method. This technique dynamically optimizes the "food source" for candidates, integrating mutual learning elements related to the initial weights. Extensive experiments were carried out on the selected dataset to ascertain the most effective parameter values, including the reward function. The findings demonstrate the superiority of our proposed model over other evaluated methods, highlighting its potential as a robust tool for EM classification in seismology. This research provides valuable insights for both seismologists and developers of EEW systems, offering a novel, efficient approach to earthquake magnitude determination.**

Keywords*—Earthquake early warning; the magnitude of the earthquake; imbalanced classification; artificial bee colony; reinforcement learning

I. Introduction

The importance of earthquake prediction research is increasing globally, aiming to mitigate earthquake impacts. Early warning systems can drastically reduce casualties and damages from severe earthquakes [1]. These systems, endorsed by the United Nations, detect seismic activities and issue alerts rather than predicting earthquakes. Utilizing the time difference between primary and secondary seismic waves, they send warnings once an earthquake is detected. This allows immediate actions like unlocking building exits, managing power grids, and adjusting operations in critical facilities like hospitals and nuclear power plants [2].

The EEW system has gained significant attention in seismology research, particularly focusing on the "excellent period method" introduced by Nakamura et al. [3] and refined by Allen and Kanamori [4], which estimates earthquake magnitude from the initial three to four seconds of P-wave data. Despite its promise, this method shows notable variability in its estimates. To improve it, Kanamori [5] applied wavelet analysis for a more detailed time-frequency analysis of seismic data, but the problem of variability remained. Lancieri and Zollo [6] proposed an alternative, the "peak ground motion displacement Pd method," which uses peak displacements measured shortly after the P- and S-waves to estimate magnitude. Like its predecessors, this method also suffers from dispersion issues. Overall, current methods for magnitude estimation offer some level of estimation capability but are often limited by significant dispersion and lack of universal applicability.

Class imbalance, where one category has more data than the other, can adversely affect the performance of classification models [7]. The challenge is greater with minority classes due to their smaller size and variability. To address this, two main strategies are employed: data-level and algorithmic-level adjustments. At the data level, one can balance classes by over-sampling the minority class or under-sampling the majority class, but these methods risk information loss and overfitting [7, 8]. While these techniques are promising, their success varies depending on the dataset and application at hand, making it crucial to consider the specific characteristics of the imbalance and the needs of the application when choosing a method [9, 10]. Deep Reinforcement Learning (DRL) [11] has shown promise in improving classification by reducing noise and enhancing features, despite increasing computation time due to complex agent-environment interactions [12, 13]. DRL has been used to improve classifiers and in ensemble pruning, yet its application in addressing class imbalance has not been sufficiently explored by researchers [14].

Deep learning models have revolutionized multiple fields with their ability to fine-tune internal parameters through learning algorithms, particularly using backpropagation [15-18]. This method adjusts the weights of the model to minimize errors, but it can suffer from problems like sensitivity to initial weight settings and getting stuck in local minima, particularly

in classification tasks [19]. To overcome these issues, researchers are looking into meta-heuristic algorithms such as the ABC algorithm [20], which searches the solution space more broadly and is thus less likely to fall into local minima. However, the ABC algorithm itself has challenges in choosing the optimal food source or solution [21]. A new approach, ML-ABC, has been developed to enhance the standard ABC [22]. This method promotes mutual learning between different elements of the algorithm, improving adaptability and possibly resolving the weight initialization issues that affect gradient-based methods [23]. By fostering information exchange within the optimization process, ML-ABC can avoid local minima and find better solutions [24].

This research presents a deep learning model designed for EEW systems, which can accurately detect seismic events with magnitudes of 4 or higher using data from the first 7 seconds after the P-wave arrival from a single seismic station. The model uses dilated convolutional layers for feature extraction, which are then used for classification into two categories: EM ≥4 or EM <4. The data is skewed, with the EM ≥4 category being overrepresented, which adversely affects the model's performance. To tackle the data imbalance, the study applies a RL approach, where an agent iteratively learns from the environment to make binary classifications, receiving rewards for correct predictions and penalties for mistakes. A higher reward is given for correctly predicting the underrepresented class to address the imbalance. The agent aims to maximize classification accuracy and cumulative reward. The paper also addresses the issues of gradient-based training methods, which are sensitive to initial weight settings. The innovative ML-ABC method is introduced as a solution. It dynamically adjusts the optimization process by mutual learning from the initial weights, which improves the model's robustness. The model demonstrates outstanding performance in EEW tasks, with classification accuracy exceeding 90%. The paper highlights the effectiveness of combining DL with RL and the novel ML-ABC method for addressing data imbalance and initialization sensitivity in classification tasks. Below is a succinct summary of the primary themes explored within this paper:

- The article addresses the issue of imbalanced classification by introducing a sequential decision-making RL algorithm.
- Rather than relying on random weight assignments for model parameters, an enhanced ABC algorithm is adopted to establish initial values and a coding strategy for these parameters.
- A unique reward mechanism is integrated into the approach, incentivizing accurate decisions while penalizing errors. This mechanism assigns higher rewards to the minority class, prompting the model to allocate adequate attention to less common data. This strategic adjustment fosters a more equitable and balanced classification process.

The subsequent sections of the paper are organized as follows: Section II presents related works, while Section III provides an in-depth exploration of the proposed approach, detailing the core methodology. Section IV presents the empirical results and their subsequent analysis. In Section V, concluding remarks are provided, along with potential directions for future research.

II. Related Work

As the pool of seismic activity data grows, deep learning is becoming a cornerstone in the development of EEW systems. Ren et al. [25] capitalized on this by implementing CNNs to estimate earthquake magnitudes, treating the process as a classification problem. Their innovative model analyzes 4-second waveform segments from individual seismic channels to determine the potential impact of an earthquake. This approach has been validated through its successful application to distinct earthquake events, such as the clusters in Changning and Tangshan, proving its effectiveness in practical scenarios. Beyond academic validation, the model has seen real-world application, having been adopted by the China Earthquake Network Center (CENC) for real-time seismic data analysis, thus contributing to more timely and reliable earthquake warnings. Wang et al. [26] proposed a sophisticated deep learning model using long short-term memory (LSTM) neural networks to enhance onsite EEW systems. By analyzing the initial P waves to predict destructive S waves, this approach aims to mitigate the limitations of fixed-threshold single indicators, addressing complex nonlinearities due to varying travel paths and site effects. The LSTM model's proficiency was validated by testing it with recent seismic events in Taiwan, achieving remarkable accuracy with a 0% missed alarm rate and a mere 2.01% false alarm rate.

Expanding the scope of DL in EEW, Hu and Zhang [27] have designed a DL model that predicts the range of earthquake magnitudes with greater precision and reliability than traditional models. Despite its sophistication, this model could be prone to overfitting when faced with limited or noisy data, which might limit its application to new seismic events. Wang et al. [28] have pushed the boundaries further with DLcav, a CNN model aimed at predicting the cumulative absolute velocity (CAV) from seismic waveforms. DLcav stands out for enhancing the prediction of earthquake-induced damage, which is a crucial component of EEW. Datta et al. [29] introduced DeepShake, a novel deep spatiotemporal recurrent neural network designed for forecasting shaking intensities using real-time ground-motion data. This network-based model predicts future shaking across an array of stations without prior knowledge of their locations, learning from data on wave propagation patterns. Tested in the 2019 Ridgecrest earthquake sequence, DeepShake successfully alerted for significant shaking events with an equal error rate of 11.4%, indicating its potential as a reliable one-step early warning system for earthquakes.

Kavitha et al. [30] advanced the field of EEW through their development of the 3S-AE-CNN model, a novel innovation in deep learning applications. This model excels in quickly assessing crucial earthquake parameters, such as size and location, immediately after detecting the initial P-wave. The integration capabilities of this model with Internet of Things (IoT) technology could revolutionize the responsiveness of EEW systems. In a similar vein, Yanwei et

al. [31] have presented EEWNet, a cutting-edge deep learning framework that promises to rapidly and accurately predict earthquake magnitudes by analyzing raw P-wave data from single seismic stations. This model is particularly adept at dealing with earthquakes in the magnitude range of 4.0 to 5.9, offering a considerable improvement over existing empirical method.

Meanwhile, Meng et al. [32] have introduced an EEW model that leverages a DenseBlock structure with a Bottleneck and Multi-Head Attention mechanism to classify earthquake magnitudes from 7-second seismic data snippets. However, the model performance could be compromised by the issue of imbalanced data, which poses a risk to its predictive capabilities, especially for less frequent but critical earthquake magnitudes. Lastly, Lin et al. [33] have developed a CNN tailored for magnitude prediction, utilizing just 3-second windows of seismic data and conceptualizing the task as a classification problem, reinforcing the trend towards integrating advanced DL techniques in EEW for enhanced magnitude estimation accuracy. Münchmeyer et al. [34] proposed TEAM-LM, a hybrid model integrating CNN and Transformer architectures for simultaneous magnitude and location prediction using 10 seconds of data. While demonstrating promise within EEW systems, its computational intensity and multi-parameter predictions could impact overall accuracy.

III. Materials and Methods

The structure of the proposed model is shown in Fig. 1. The proposed model is particularly designed to enhance the classification process in seismic data analysis, where the intricacies of imbalanced classes and the necessity for precise initial weight settings are crucial. The adoption of ML-ABC and RL within our model directly addresses these critical areas where existing models fall short. Traditional algorithms often fail to offer a systematic method for initial weight selection, which can hinder the learning phase by causing slower convergence and the risk of settling on suboptimal minima. This can be particularly problematic in seismology, where the timeliness and accuracy of predictions are paramount.

Moreover, current models are not adequately equipped to deal with class imbalance, a frequent issue in seismic data where large magnitude events are rare. These models tend to be biased towards the majority class, resulting in a significant oversight of minority classes which are often the most critical to detect accurately in the context of earthquake early warning systems. Our approach, through ML-ABC, provides a carefully considered and diverse set of initial weights, enhancing the model's ability to escape local minima and converge more efficiently to a global solution.

Furthermore, the RL component of our model is tailored to assign a higher reward to the correct classification of the minority class, effectively shifting the model's focus towards these critical predictions. This is a substantial improvement over traditional supervised learning methods, which may not have sufficient representative data to train effectively across all classes. RL's adaptability in the learning policy allows for a more balanced exploration of the decision space, leading to strategies that favor the accurate classification of underrepresented classes. This adaptive nature of RL in our model sets it apart from existing methods, equipping it to overcome the inherent limitations faced by conventional classification approaches in the context of EEW systems.

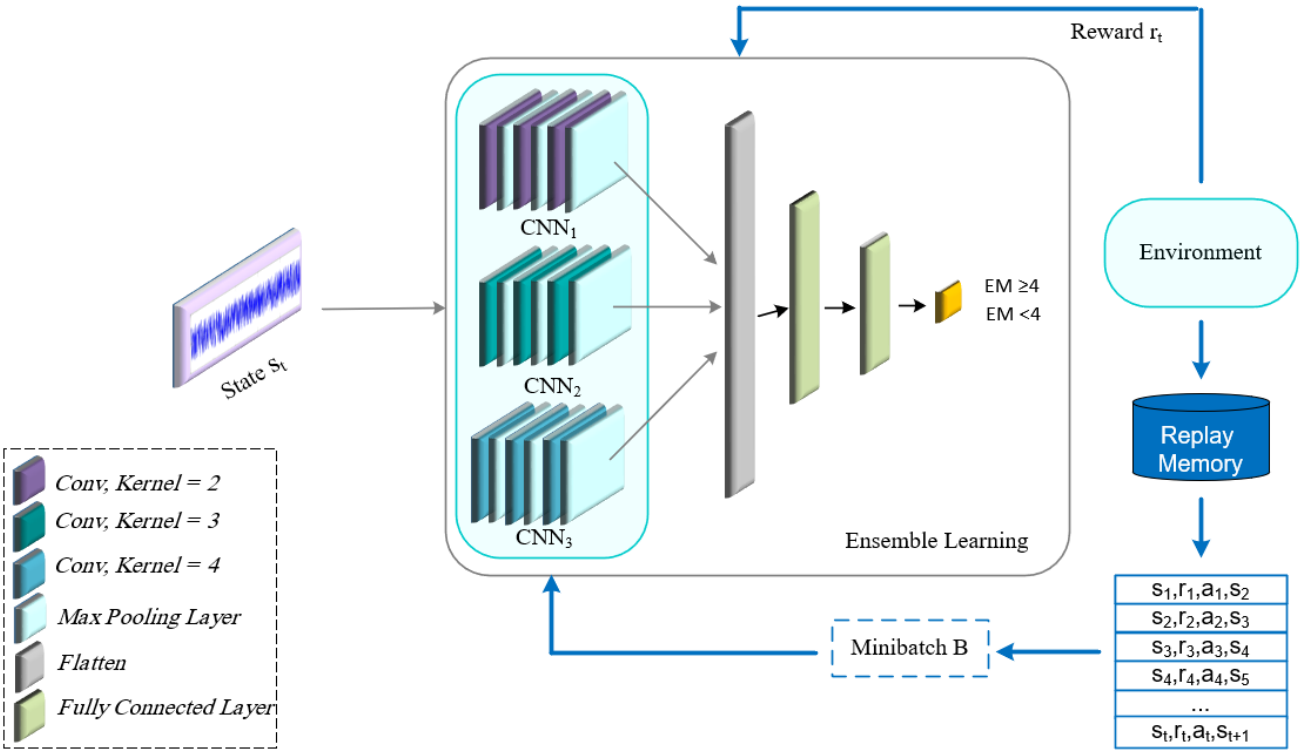


Fig. 1. The proposed model.

A. Pre-training Phase

The accuracy of network weights' initial values holds significant importance for deep models. If these initial settings are inaccurate, their repercussions can cascade throughout the model's training process, potentially leading to challenges related to convergence. Acknowledging the pivotal role of weight initialization, the initial phase of this study embarks on establishing the optimal configurations for both the CNN and the feed-forward neural network. To address the predicament of weight initialization, we introduce a groundbreaking approach called the ML-ABC technique. This strategy seeks to revolutionize the conventional process of determining initial values for network weights. In typical scenarios, the selection of initial weights relies on various methods, often incorporating heuristic or random techniques. However, the ML-ABC method introduces a higher level of intelligence into this process. By combining the principles of artificial bee colony optimization with the concept of mutual learning, the ML-ABC approach draws inspiration from the cooperative behaviors observed in natural bee colonies. This innovative approach aims to generate optimal initial weight seeds by enabling the network to learn and evolve collaboratively through the exchange of insights and knowledge. This facilitates a more informed and efficient weight initialization process.

1) The ML-ABC algorithm: The ABC algorithm, drawing inspiration from the intricate behaviors of honeybees when foraging for food, mirrors the collective intelligence and intricate actions observed in nature. This algorithm provides a methodical and intuitive approach to solving optimization problems. At its core, the ABC algorithm comprises four essential components:

a) Worker Bees: Functioning as the initial explorers, these bees venture into a designated area or target zone to identify potential food sources. Their actions are not random but are guided by their knowledge of previous sources, evaluating the quantity and quality of nectar available. After gathering information, they return to the hive to share their findings.

b) Observer Bees: Situated within the hive, observer bees absorb and analyze the information brought by the worker bees. Their decisions are based on shared information, particularly the quality and quantity of nectar discovered. If a worker bee's dance (used to communicate food source information) is compelling and suggests a rich source, observer bees are influenced to visit that location. This dance-based decision-making process ensures the swift exploitation of promising food sources.

c) Scout Bees: These bees act as adventurers, stepping in when a food source becomes depleted or no longer viable. Rather than relying on dance-based information, scouts explore the surroundings to find fresh food sources in a stochastic manner. Their role ensures the hive's adaptability and resilience to changing conditions.

d) Food Sources: Representing potential solutions to the optimization problem, each food source possesses a nectar quantity that reflects the quality or fitness of the associated solution. The collective aim of the worker bees, observer bees, and scouts is to maximize nectar accumulation, analogous to seeking optimal solutions in a computational context.

The brilliance of the ABC algorithm lies in its adaptability. By mimicking the foraging behaviors of honeybees, it achieves a balance between exploring new solutions and exploiting known ones. This intricate interplay of roles and responsibilities, guided by nature-inspired principles, positions the ABC algorithm as a robust tool in the realm of optimization [35].

Eq. (1) elucidates the process of generating a new position, utilizing spatial information from the worker bee. If the nectar found at the new position surpasses the quality of the previous location, the bee will adopt the new position, abandoning the former one. Conversely, if the quality of the nectar is subpar, the bee retains the memory of its previous location.

$$v_i^j = s_i^j + \varphi_i^j (x_i^j - x_k^j) \tag{1}$$

The equation employs the index i to represent the i-th position, where each solution s_i comprises a set of D parameters. The parameter D signifies the count of parameters subjected to optimization, while k signifies an alternative random solution distinct from i. The variable φ_i^j is randomly chosen from the interval [0, 1]. Introducing a change to one parameter of s_i generates a novel solution v_i, which may differ from the original one.

During an optimization process in "D" dimensions, a dimension is randomly selected, and its value is adjusted, with the fitness value determining the superior outcome in each iteration. As indicated in Eq. (1), the novelty and irregularity of the fresh food source v_i^j stem from its dependence on two variables, s_i^j and s_k^j.

To create a food source with enhanced fitness, we leverage knowledge derived from mutual learning between the present and adjacent food sources. This aspect is pivotal in the ABC algorithm, as it necessitates a food source with a heightened fitness value [36].

$$v_i^j = \begin{cases} s_i^j + \varphi_i^j (s_k^j - s_i^j), & Fit_i < Fit_k \\ s_k^j + \varphi_i^j (s_i^j - s_k^j), & Fit_i \geq Fit_k \end{cases} \tag{2}$$

where, Fit_i and Fit_k denote the fitness values of the neighboring and current food sources, respectively. The variable φ_i^j is a uniformly distributed random integer ranging from 0 to F, where F represents a positive mutual learning factor. Novel solutions enhance their fitness by evaluating nearby and current food sources and gravitating toward superior choices. If the current food source offers better fitness, the candidate solution aligns with it; otherwise, the solution moves toward a neighboring source. The stability of food positions hinges on the value of F, which must be a non-negative positive number yielding improved solution.

As the value of F incrementally rises from zero to a specific threshold, the disturbances observed in the corresponding point diminish. This signifies that the fitness value of the alternative food source is converging to, or

closely approaching, the higher fitness value. However, an elevated F value disrupts the delicate balance between exploration and exploitation. In order to illuminate the potential solution in the ML-ABC algorithm, our research's encoding technique vectorizes the CNN and feed-forward weights. It is challenging to provide precise weights, but through several attempts, we devised a method of encoding that strives for the utmost accuracy. An example of encoding a feed-forward network with three hidden layers and a CNN network with three layers, each layer comprising three filters, is shown in Fig. 2. The array's weight matrices are all referenced as rows, and this is crucial information to remember.

For the purpose of determining how effectively a solution fits into the developed DE algorithm, the fitness factor is constructed as follows:

$$Fitness = \frac{1}{1+\sum_{i=0}^{N}(y_i-\tilde{y}_i)^2} \quad (3)$$

The y_i and $\tilde{y}_i$ represent the target and projected labels, respectively, for the i-th dataset, while *N* indicates the total number of cases.

B. Deep Reinforcement Learning

DRL presents a robust approach within the realm of deep learning, where an agent dynamically engages with its environment to optimize rewards. This dynamic learning process equips the agent to navigate uncertain situations and make a sequence of decisions, proving particularly valuable in diverse domains like robotics, healthcare, and finance [37]. DRL's competence in handling tasks necessitating sequential decisions and its adaptability to unpredictable scenarios underscore its versatile, practical applicability. A primary challenge in tasks involving categorization arises from imbalanced datasets, where one category significantly outweighs others. This imbalance can lead to biased learning, as conventional categorization methods tend to prioritize the dominant category heavily, consequently hindering the recognition of less prominent ones. In such cases, DRL emerges as a more efficacious strategy for training neural networks compared to conventional methods. DRL tackles the uneven categorization issue by incorporating a reward-based system. By judiciously assigning rewards, the agent's attention can be steered towards instances from less prevalent categories, thereby enhancing the accurate identification of these infrequent groups. This incentive-driven model establishes a comprehensive decision-making approach, giving prominence to the detection and categorization of rare incidents or less common categories.

In the domain of deep Q-learning, the agent's objective is centered on selecting actions that maximize the potential rewards in forthcoming situations. The rewards expected in future scenarios, denoted by the reward value, decrease over time with the discount rate γ, as depicted in Eq. (3). In this context, T represents the final time step of an episode.

$$R_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'} \quad (4)$$

Q-values stand for the measure of state-action interactions' effectiveness and symbolize the predicted result of policy π when action a is taken in state s. This calculation is illustrated in Eq. (4).

$$Q^{\pi}(s,a) = E[R_t|s_t = s, a_t = a, \pi] \quad (5)$$

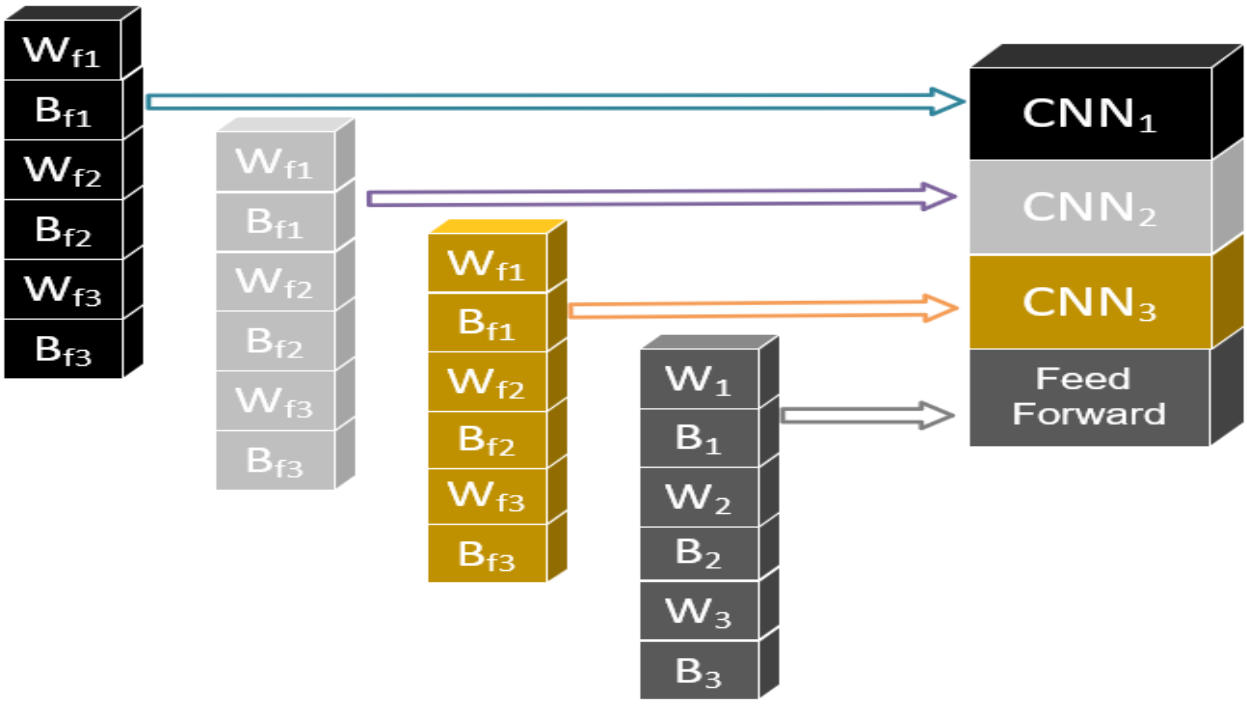


Fig. 2. Strategy for encoding employed in the algorithm being suggested.

The action-value function that yields the maximum expected reward among all strategies following the observation of state s and the execution of action a is computed as illustrated in Eq. (5).

$$Q^*(s,a) = max_\pi\, E[R_t|s_t = s, a_t = a, \pi] \quad (6)$$

The function applies the Bellman equation [38], which asserts that the highest expected result for a specific action is the combination of the current action's rewards and the maximum expected outcome from subsequent actions in the following step. This idea is illustrated in Eq. (6).

$$Q^*(s,a) = E[r + \gamma\, max_{a'} Q^*(s',a')|s_t = s, a_t = a] \quad (7)$$

The calculation of the optimal action-value function is conducted in a step-by-step manner utilizing the Bellman equation, as demonstrated in Eq. (7).

$$Q_{i+1}(s,a) = E[r + \gamma\, max_{a'} Q_i(s',a')|s_t = s, a_t = a] \quad (8)$$

During the training process, as the network encounters state s, it generates a corresponding action for that state. Subsequently, the environment provides a reward r and transitions to the next state s'. These variables are combined into a tuple (s, a, r, s'), which is then stored in memory M. Collections of these tuples, referred to as Batches B, are selected to perform gradient descent. The formula to compute the loss is described as shown in Eq. (8).

$$L_i(\theta_i) = \sum_{(s,a,r,s') \in B} (y - Q(s,a;\theta_i))^2 \quad (9)$$

In this context, θ represents the weights of the model, while y represents the estimated target for the Q function. This estimation involves summing the reward associated with the state-action combination and the discounted highest Q value in subsequent time steps, as demonstrated in Equation 9.

$$y = r + \gamma\, max_{a'} Q(s', a';\ \theta_{k-1}) \quad (10)$$

It is worth highlighting that the Q value assigned to the terminal state is set to 0. The magnitude of the gradient for the loss function at iteration i can be computed using Eq. (10).

$$\nabla_{\theta_i} L(\theta_i) = -2 \sum_{(s,a,r,s') \in B} (y - Q(s,a;\theta_i)) \nabla_{\theta_i} Q(s,a;\theta_i) \quad (11)$$

Through executing a gradient descent iteration on the loss function, the model's weights are updated as per Eq. (11). This adjustment aims to reduce the disparity, with α representing the learning rate that governs the degree of progress in the optimization process.

$$\theta_{i+1} = \theta_i + \alpha \nabla_{\theta_i} Q(s,a;\theta_i) \quad (12)$$

In this paper, our emphasis lies in applying the RL-driven algorithm in the realm of EEW. The following explanation elucidates the functioning of the methodology and imparts comprehension of each constituent element:

- State $\boldsymbol{s_t}$: This corresponds to the image observed at the time step t.
- Action $\boldsymbol{a_t}$: The classification performed on the image is treated as an action. This represents a decision made by the network based on its existing understanding of the goal.
- Reward $\boldsymbol{r_t}$: A reward is provided for each classification, designed to guide the network toward precise categorization. The formulation of this reward mechanism is presented as follows:

$$r_t(s_t, a_t, y_t) = \begin{cases} +1\,, a_t = y_t \text{ and } s_t \in D_O \\ -1\,, a_t \neq y_t \text{ and } s_t \in D_O \\ \lambda\,, a_t = y_t \text{ and } s_t \in D_N \\ -\lambda\,, a_t \neq y_t \text{ and } s_t \in D_N \end{cases} \quad (13)$$

Here, D_O and D_N respectively refer to the majority and minority classes. Accurately/incorrectly categorizing a sample from the majority class results in a gain/loss of $+\lambda/-\lambda$. The proposed method directs the network towards prioritizing the precise classification of instances from the rarer class, assigning a higher absolute value as a reward. Simultaneously, incorporating the majority class and a versatile reward parameter within the interval of $0 < \lambda < 1$ introduces intricacy to the reward structure, enabling precise tuning of the network's emphasis between the more prevalent and less common classes.

IV. Empirical Evaluation

A. Dataset

The data utilized in this investigation consists of three-component waveform data sampled at a frequency of 100 Hz. These data were collected by CENC from 1104 stations located in China over the time span from 2009 to 2017 [32]. As seismic waves propagate, their energy weakens as the distance of propagation increases. Consequently, the characteristics of waveforms tend to deteriorate. To address this, we have chosen seismic waveforms with epicentral distances shorter than 200 km. These waveforms might display disruptions and overlaps due to issues like network communication anomalies and equipment malfunctions. Therefore, we employ the "obspy" library to reprocess the data in this study. The composition of the training and test sets is presented in Table I.

TABLE I. Waveform Distributions for the Unbalanced Dataset Presented by CENC

Magnitude range	Waveforms in the CENC dataset
1-2	214963
2-3	91206
3-4	14080
≥4	4017

B. Metrics

We employ various metrics such as Accuracy, F-measure, and G-means, which are defined as follows:

$$\text{Accuracy} = \frac{TP+TN}{Total\ number\ of\ samples} \quad (14)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (15)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (16)$$

$$Specificity = \frac{TN}{TN+FP} \quad (17)$$

$$\text{G} - \text{means} = \sqrt{\text{Recall} \times Specificity} \quad (18)$$

Here, TP represents the true positives, or the count of actual positives accurately identified by the model. TN denotes the true negatives, referring to the actual negatives that the model correctly predicts. FP stands for false positives, which are the actual negatives that the model erroneously labels as positive. Lastly, FN indicates the false negatives, meaning the actual positives that the model incorrectly classifies as negative.

C. Comparator Models

During the evaluation phase, a comprehensive comparative analysis was conducted to assess the performance of our proposed model in comparison to five distinct deep learning frameworks: SeisNet [25], EEWMagNet [32], MagEstNet [39], QuakeClassNet [40], QuakeNet [41]. This rigorous examination aimed to offer a holistic perspective on our model's efficacy in relation to contemporary methodologies. To further explore various adaptations of our proposed model, two modified versions were introduced into the analysis. The first variant, referred to as "Proposed with RL," maintained an architectural design akin to our original model while excluding the application of the ML-ABC technique. Conversely, the second variant, "Proposed with ML-ABC," omitted the use of the RL technique for classification. To thoroughly evaluate the performance of these models, established metrics were employed, with a specific focus on metrics like the F-measure and geometric mean, known for their suitability in cases of skewed data distributions.

D. Results

The outcomes of this evaluation, as detailed in Table II, distinctly showcase the superior performance of our proposed model when compared to all other contenders, even including well-established models such as EEWMagNet and SeisNet. Across all evaluation criteria, our model consistently demonstrated its superiority over its counterparts. Remarkably, the model achieved substantial reductions in errors, showcasing improvements of 25% and 10% in the F-measure and G-means metrics, respectively. These noteworthy enhancements underscore the model's ability to effectively address challenges posed by imbalanced data distributions and its proficiency in delivering more accurate predictions.

A particularly illuminating comparison emerges when we consider the model's modified versions - "Proposed with RL" and "Proposed with ML-ABC". The advantages inherent in the integration of ML-ABC with the RL approach become evident. Our primary model achieved a substantial reduction in errors, nearly 30%, relative to its modified versions. This emphasizes the pivotal roles of both ML-ABC and RL in augmenting the model's performance, underscoring their significance in advancing the development of cutting-edge deep learning systems.

TABLE II. PERFORMANCE METRICS OF THE PROPOSED MODEL VERSUS COMPARATOR DEEP MODELS

	Accuracy	**F-measure**	**G-means**
EEWMagNet	0.8914±0.1405	0.8622±0.2001	0.8515±0.0104
SeisNet	0.8510±0.1553	0.7471±0.1054	0.8002±0.0402
MagEstNet	0.8105±0.1404	0.7302±0.1404	0.8006±0.1201
QuakeClassNet	0.7901±0.1006	0.5502±0.0953	0.6505±0.2502
QuakeNet	0.6703±0.1205	0.6702±0.2103	0.7503±0.0051
Proposed with RL	0.8104 ± 0.0627	0.7918 ± 0.1623	0.8303 ± 0.2622
Proposed with ML-ABC	0.8615 ± 0.0243	0.8506 ± 0.0517	0.8609 ± 0.0921
Proposed	0.9217 ± 0.0384	0.8814 ± 0.0297	0.9066 ± 0.0423

Table III presents a comparative analysis of the performance metrics for various deep learning models, including the proposed model, across two classes: EM <4 and EM ≥4. In the EM <4 class, the proposed model exhibits a superior performance with an accuracy of 0.9012±0.0450, indicating its robustness in this category. This performance is notably higher than that of EEWMagNet and SeisNet, which have accuracies of 0.8714±0.1210 and 0.8616±0.1020, respectively. MagEstNet and QuakeClassNet show moderate performance, while QuakeNet trails with the lowest accuracy in this class. For the EM ≥4 class, the proposed model again leads with an impressive accuracy of 0.9415±0.0056, significantly outperforming all comparator models. EEWMagNet, with an accuracy of 0.9012±0.0255, follows as the second most accurate model. SeisNet and MagEstNet show relatively similar performances, while QuakeClassNet and QuakeNet lag behind, indicating potential challenges in accurately classifying instances in this higher EM category. The consistency in the leading performance of the proposed model across both classes underscores its effectiveness and

reliability in binary classification tasks within this context. This superiority is particularly notable in the EM ≥4 class, where the margin of its lead suggests a higher degree of precision and confidence in its predictive capabilities. These results highlight the potential of the proposed model as a robust tool for classification in the studied domain.

Fig. 3 depicts the receiver operating characteristic (ROC) curves for the techniques outlined in Table I. The primary metric employed for assessing the effectiveness of classifiers is the area under the curve (AUC). An AUC value of 1 signifies perfect differentiation, while a value of 0.5 indicates performance equivalent to random chance. Among the various models, SeismoNet stands out with an impressive AUC score of 0.81. This remarkable result showcases its exceptional ability to distinguish between positive and negative outcomes, providing further evidence of the method's robust predictive prowess.

In contrast, both EEWMagNet and SeisNet demonstrated average AUC scores of 0.69 and 0.51, respectively, falling short of the performance achieved by our proposed model. The remaining models, including QuakeNet, QuakeClassNet, and QuakeNet, yielded less favorable results, exhibiting AUC scores ranging from 0.48 to 0.51. Notably, the performance of the QuakeNet model was notably lower, achieving a meager AUC of 0.48, just slightly surpassing the threshold for random prediction. The ROC curves vividly illustrate the disparities in performance across the evaluated methods, with our proposed model showcasing superior performance in discriminating between different outcomes.

Fig. 4 illustrates the error progression observed in the proposed model over a span of 500 epochs. Commencing with an initial error measurement of 10, a consistent downward trajectory becomes evident as the epochs advance. This sustained decline in error validates the model's ability to adapt and enhance its predictive capabilities with the accumulation of training iterations.

It is noteworthy that the most significant reduction in error occurs during the earlier epochs, gradually tapering off as the epochs accumulate. This pattern suggests diminishing returns in terms of error reduction as training prolongs, particularly after approximately the 425th epoch. Around this point, the error stabilizes, maintaining a relatively constant value of approximately 4.2962 across subsequent epochs. This plateau implies that continuing training beyond this juncture might not yield substantial performance improvements, implying the likelihood of the model reaching convergence.

Furthermore, this stability could serve as an indicator of potential overfitting, especially if there is no subsequent improvement observed on external validation or testing datasets. This underscores the importance of monitoring the model's behavior and performance to strike the right balance between training duration and preventing overfitting.

TABLE III. PERFORMANCE METRICS OF THE PROPOSED MODEL VERSUS COMPARATOR DEEP MODELS BY CLASS

Model / Class	EEWMagNet	SeisNet	MagEstNet	QuakeClassNet	QuakeNet	Proposed
EM <4	0.8714±0.1210	0.8616±0.1020	0.7914±0.2356	0.8015±0.1148	0.7014±0.1263	0.9012 ± 0.0450
EM ≥4	0.9012±0.0255	0.8423±0.1026	0.8216±0.1462	0.7726±0.1065	0.6425±0.1105	0.9415± 0.0056

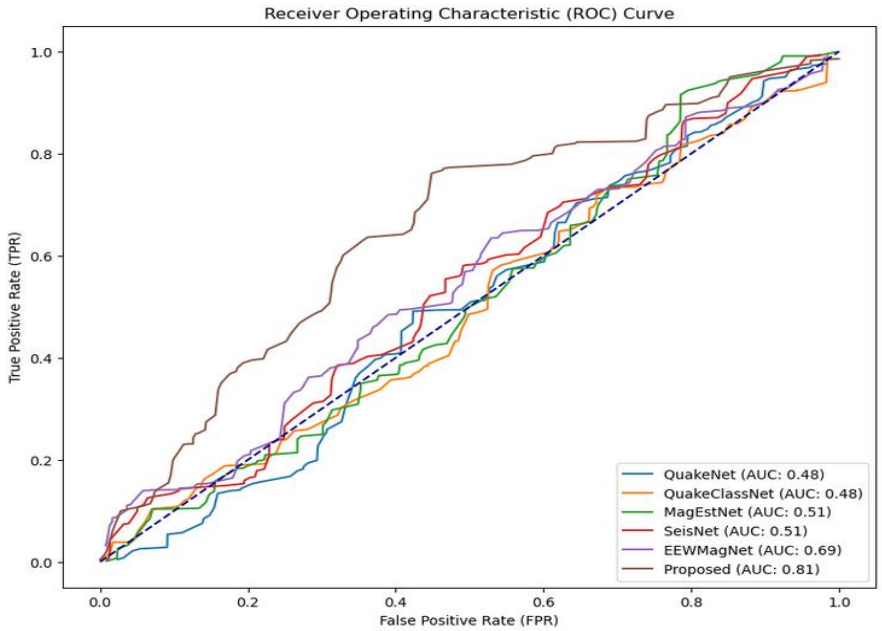


Fig. 3. The ROC chart illustrates the ROC curves for the proposed model and other benchmark techniques. The blue dashed line on the graph represents the ROC curve corresponding to random guessing.

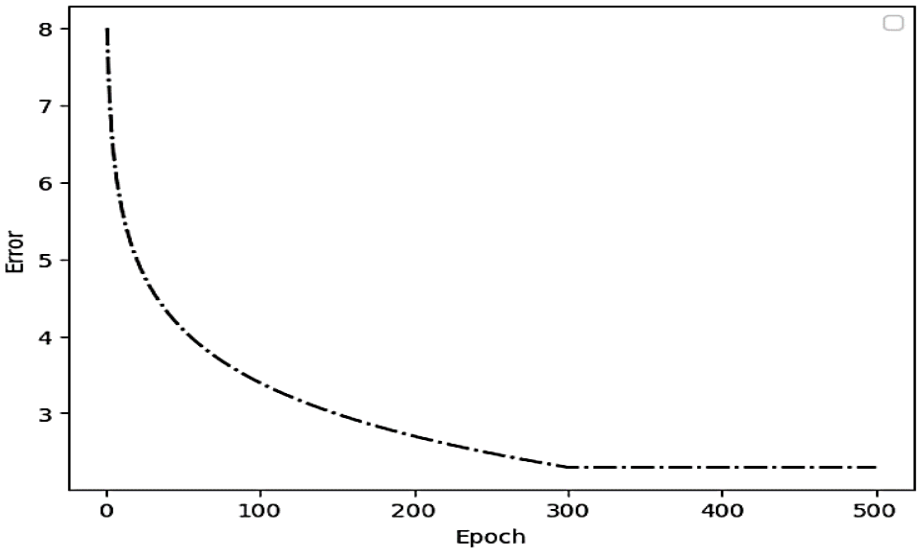


Fig. 4. Diagram illustrating the comparative dynamics of errors.

1) Impact of the reward function: Within our devised model, we introduced a reward mechanism to handle class imbalances effectively. In the case of the majority class, correct predictions were rewarded with a score of +1, while incorrect ones incurred a penalty of -1. For the minority class, accurate predictions garnered rewards of +λ, while incorrect ones resulted in penalties of -λ. The specific value of λ was adjusted based on the ratio between majority and minority samples. Notably, as this ratio increased, the optimal λ tended to decrease.

To explore the impact of λ, we thoroughly analyzed the model's performance across a spectrum of λ values ranging from 0 to 1, incrementing by intervals of 0.1. Throughout this analysis, the reward for the minority class remained consistent. The findings are presented in Fig. 5. At λ = 0, the influence of the majority class was minimal, while at λ = 1, both classes carried equal significance.

The data underscores that the model's most optimal performance is achieved when λ is set to 0.6, according to all evaluation criteria. This indicates that the most suitable value for λ resides between the extremes of 0 and 1. It is essential to recognize that while adjusting λ to diminish the dominance of the majority class is essential, setting it too low can potentially undermine the overall model performance.

This investigation highlights the critical role that the choice of λ plays in shaping the model's outcomes. Determining the ideal value of λ hinges on the relative frequency of majority and minority samples, emphasizing the significance of meticulous parameter adjustments to attain optimal result.

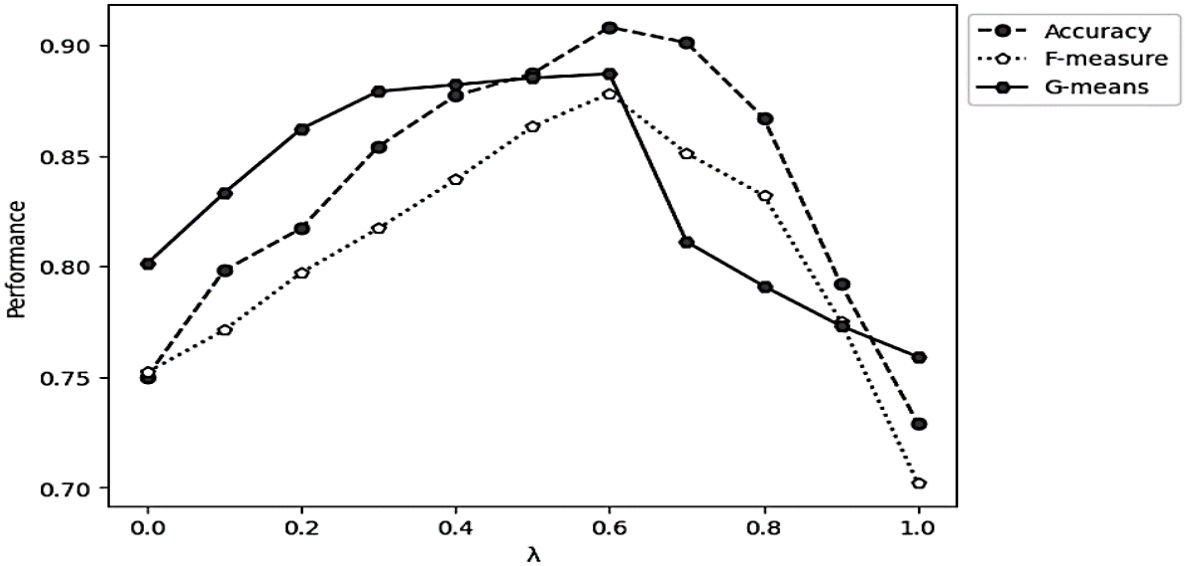


Fig. 5. Generating graphs illustrating the performance metrics of the proposed model as they vary in response to different λ values incorporated within the reward function.

2) Impact of the loss function: A wide array of strategies is available to tackle data imbalances within the realm of machine learning [42]. These encompass enhancements in data augmentation methodologies as well as the careful selection of an appropriate loss function. The pivotal role of the chosen loss function in enabling the model to capture the intricacies of the underrepresented class comprehensively cannot be overstated. Our investigation delved into the effectiveness of five distinct loss functions: Weighted cross-entropy (WCE) [43], balanced cross-entropy (BCE)] [44], Dice loss (DL) [45], Tversky loss (TL) [46], and Combo Loss (CL) [47].

While both WCE and BCE commonly utilized loss functions that treat positive and negative instances with equal weight, their performance may falter when faced with datasets exhibiting significant imbalances that heavily favor the minority class. On the contrary, DL and TL emerge as more suitable choices for datasets characterized by pronounced imbalances, showcasing notably improved results in relation to the minority class. Among these, CL takes a distinctive position as an exceedingly effective loss function, particularly tailored to datasets characterized by skewed distributions. Through judicious adjustments of the loss function's weights, CL elevates the significance of intricate samples, thereby conferring them a greater influence than straightforward ones.

Our meticulous analysis of these loss functions has been meticulously detailed in Table IV, yielding illuminating insights. The outcomes of this study underscore the exceptional performance of CL in comparison to TL, leading to an impressive 13% reduction in error rate accuracy and a substantial 25% enhancement in the F-measure, a pivotal metric for model evaluation. However, it is imperative to note that, despite its remarkable achievements, CL falls short by 20% when measured against our proposed model, which has been intricately designed to tackle binary classification challenges. This underscores the context-specific nature of model design and the need to tailor solutions to the particular problem at hand.

TABLE IV. COMPARISON OF THE PERFORMANCE METRICS BETWEEN THE SUGGESTED MODEL AND THE EMPLOYED LOSS FUNCTIONS

	Accuracy	F-measure	G-means
WCE	0.7432± 0.1132	0.7215± 0.0130	0.7510± 0.1255
BCE	0.8001± 0.1021	0.7622± 0.0102	0.8025± 0.0101
DL	0.8042± 0.2033	0.7992± 0.0106	0.826± 0.1000
TL	0.8320± 0.1203	0.8148± 0.0436	0.8450± 0.2062
CL	0.8630± 0.0247	0.8440± 0.0152	0.8639± 0.0152

3) Impact of the number of CNNs: The proposed architecture employs multiple CNNs to derive feature vectors from input images concurrently. The quantity of these CNN-based feature extractors significantly influences the model's performance. Utilizing an insufficient number of CNNs can lead to subpar feature extraction, while an excessive number might result in overfitting or the inclusion of redundant information. Both these extremes have the potential to undermine the efficacy of the model. To identify the ideal number of CNNs, an evaluation of the proposed model was conducted across a range of one to seven CNN feature extractors.

The outcomes of this investigation reveal that the model attains its peak performance when equipped with three CNNs, as clearly depicted in Fig. 6. Interestingly, the model's performance exhibited a decline when six and seven CNNs were employed. Strikingly, these cases even yielded results inferior to those obtained using a solitary CNN. This finding underscores the delicate balance required in selecting the number of CNNs for optimal performance, where an excess can prove detrimental rather than beneficial.

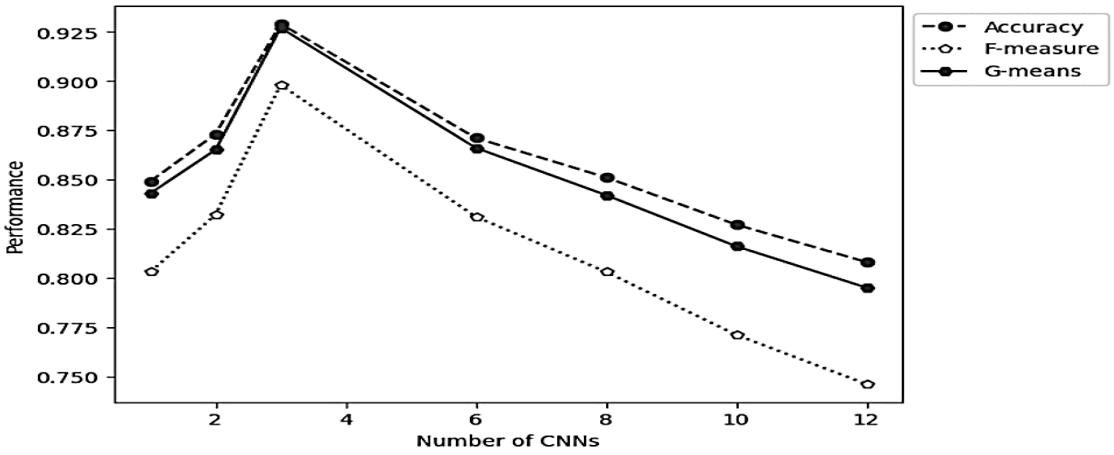


Fig. 6. Graphing the performance measurements of the suggested model concerning changes in the count of convolutional feature extractors.

4) Discussion: The study unveiled an innovative EEW model that was specifically designed to classify earthquake magnitudes rapidly. This classification was achieved through the analysis of a 7-second, three-component seismic waveform record acquired from the CENC. This novel model architecture featured a distinctive incorporation of three concurrent dilated convolution layers. These layers played a pivotal role in extracting crucial feature vectors that were subsequently employed for effective classification tasks.

Addressing the inherent challenges tied to data imbalances often encountered in conventional EEW models, the research integrated an algorithm based on RL. This strategic approach aimed to enhance the accuracy of sample classifications by providing appropriate rewards to the model. In the training process, individual samples were visualized as distinct states within a sequence of interconnected decision points. Functioning as an agent, the neural network received rewards or penalties based on its performance in accurately distinguishing between various classes.

To initiate the weight pre-training of the model, a unique approach known as ML-ABC was introduced. This pioneering technique dynamically adjusted the optimal "food source" for candidate solutions, encompassing components of mutual learning that were intricately linked to the starting weights. This innovative initialization process contributed to the ability of the model to learn and adapt during subsequent training phases.

The model proposal, crafted using seismic waveform records obtained from the CENC, exhibits potential limitations when attempting to extend its applicability to various global datasets. Its fundamental design, intricately linked to the distinct seismic characteristics unique to the CENC, raises substantial concerns regarding its performance when applied to regions such as California, renowned for the San Andreas Fault, or Japan, a significant participant in the Pacific Ring of Fire [48]. This potential challenge unfolds in two aspects: firstly, the model might internalize and overly emphasize the particular attributes of the CENC dataset, and secondly, it could inadvertently absorb any biases inherently present within that dataset. Such an excessively specialized adaptation could significantly hinder the capacity of the model to adapt to broader, more diverse contexts, impeding its generalization capabilities [32].

Furthermore, while the algorithms incorporated within the proposed model, such as RL and dilated convolution layers have been meticulously fine-tuned to accommodate CENC data, their flexibility when applied to unfamiliar datasets, remains questionable. A common issue observed with specialized algorithms is their susceptibility to overfitting on the training data, rendering them less adaptable and responsive when presented with previously unseen data. Take the RL algorithm, for instance, which has been structured with reward mechanisms tailored to the unique attributes of the CENC dataset [49]. This design might lead to challenges when confronted with datasets that possess distinct seismic event distributions. Similarly, the dilated convolution layers, optimized specifically to decipher intricate patterns inherent in CENC data, could either overlook or misinterpret vital features originating from divergent geographic regions [50].

The complexities inherent in seismic data are multifaceted, and the presence of noise or anomalies can wield substantial influence over the accuracy and dependability of any predictive model [51]. In regions marked by dense urbanization or significant industrial undertakings, the vibrations stemming from everyday human activities like traffic movement, construction endeavors, or even the operations of underground subway systems can introduce supplementary signals into the seismic data. Likewise, natural geophysical disturbances such as volcanic eruptions, landslides, or even robust winds and oceanic waves can engender disturbances that might be misconstrued as seismic activity by a model that hasn't been primed to contend with such factors [52].

The susceptibility of the proposed model to such extraneous signals takes on heightened concern when pondering the repercussions of erroneous alerts or missed predictions. The occurrence of false alarms, where the model inaccurately forecasts a seismic event due to noise, holds the potential to induce unwarranted panic and economic disruptions and erode public trust in the reliability of the system. Conversely, if the model fails to identify a genuine seismic event owing to its confusion with noise, the ramifications could be gravely severe, encompassing loss of lives and property.

Furthermore, managing noise isn't merely a matter of filtering out undesired signals; it also necessitates the ability to discriminate between noise and legitimate seismic activities that might exhibit similar characteristics. The challenge lies in striking a delicate equilibrium wherein the model remains attuned to authentic seismic events while discarding irrelevant data. Achieving this equilibrium mandates the training of the model on diverse datasets that encapsulate a broad spectrum of noise scenarios and authentic seismic patterns.

Moreover, it is imperative to recognize that the refinement and fine-tuning of the model should remain an ongoing process. This is particularly significant due to the dynamic nature of noise sources, which can evolve over time. As cities expand, infrastructure undertakings progress and the landscape of human activity evolves, the nature of noise embedded within seismic data will inevitably undergo transformations. This dynamic characteristic underscores the necessity for the model to be adaptable rather than static, prepared to accommodate and surmount the challenges that emerge from evolving noise profiles [53].

V. CONCLUSION

This manuscript introduces SeismoNet, an innovative EEW model that capitalizes on a 7-second, three-component seismic waveform record sourced from the CENC. The architecture of the model is characterized by the inclusion of three concurrent dilated convolution layers, which play a pivotal role in the extraction of feature vectors. These vectors are subsequently amalgamated to facilitate the classification process. In response to the challenges stemming from data imbalances within the realm of EEW models, an RL-based

algorithm has been seamlessly integrated. This algorithm rewards the model agent more prominently for precise sample classifications, thereby enhancing its learning process.

The training mechanism of the model is visualized as a sequence of interconnected decision points, with each individual sample treated as a discrete state. Within this framework, the neural network assumes the role of an agent, and its performance in distinguishing between the minority and majority classes is met with corresponding rewards or penalties.

For the initial weight pre-training of the model, a groundbreaking ML-ABC technique is introduced. This method exhibits dynamic adjustments to the optimal "food source" for candidate solutions, ingeniously incorporating mutual learning components that are intricately linked to the initial weights. This innovative initialization process lends the model adaptability and enhances its capacity to learn and adapt during subsequent training phases.

A comprehensive suite of experiments has been conducted using the designated dataset to discern optimal parameter values, including the formulation of the reward function. The outcomes of these experiments resoundingly underscore the efficacy of our proposed methodology when juxtaposed with alternative approaches that were explored within the confines of this study. This affirmation further solidifies the viability and potential of the proposed model as a robust solution for Earthquake Early Warning systems.

ACKNOWLEDGMENT

This study is sponsored by the Key projects of natural science research in Anhui Colleges and Universities: “Research on seismic behavior of vertical irregular steel frame structures based on new composite dampers” (2022AH052653)

This research is sponsored by horizontal topics at the school level: "Research on Energy Dissipation and Damping Technology of Building Structures"(CZZY-HX-2023-08)

REFERENCES

[1] M. R. Jenkins, S. K. McBride, M. Morgoch, and H. Smith, "Considerations for creating equitable and inclusive communication campaigns associated with ShakeAlert, the earthquake early warning system for the West Coast of the USA," Disaster Prevention and Management: An International Journal, vol. 31, no. 1, pp. 79-91, 2022.

[2] E. Zuccolo, G. Cremen, and C. Galasso, "Comparing the performance of regional earthquake early warning algorithms in Europe," Frontiers in Earth Science, vol. 9, p. 686272, 2021.

[3] Y. Nakamura, "On the urgent earthquake detection and alarm system (UrEDAS)," in Proc. of the 9th World Conference on Earthquake Engineering, 1988, vol. 7, pp. 673-678.

[4] R. M. Allen and H. Kanamori, "The potential for earthquake early warning in southern California," Science, vol. 300, no. 5620, pp. 786-789, 2003.

[5] H. Kanamori, "Real-time seismology and earthquake damage mitigation," Annu. Rev. Earth Planet. Sci., vol. 33, pp. 195-214, 2005.

[6] A. Zollo et al., "Earthquake early warning system in southern Italy: Methodologies and performance evaluation," Geophysical research letters, vol. 36, no. 5, 2009.

[7] S. V. Moravvej, S. J. Mousavirad, M. H. Moghadam, and M. Saadatmand, "An LSTM-based plagiarism detection via attention mechanism and a population-based approach for pre-training parameters with imbalanced classes," in Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part III 28, 2021: Springer, pp. 690-701.

[8] M. Soleimani, Z. Forouzanfar, M. Soltani, and M. J. Harandi, "Imbalanced Multiclass Medical Data Classification based on Learning Automata and Neural Network," EAI Endorsed Transactions on AI and Robotics, vol. 2, 2023.

[9] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," Expert systems with applications, vol. 73, pp. 220-239, 2017.

[10] S. Moravvej, M. Maleki Kahaki, M. Salimi Sartakhti, and M. Joodaki, "Efficient GAN-based method for extractive summarization," Journal of Electrical and Computer Engineering Innovations (JECEI), vol. 10, no. 2, pp. 287-298, 2022.

[11] S. V. Moravvej, S. J. Mousavirad, D. Oliva, and F. Mohammadi, "A Novel Plagiarism Detection Approach Combining BERT-based Word Embedding, Attention-based LSTMs and an Improved Differential Evolution Algorithm," arXiv preprint arXiv:2305.02374, 2023.

[12] M. S. Sartakhti, M. J. M. Kahaki, S. V. Moravvej, M. javadi Joortani, and A. Bagheri, "Persian language model based on BiLSTM model on COVID-19 corpus," in 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA), 2021: IEEE, pp. 1-5.

[13] L. Hong et al., "GAN-LSTM-3D: An efficient method for lung tumour 3D reconstruction enhanced by attention-based LSTM," CAAI Transactions on Intelligence Technology, 2023.

[14] S. V. Moravvej, M. J. M. Kahaki, M. S. Sartakhti, and A. Mirzaei, "A method based on attention mechanism using bidirectional long-short term memory (BLSTM) for question answering," in 2021 29th Iranian Conference on Electrical Engineering (ICEE), 2021: IEEE, pp. 460-464.

[15] S. V. Moravvej, M. Joodaki, M. J. M. Kahaki, and M. S. Sartakhti, "A method based on an attention mechanism to measure the similarity of two sentences," in 2021 7th International Conference on Web Research (ICWR), 2021: IEEE, pp. 238-242.

[16] S. V. Moravvej, A. Mirzaei, and M. Safayani, "Biomedical text summarization using conditional generative adversarial network (CGAN)," arXiv preprint arXiv:2110.11870, 2021.

[17] M. Marani, M. Soltani, M. Bahadori, M. Soleimani, and A. Moshayedi, "The Role of Biometric in Banking: A Review," EAI Endorsed Transactions on AI and Robotics, vol. 2, no. 1, 2023.

[18] M. Bahadori, M. Soltani, M. Soleimani, and M. Bahadori, "Statistical Modeling in Healthcare: Shaping the Future of Medical Research and Healthcare Delivery," in AI and IoT-Based Technologies for Precision Medicine: IGI Global, 2023, pp. 431-446.

[19] S. Vakilian, S. V. Moravvej, and A. Fanian, "Using the cuckoo algorithm to optimizing the response time and energy consumption cost of fog nodes by considering collaboration in the fog layer," in 2021 5th International Conference on Internet of Things and Applications (IoT), 2021: IEEE, pp. 1-5.

[20] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," Applied soft computing, vol. 8, no. 1, pp. 687-697, 2008.

[21] S. V. Moravvej, S. J. Mousavirad, D. Oliva, G. Schaefer, and Z. Sobhaninia, "An improved de algorithm to optimise the learning process of a bert-based plagiarism detection model," in 2022 IEEE Congress on Evolutionary Computation (CEC), 2022: IEEE, pp. 1-7.

[22] S. Vakilian, S. V. Moravvej, and A. Fanian, "Using the artificial bee colony (ABC) algorithm in collaboration with the fog nodes in the Internet of Things three-layer architecture," in 2021 29th Iranian Conference on Electrical Engineering (ICEE), 2021: IEEE, pp. 509-513.

[23] P. Saeid, M. Pazoki, and M. Zeinolabedini, "Optimization of biomass production from sugar bagasse in anaerobic digestion using genetic algorithm," Modeling Earth Systems and Environment, vol. 9, no. 2, pp. 2183-2198, 2023.

[24] S. V. Moravvej et al., "RLMD-PA: A reinforcement learning-based myocarditis diagnosis combined with a population-based algorithm for

pretraining weights," Contrast Media & Molecular Imaging, vol. 2022, 2022.

[25] T. Ren et al., "Seismic severity estimation using convolutional neural network for earthquake early warning," Geophysical Journal International, vol. 234, no. 2, pp. 1355-1362, 2023.

[26] C. Y. Wang, T. C. Huang, and Y. M. Wu, "Using LSTM neural networks for onsite earthquake early warning," Seismological Society of America, vol. 93, no. 2A, pp. 814-826, 2022.

[27] A. Hu and H. Zhang, "Application of machine learning to magnitude estimation in earthquake emergency prediction system," Chinese Journal of Geophysics, vol. 63, no. 7, pp. 2617-2626, 2020.

[28] Y. Wang, Q. Zhao, K. Qian, Z. Wang, Z. Cao, and J. Wang, "Cumulative absolute velocity prediction for earthquake early warning with deep learning," Computer-Aided Civil and Infrastructure Engineering, 2023.

[29] A. Datta, D. J. Wu, W. Zhu, M. Cai, and W. L. Ellsworth, "DeepShake: Shaking intensity prediction using deep spatiotemporal RNNs for earthquake early warning," Seismological Society of America, vol. 93, no. 3, pp. 1636-1649, 2022.

[30] C. Kavitha and V. Gnanadesigan, "Predicting Earthquake Measurements using Deep Learning," in 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 2023: IEEE, pp. 1-7.

[31] Y. Wang, X. Li, Z. Wang, and J. Liu, "Deep learning for magnitude prediction in earthquake early warning," Gondwana Research, 2022.

[32] F. Meng, T. Ren, Z. Liu, and Z. Zhong, "Toward earthquake early warning: A convolutional neural network for repaid earthquake magnitude estimation," Artificial Intelligence in Geosciences, vol. 4, pp. 39-46, 2023.

[33] B. Lin et al., "The research of earthquake magnitude determination based on Convolutional Neural Networks," Chinese Journal of Geophysics, vol. 64, no. 10, pp. 3600-3611, 2021.

[34] J. Münchmeyer, D. Bindi, U. Leser, and F. Tilmann, "Earthquake magnitude and location estimation from real time seismic waveforms with a transformer network," Geophysical Journal International, vol. 226, no. 2, pp. 1086-1104, 2021.

[35] H. Zareiamand, A. Darroudi, I. Mohammadi, S. V. Moravvej, S. Danaei, and R. Alizadehsani, "Cardiac Magnetic Resonance Imaging (CMRI) Applications in Patients with Chest Pain in the Emergency Department: A Narrative Review," Diagnostics, vol. 13, no. 16, p. 2667, 2023.

[36] S. Danaei et al., "Myocarditis Diagnosis: A Method using Mutual Learning-Based ABC and Reinforcement Learning," in 2022 IEEE 22nd International Symposium on Computational Intelligence and Informatics and 8th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics (CINTI-MACRo), 2022: IEEE, pp. 000265-000270.

[37] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, "An introduction to deep reinforcement learning," Foundations and Trends® in Machine Learning, vol. 11, no. 3-4, pp. 219-354, 2018.

[38] E. Barron and H. Ishii, "The Bellman equation for minimizing the maximum cost," NONLINEAR ANAL. THEORY METHODS APPLIC., vol. 13, no. 9, pp. 1067-1090, 1989.

[39] J. Zhu, S. Li, and J. Song, "Magnitude estimation for earthquake early warning with multiple parameter inputs and a support vector machine," Seismological Research Letters, vol. 93, no. 1, pp. 126-136, 2022.

[40] M. A. Meier et al., "Reliable real-time seismic signal/noise discrimination with machine learning," Journal of Geophysical Research: Solid Earth, vol. 124, no. 1, pp. 788-800, 2019.

[41] J. Huang, X. Wang, Y. Zhao, C. Xin, and H. Xiang, "LARGE EARTHQUAKE MAGNITUDE PREDICTION IN TAIWAN BASED ON DEEP LEARNING NEURAL NETWORK," Neural Network World, no. 2, 2018.

[42] H. Gharagozlou, J. Mohammadzadeh, A. Bastanfard, and S. S. Ghidary, "RLAS-BIABC: A reinforcement learning-based answer selection using the bert model boosted by an improved ABC algorithm," Computational Intelligence and Neuroscience, vol. 2022, 2022.

[43] Ö. Özdemir and E. B. Sönmez, "Weighted cross-entropy for unbalanced data with application on covid x-ray images," in 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 2020: IEEE, pp. 1-6.

[44] F. Huang, J. Li, and X. Zhu, "Balanced Symmetric Cross Entropy for Large Scale Imbalanced and Noisy Data," arXiv preprint arXiv:2007.01618, 2020.

[45] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced NLP tasks," arXiv preprint arXiv:1911.02855, 2019.

[46] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in Machine Learning in Medical Imaging: 8th International Workshop, MLMI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Proceedings 8, 2017: Springer, pp. 379-387.

[47] S. A. Taghanaki et al., "Combo loss: Handling input and output imbalance in multi-organ segmentation," Computerized Medical Imaging and Graphics, vol. 75, pp. 24-33, 2019.

[48] N. A. Pambudi, "Geothermal power generation in Indonesia, a country within the ring of fire: Current status, future development and policy," Renewable and Sustainable Energy Reviews, vol. 81, pp. 2893-2901, 2018.

[49] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," IEEE access, vol. 7, pp. 53040-53065, 2019.

[50] Y. Qiu, C. Xu, Z. Xiao, and J. Wang, "Seismogenic Structure of the 2017 M s 6.6 Jinghe, China, Earthquake Inferred from Seismic Detection and Relocation," Seismological Society of America, vol. 93, no. 5, pp. 2612-2624, 2022.

[51] T. Jordan et al., "Operational Earthquake Forecasting: State of Knowledge and Guidelines for Implementation," Annals of Geophysics, 2011.

[52] G. C. Beroza, M. Segou, and S. Mostafa Mousavi, "Machine learning and earthquake forecasting—next steps," Nature communications, vol. 12, no. 1, p. 4761, 2021.

[53] W. Marzocchi and J. D. Zechar, "Earthquake forecasting and earthquake prediction: different approaches for obtaining the best model," Seismological Research Letters, vol. 82, no. 3, pp. 442-448, 2011.

Information Retrieval System for Scientific Publications of Lampung University by using VSM, K-Means, and LSA

Rahman Taufik, Didik Kurniawan*, Anie Rose Irawati, Dewi Asiah Shofiana
Dept. of Computer Science, University of Lampung, Bandar Lampung, Indonesia

***Abstract*—The Lampung University repository system is a repository of data related to study, community service, and other scientific works, currently has 37242 documents accessible through repository.lppm.unila.ac.id. Despite the amount of data, its optimal use as an information retrieval remains unrealized, hindering the effective promotion of Lampung University's scientific publication excellence. Recognizing the limitations of existing information retrieval systems that are limited to specific methods for topic identification through clustering, this study aims to develop a retrieval system for Lampung University's repository using Vector Space Model (VSM), K-Means and Latent Semantic Analysis (LSA) that generates clusters and study expertise at the level of study program, faculty and Lampung University. The methodology includes data collection, preprocessing, modeling, evaluation and system deployment. The results show that the number of clusters obtained for the university level is 7 clusters, for the faculty level are 6, 7, 8 and 10 clusters, and for the program level are 3 to 5 clusters. In addition, the finding topic identification indicate that the expertise topics at Lampung University, which are agriculture, soil, education, plants, learning, society, Lampung. This study contributes to optimizing the information retrieval system, promoting academic excellence, and advancing the understanding of study expertise at Lampung University.**

Keywords*—Information retrieval; Vector Space Model (VSM); k-means; Latent Semantic Analysis (LSA); clustering; topic identification; scientific publication information

I. INTRODUCTION

The Lampung University repository system is a data repository accessible to all lecturers and the academic community through the address repository.lppm.unila.ac.id. According to Google index data, the publication data already contains 37242 documents. Additionally, the University of Lampung Repository System is utilized to store lecturer data related to study activities, publications, community service, and other achievements. All data stored in this repository is traceable based on divisions at the University of Lampung, namely faculties, study programs, authors and years of study. The data stored in this repository can serve various purposes, including as a source of information, for literature reviews, accreditation, self-evaluation, institutional achievements, and other *tri dharma*-related purposes.

However, the data in the repository has not been effectively utilized as meaningful information due to the inadequacy of the system in accessing and managing data in an informative manner. Consequently, units within Lampung University encounter difficulties when searching for relevant scientific publications in their areas of expertise. Furthermore, understanding the strength of study fields based on the obtained information is crucial to assess the excellence of study at the University of Lampung.

One of the technologies that can be used to extract information from a repository system is an information retrieval system. Several studies such as Information Retrieval for Digital Library [1], Information Retrieval for Faculty Study Repositories [2], Information Retrieval for Bibliographic Control and Institutional Repositories [3], Multilingual Information Retrieval [4], have proven that information retrieval systems can effectively collect and determine information from various repositories. In study [1], they improved the existing digital library system, namely Sowiport, by integrating heterogeneous databases through an information retrieval system approach. As a result, they obtained 513,000 data entries from 25 different thesauri, enhancing keyword search capabilities. In addition to digital library system [1], there are other studies related to educational repositories which are studies [2][3]. The study [2] aims to evaluate faculty study repositories used in higher education institutions. One aspect of the evaluation is to identify relevant articles from nine academic databases using an information retrieval system approach. The results indicate that the evaluated and redesigned system improves the preservation of scientific study results. While study [3] improves the quality of their data, metadata, and semantic data through an information retrieval system approach, the results support their control over the bibliography, including a repository that is rich and clean. In terms of diversity of information data, compared to other studies that collect data in various languages through information retrieval systems [4], they successfully gathered over 77 thousand Wikipedia queries in 18 languages. This supports improvements in data retrieval across multiple languages.

*Corresponding Author

The implementation of an information retrieval system must be accompanied by the utilization of techniques. The follow-up regarding the application of techniques can be observed in studies [5] [6] [7]. Study [5] employs data mining techniques to enhance information retrieval within their digital institutional repository, yielding personalized profiles specific to each user group in their repository. In study [6], an information retrieval system and data mining techniques were employed, specifically an Expertise Recommender System. This approach was utilized to develop a prototype system aimed at identifying and recommending thesis advisors for specific subjects. Unfortunately, specific data mining methods used in studies [5] [6] are not mentioned. On the other hand, study [7] developed an information retrieval system using hybrid deep fuzzy hashing algorithm to address issues related to similarity measurement procedures. Experimental result showed that the proposed model achieved higher retrieval accuracy compared to conventional models, although it is limited in feature selection.

Additionally, other approaches, such as Natural Language Processing (NLP), can be employed, as demonstrated in study [8] [9]. Study [8] developed an information retrieval system using NLP techniques to address employee queries based on knowledge from their internal site. The results showed that the queries were successfully responded to with relevant answers, although there is potential for further acceleration in the process. Furthermore, study [9] developed a search system by implementing information retrieval system techniques, Jaccard and cosine similarity matrix techniques. These methods were utilized to clean and standardize data and information related to academic documents. The results indicated that the system could be utilized to recommend documents, albeit with limitations in searches based on divergence metrics.

Despite significant advancements in the utilization of information retrieval systems across various domains [1] [2] [3] [4], there appears to be a noticeable gap in explicit discussions or study concerning the optimization of data generated by information retrieval systems. Several studies [5] [6] [7] [8] [9] mention the use of data mining techniques and NLP, but specific methods for topic identification through clustering have not been proposed. The implementation of clustering for topic identification is crucial, particularly in the context of academic data.

The present study aims to develop an information retrieval system for scientific publications of Lampung University using the Vector Space Model (VSM), K-Means and Latent Semantic Analysis (LSA). We propose VSM to obtain the similarity in vector representations, K-Means to cluster study topics, while LSA is used to generate the identification of study topics. The use of Vector Space Model (VSM) method can improve the information obtained from information retrieval systems. A number of studies have developed information retrieval system that applied VSM to determine information [10] [11]. The study in [10] proposed the combination of two methods, namely VSM and description logic for concept extraction, which increases the level of similarity between documents and certain concepts in information retrieval systems. While, the study [11] proposed VSM based on TF-IDF weights and word vectors to calculate semantic similarity for implicit citation detection problem. On the other hand, studies such as [12] [13] show that K-means method can be proposed for clustering scientific publication documents. The study [12] proposes a multi-verse optimizer and k-means algorithm for extracting topics from clustered documents. K-means can be used especially in cases where the dataset is normally distributed such as scientific publication documents [13]. Furthermore, regarding LSA, study [14] proposed a method for information retrieval systems using LSA method to retrieve important information from questions asked by users or mass documents, and the results showed a positive contribution. Therefore, the selection of VSM, K-means, and LSA is proposed based on a literature review [10] [11] [12] [13] [14], given their effectiveness in addressing the specific aspects required for the study, the use of alternative methods may not align with our objectives.

The objective of this study is to develop an information retrieval system that generates clusters and study expertise at the level of study program, faculty and Lampung University. The study is driven by the following three research questions: (1) How to develop the information retrieval system for scientific publications using the Vector Space Model (VSM), K-Means and Latent Semantic Analysis (LSA)? (2) How many clusters are formed based on data generated by the information retrieval system? (3) What is the most dominant study topics based on the results of the information retrieval system? Answers to these study questions are provided and discussed in the remaining sections of this paper, particularly in the results and discussion section. We aim to contribute to the ongoing discourse on information retrieval system development and provide practical insights on VSM, K-Means and LSA implementation to determine study expertise.

II. STUDY METHODOLOGY

In developing the information retrieval system (see Fig. 1), various stages are proposed, which include data collection, preprocessing, modeling, evaluation, and system distribution. Initially, 37242 scientific publication documents from lecturers were collected from the repository of the Institute for Study and Community Service (LPPM), University of Lampung. After collection, the data was preprocessed by keyword extraction, cleaning, language detection, translation, stemming, and stopword removal. The data was then modeled to achieve clustering and classification of study topics. The evaluation was done using clustering test methods such as Silhouette score [15], Calinski-Harabasz score [16], and Davies-Bouldin score [17]. Finally, the obtained study data on clusters and scientific publication expertise were distributed as meaningful information for Lampung University.

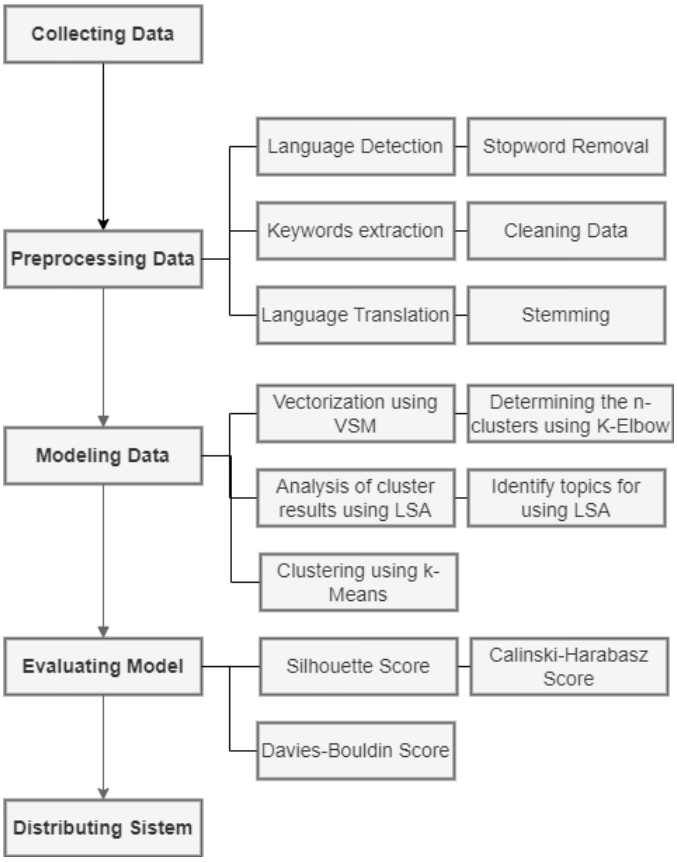


Fig. 1. Stages of information retrieval system development.

III. Results and Discussion

A. Data Collection

The first stage in the development of the Lampung University Lecturer Publication Information Retrieval System is literature study and data collection. Data collection is done by first creating a program using the Python programming language which can be seen in Fig. 2. This program was created to retrieve raw data from the Lampung University's LPPM repository (http://repository.lppm.unila.ac.id/) into collection data. The rest, this program is used to collect data from each URL that has been registered, clean the data collection and convert it into csv.

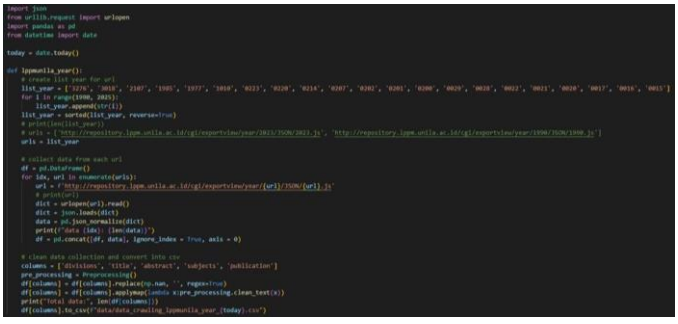


Fig. 2. Code of the study publication data collection program.

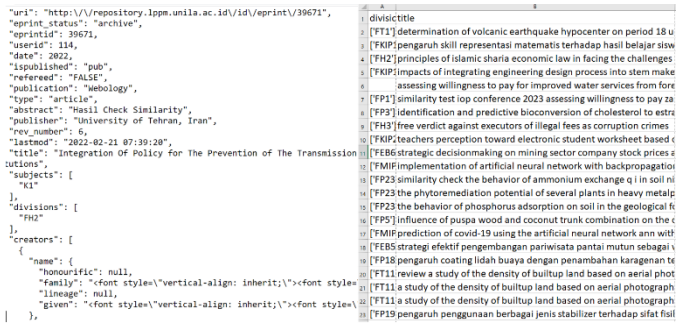


Fig. 3. Data collection of study publication.

Using the above program, 37242 data were retrieved. These data are scientific publication data of Lampung University lecturers, including division, title, abstract, subjects, publication, publisher, date, keywords. Fig. 3 shows raw data from Lampung University's LPPM repository and collection data in the form of .csv files processed by the program.

B. Data Preprocessing

The next step is data preprocessing. This stage aims to clean the collection data prepared for data modeling. This stage includes several activities, including keyword extraction, data cleaning, language detection, language translation, stemming, and stopword removal. These preprocessing stages are performed sequentially.

The collection data obtained from the Lampung University's LPPM repository does not have good consistency, some data have keywords and some do not, so the first preprocessing stage performed is the separation of keywords from abstract data using the Python library and regex. After the keywords data are obtained, the title and abstract data are cleaned by separating some unimportant characters and symbols. In addition to inconsistencies related to keywords, the collection data obtained is also inconsistent regarding language, therefore the next stage is to detect the language, and if there is data in English then it is translated into Indonesian. After all the data is in Indonesian, the next data is stemming which converts a word to its root word by removing the phrase prefix [18]. The last preprocessing stage performed is stopword removal; this stage is performed to remove unimportant words. The stopword removal can be seen in Fig. 4. Furthermore, the preprocessing results of keywords, abstract and title data are combined into a feature that is ready to be used for data modeling.

Fig. 4. The stopword removal to remove unimportant words..

C. Data Modeling

The next stage is data modeling which is the core of the development of this information retrieval system. Data modeling in this study is aimed at obtaining clustering and topic classification from study data using various machine learning methods. The modeling stages include text vectorization using the VSM model, determining the number of clusters using the k-elbow method, clustering using the k-means method, visual analysis of cluster results using the Umap library and the LSA method, and topic classification using the LSA method. This data modeling stage is performed on three different types of data, first for university level data, second for faculty data and third for study program data.

The collection data are preprocessed and then converted into vector form using VSM. One of the VSM methods used is TF-IDF (Term Frequency - Inverse Document Frequency). TF-IDF produces a vector value based on measuring the authenticity of a word by comparing the number of occurrences of a word in a document with the number of occurrences of a document containing that word. The vector results obtained from the University of Lampung study data have vector dimensions (37242, 3094428). The result of the vector is then used to determine the number of clusters and the clustering process, k-elbow and k-means methods are proposed in this case. In the results, the number of clusters obtained for the university level is 7 clusters, while for the faculty level is 6, 7, 8 and 10 clusters, while for the study program level is 3 to 5 clusters. The details for the faculty and study program levels can be seen in Tables I and II. In addition, the results of clustering at the university level can be seen visually in Fig. 5.

TABLE I. FACULTY-LEVEL CLUSTERING INFORMATION

Faculty	Total Data	Cluster Number	Topic
FKIP (Faculty of Teacher Training and Education)	4710	8	student, learning, teacher, education, teaching, Lampung, language,class
FMIPA (Faculty of Mathematics and Natural Sciences)	5042	7	plant, orchid, plant, extract, method, acid, virus
FEB (Faculty of Economics and Business)	2207	8	business, work, Lampung, indonesia, employee, consumer, tax,money
FT (Faculty of Engineering)	4401	6	land, water, city, oil, material, earth
FP (Faculty of Agriculture)	9022	10	plant, forest, agriculture, fertilizer, tree, food, water, plant, food,Lampung
FISIP (faculty of Social Sciences and Political Sciences)	2912	7	lampung, village, community, travel, tourism, group, work, district
FK (Faculty of Medical)	6569	8	patient, health, mind, development, behavior, school
FH (Faculty of Law)	2466	7	law, custom, community, protected, Lampung, tax, data

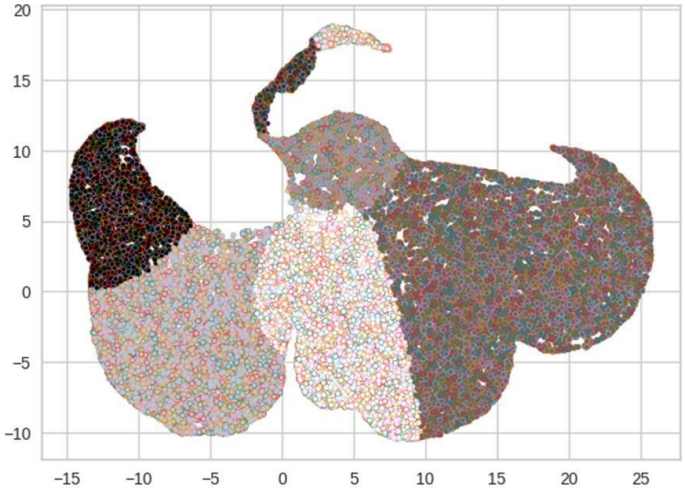


Fig. 5. Visualization of University-level clustering.

In addition to clustering, the results of this study also generate a set of topic identification from each clustering obtained. These topics identifications are generated using the LSA method. For each topic identification from the clusters obtained, then a topic is selected that is very relevant in the context of the topic in manually, for example from the Faculty of Agriculture 2 clusters are produced, the first cluster has classified topics including study, results, farming, while the second cluster is soil, processing, farming, then from the first cluster the topic taken is farming and the second cluster the topic taken is soil. This is because the topic represents the cluster; in addition to the words farming and soil have a context related to the study topics of the Faculty of Agriculture. The results of topic classification for the university level include agriculture, soil, education, plants, learning, society, Lampung. As for faculties and study programs can be seen in Table I and Table II.

D. Data Evaluation

The next stage is model evaluation. The model evaluation stage is used to test the quality of the clustering results of the proposed model. Several methods are used to obtain the evaluation results, including Silhouette score, Calinski-Harabasz score, and Davies Bouldin score. The Silhouette score with a value close to 1 means that the data point is in the correct cluster, close to -1 means that the data points is in wrong cluster [15]. While in Calinski-Harabasz calculations, the higher the value, the better the grouping is [16]. On the other hand, for the Davies Bouldin score, the smaller the value produced, the more optimal the clustering model [17]. The clustering model evaluation results obtained for the university level are 0.53 for the Silhouette score, 107298.36 for the Calinski Harabasz score, and 0.60 for the Davies Boulding score. The plot of these scores shows that the cluster model formed is quite good. Meanwhile, the results of the clustering model evaluation for the faculty and study program levels can be seen in Table III and Table IV.

TABLE II. STUDY PROGRAM-LEVEL CLUSTERING INFORMATION

Faculty	ID Study Program	Study Program	Total Data	Cluster Number	Topic
FEB	FEB1	Master of Accounting ScienceProgram	17	5	managerial, organization, service, government, audit
FEB	FEB2	Master of Economics Program	35	3	economy, employment,decentralization
FEB	FEB3	Master of Management Program	72	3	investment, loyalty, funding
FEB	FEB4	Accounting Study Program	731	4	finance, work, management,tax, financial
FEB	FEB5	Development Economics StudyProgram	465	5	economy, umkm, labor,investment, finance
FEB	FEB6	Management Study Program	899	4	ownership, consumer, production, business, work
FH	FH1	Doctoral Program in Law	880	3	law, politics, authority
FH	FH2	Master of Law Program	316	4	property rights, law, certificate,law
FH	FH3	Law Study Program	1271	4	customary law, rights, tax,village
FISIP	FISIP1	Business Administration Program	319	5	business, quality, banking,management, work
FISIP	FISIP2	State Administration Program	980	5	tourism, development, village,policy, program
FISIP	FISIP3	Government Science Program	621	4	government, politics,corruption, facilities
FISIP	FISIP4	Communication Study Program	379	5	communication, information, regression, behavior, media
FISIP	FISIP5	Master of Administrative ScienceProgram	13	4	management, work, fluctuation,policy
FISIP	FISIP6	Master of Government ScienceProgram	125	5	organization, culture, politics,conflict, strategy
FISIP	FISIP7	Sociology Study Program	356	4	society, conflict, behavior,crime
FISIP	FISIP8	International Relations Program	124	4	cooperation, country, export,relationship
FK	FK1	Medical Education Study Program	1860	4	patient, disease, treatment,doctor
FKIP	FKIP1	English Language Program	2202	4	students, classes, teachingtechniques, learning
FKIP	FKIP10	Master's Program in SocialStudies Education	96	4	development, teaching, social,student
FKIP	FKIP11	Master's Program in EducationalTechnology	177	4	students, learning model,motivation, education
FKIP	FKIP12	Civics Study Program	161	4	character, nation, education,culture
FKIP	FKIP14	Indonesian and Regional Language and Literature Education Study Program	174	5	language, speech, society,teaching, indonesia
FKIP	FKIP15	Biology Education Program	271	4	students, skills, science,learning model
FKIP	FKIP16	Social Studies EconomicsEducation Program	159	3	students, economy,entrepreneurship
FKIP	FKIP17	Physics Education Program	386	5	students, physics, skills, development, learning model
FKIP	FKIP18	Social Studies GeographyEducation Program	191	4	student, soil, factor, landslide
FKIP	FKIP19	Elementary School Teacher Education Study Program (PGSD)	143	4	students, education, development, teaching
FKIP	FKIP2	Guidance and Counseling StudyProgram	1386	4	students, learning model,concentration, education
FKIP	FKIP20	Physical Education, Health andRecreation	54	5	student, sport, ball, athlete,learning model
FKIP	FKIP21	Chemistry Education Program Management	476	4	students, experiment, teaching,learning model
FKIP	FKIP22	Mathematics Education Program	265	4	students, math, diagram,learning model
FKIP	FKIP23	PG-PAUD Education Program	183	4	child, development, social,behavior
FKIP	FKIP24	Social Studies History EducationProgram	159	3	history, students, education
FKIP	FKIP25	Drama, Dance and MusicEducation Program	150	5	music, dance, art, performance,works
FKIP	FKIP3	Master's Program in ElementaryTeacher Education	111	4	empirical, student, development, teaching
FKIP	FKIP4	Master of Science TeacherTraining Program	169	4	students, systems, learningmodels, skills
FKIP	FKIP5	Master's Program in Education	150	4	school, leadership, management, student
FKIP	FKIP6	Master's Program in EducationManagement	20	3	school, teacher, education
FKIP	FKIP7	Master's Program in Regional Language and Literature Education	62	4	language, semantics, tradition,teaching
FKIP	FKIP8	Master's Program in Indonesian Language and Literature Education	103	4	teaching, language, students,literature
FKIP	FKIP9	Master's Program in PhysicsEducation	508	4	students, learning models,development, skills
FMIPA	FMIPA1	Physics Study Program	415	4	temperature, data, method,concentration
FMIPA	FMIPA10	Information Systems Program	18	4	system, application, information, classification

Faculty	ID Study Program	Study Program	Total Data	Cluster Number	Topic
FMIPA	FMIPA2	Biology Study Program	1724	4	plant, extract, plant, orchid
FMIPA	FMIPA3	Computer Science Program	518	5	system, information, medicine,recommendation
FMIPA	FMIPA4	Chemistry Study Program	1079	4	compound, ion, village,adsorption
FMIPA	FMIPA5	Master's Program in BiologicalSciences	127	4	extract, steroid, taurine, benzo
FMIPA	FMIPA6	Master's Program in PhysicalScience	33	3	sample, temperature, point
FMIPA	FMIPA7	Master's Program in ChemicalScience	75	3	mushroom, silica, synthesis
FMIPA	FMIPA8	Master's Program in MathematicalSciences	46	3	model, probability, test
FMIPA	FMIPA9	Mathematics Study Program	1033	4	method, model, data, students
FP	FP1	Agribusiness Study Program	3758	5	food, community, agriculture,water, forest
FP	FP11	Master of Forestry ScienceProgram	217	4	forest, management, community, diversity
FP	FP12	Master of Environmental ScienceProgram	72	5	health, environment, social,community, ecotourism
FP	FP13	Master's Program in Development Counseling / Community Empowerment	18	3	sanitation, counseling,community
FP	FP14	Master's Program in Agricultural Extension and Communication	12	5	agriculture, village, extension,farmer, citizen
FP	FP15	Master's Program in AgriculturalIndustrial Technology	60	4	processing, coffee, flowers,technology analysis
FP	FP17	Animal Husbandry Study Program	389	4	livestock, livestock products, methods, animal diseases
FP	FP18	Agricultural Engineering Program	973	5	fertilization, moisture content, cropping techniques, agricultural yield, processing
FP	FP19	Agricultural Product TechnologyProgram	784	4	products, processing, agricultural products, processing
FP	FP2	Agro technology Study Program	2941	4	fertilization, agriculture, yieldprocessing, tillage
FP	FP20	Aquatic Resources Program	108	4	fish, water, fry, care
FP	FP21	Marine Science Program	95	4	sea, fish, aquaculture, care
FP	FP22	Agricultural IndustrialTechnology Program	136	3	agricultural products, processing, industrial innovation
FP	FP23	Soil Science Program	516	4	soil, fertilization, tillage,planting method
FP	FP24	Plant Protection Study Program	458	4	plant diseases, care, treatment,prevention
FP	FP25	Agronomy and Horticulture StudyProgram	614	4	crops, fertilization, seeds, pests
FP	FP26	Agricultural Extension Program	507	4	agriculture, extension, farmers,agricultural products
FP	FP27	Animal Nutrition and FeedTechnology Program	97	4	livestock products, nutrition,innovation, processing
FP	FP3	Aquaculture Study Program	375	4	aquaculture, fish, shrimp, care
FP	FP4	Doctoral Program in AgriculturalSciences	12	3	treatments, plant diseases, riskanalysis
FP	FP5	Forestry Study Program	1871	4	forest, community, forestproducts, fauna
FP	FP6	Master of Agribusiness Program	31	3	organization, result analysis,method analysis
FP	FP7	Master Program in Agroecotechnology	17	4	plants, replication, innovationresults, maintenance
FP	FP8	Master Program in Agronomy	32	4	crop, experiment, fertilizerprocessing, development
FT	FT1	Geophysical Engineering Program	1353	4	soil, rock, reservoir, water
FT	FT11	Geodetic Engineering StudyProgram	146	4	land, measurement, mapping,data collection
FT	FT12	D3 Civil Engineering StudyProgram	13	5	settlement, community, social, development, sustainability
FT	FT13	D3 Survey and Mapping StudyProgram	41	4	mapping, survey, community,evaluation
FT	FT14	D3 Mechanical Engineering StudyProgram	34	3	electrical, engine, testing
FT	FT2	Master Program in CivilEngineering	116	4	concrete, measurement, testing,material
FT	FT3	Master's Program in MechanicalEngineering	41	4	mechanism, machine,processing, material
FT	FT4	Civil Engineering Program	843	4	soil, rainfall, drainage, roads
FT	FT5	Electrical Engineering Program	528	4	electrical, voltage, network,sensor
FT	FT6	Chemical Engineering Program	553	3	waste, treatment, processing
FT	FT7	Mechanical Engineering Program	549	5	machinery, materials, temperature regulation, processing, measurement
FT	FT8	Informatics Engineering Program	231	4	information system, development, sensor, network
FT	FT9	Architecture Program	200	4	building, architecture, publicspace, tourism

TABLE III. FACULTY-LEVEL CLUSTERING EVALUATION RESULTS

Faculty	*Silhouette score*	*Calinski Harabasz score*	*Davies Bouldin Score*
FKIP (Faculty of Teacher Training and Education)	0.4109131575	8379.067221	0.7750398559
FMIPA (Faculty of Mathematics and Natural Sciences)	0.4897095986	19386.7335	0.6157010033
FEB (Faculty of Economics and Business)	0.4059269475	4079.690594	0.7504302209
FT (Faculty of Engineering)	0.4865720672	13639.76272	0.5176474339
FP (Faculty of Agriculture)	0.4546237817	13403.42462	0.7251143872
FISIP (faculty of Social Sciences and Political Sciences)	0.5142060504	8154.666587	0.5947357487
FK (Faculty of Medical)	0.4211412604	11129.5431	0.7477032403
FH (Faculty of Law)	0.476002308	7443.539066	0.7927361434

TABLE IV. STUDY PROGRAM-LEVEL CLUSTERING EVALUATION RESULTS

No	Study Program	*Silhouette score*	*Calinski Harabasz score*	*Davies Bouldin Score*
1	Master of Accounting ScienceProgram	0.44474428	26.25916184	0.5797495286
2	Master of Economics Program	0.9106475354	455.880385	0.07888671825
3	Master of Management Program	0.6558252049	195.8234061	0.4961123569
4	Accounting Study Program	0.4686553933	1168.559533	0.6627273443
5	Development Economics StudyProgram	0.5203163725	1013.045096	0.5052247982
6	Management Study Program	0.5370269291	2506.570266	0.5440162171
7	Doctoral Program in Law	0.712811619	3209.953671	0.3392638925
8	Master of Law Program	0.6382238705	716.6067985	0.4373224099
9	Law Study Program	0.4513430856	2772.691672	0.6360769562
10	Business Administration Program	0.4743366388	379.7126064	0.6451603922
11	State Administration Program	0.4971081968	2627.913425	0.6290085573
12	Government Science Program	0.5262378928	1065.168098	0.5217246815
13	Communication Study Program	0.5603440786	647.3738657	0.5011094346
14	Master of Administrative ScienceProgram	0.3540879556	40.03367626	0.7137427753
15	Master of Government ScienceProgram	0.5229085354	290.8637055	0.5319551749
16	Sociology Study Program	0.484876088	739.6636388	0.6474297144
17	International Relations Program	0.5733409428	498.2796745	0.4034342429
18	Medical Education Study Program	0.4992342555	2733.869864	0.5854277671
19	English Language Program	0.4825049699	3808.495565	0.6538096004
20	Master's Program in SocialStudies Education	0.4877305057	246.2186114	0.5998584388
21	Master's Program in EducationalTechnology	0.5661563429	360.5946676	0.5367123514
22	Civics Study Program	0.4759859002	256.7458577	0.5885390139
23	Indonesian and Regional Language and Literature Education Study Program	0.4261095962	292.1032256	0.7184361478
24	Biology Education Program	0.47853607	333.9577153	0.7308542766
25	Social Studies EconomicsEducation Program	0.5330673746	225.2167718	0.5502047002
26	Physics Education Program	0.4742562866	1062.980569	0.6636898044
27	Social Studies GeographyEducation Program	0.4923488311	448.1597782	0.6056313622
28	Elementary School Teacher Education Study Program (PGSD)	0.4217206704	194.523634	0.7536003486
29	Guidance and Counseling StudyProgram	0.5102595257	2887.425063	0.5507603067
30	Physical Education, Health andRecreation	0.6649952221	196.0591115	0.3759244643
31	Chemistry Education Program Management	0.4990151929	1161.093757	0.5890013265
32	Mathematics Education Program	0.5470512319	512.7211607	0.5589964858
33	PG-PAUD Education Program	0.5113909281	317.8699604	0.5416860298
34	Social Studies History EducationProgram	0.5890686005	479.1019235	0.4484316617
35	Drama, Dance and Music Education Program	0.3787565327	158.6606907	0.8101570708
36	Master's Program in Elementary Teacher Education	0.5726408051	244.2289309	0.4642264092
37	Master of Science TeacherTraining Program	0.5083243542	390.5031415	0.569395827
38	Master's Program in Education	0.526775473	410.7632211	0.5189634631
39	Master's Program in EducationManagement	0.7068854737	154.2750311	0.3192405471

No	Study Program	*Silhouette score*	*Calinski Harabasz score*	*Davies Bouldin Score*
40	Master's Program in Regional Language and Literature Education	0.478600304	74.0290282	0.7254216191
41	Master's Program in Indonesian Language and Literature Education	0.5704257078	357.7135453	0.4520157486
42	Master's Program in Physics Education	0.5232653522	855.1331556	0.5248570726
43	Physics Study Program	0.5523061148	1565.782463	0.5020236427
44	Information Systems Program	0.5849994556	126.7535908	0.4350038006
45	Biology Study Program	0.4582402589	4449.41187	0.6466400134
46	Computer Science Program	0.4460617374	1234.057519	0.6570492754
47	Chemistry Study Program	0.5402759459	3901.174183	0.552069023
48	Master's Program in BiologicalSciences	0.7173196489	684.3202966	0.3836404387
49	Master's Program in PhysicalScience	0.6814721065	337.0048772	0.3096593157
50	Master's Program in ChemicalScience	0.7477607198	363.1038902	0.3995511204
51	Master's Program in MathematicalSciences	0.8640376694	382.5221819	0.2228325984
52	Mathematics Study Program	0.5642975938	3243.65023	0.6303855749
53	Agribusiness Study Program	0.5565772378	7450.062786	0.5477881318
54	Master of Forestry ScienceProgram	0.4856209095	235.6277704	0.5907029633
55	Master of Environmental ScienceProgram	0.5318841607	160.8093078	0.5091538986
56	Master's Program in Development Counseling / Community Empowerment	0.8379093817	168.038297	0.2467429631
57	Master's Program in AgriculturalExtension and Communication	0.5355833326	57.33446831	0.380906379
58	Master's Program in AgriculturalIndustrial Technology	0.7088144844	259.8514142	0.3333115891
59	Animal Husbandry Study Program	0.4931679828	515.4441551	0.7053134603
60	Agricultural Engineering Program	0.5249132742	1683.430794	0.5559149788
61	Agricultural Product TechnologyProgram	0.5397172434	4294.245412	0.5226350002
62	Agrotechnology Study Program	0.5088764145	4340.843745	0.6650451961
63	Aquatic Resources Program	0.4743389275	229.4782722	0.5870737257
64	Marine Science Program	0.4465041272	153.5480053	0.7162108935
65	Agricultural IndustrialTechnology Program	0.8567250436	676.9069297	0.2142528295
66	Soil Science Program	0.5467371165	1033.897392	0.5947156204
67	Plant Protection Study Program	0.528859842	663.2001914	0.6909396104
68	Agronomy and Horticulture StudyProgram	0.5051865457	1086.985709	0.5767930031
69	Agricultural Extension Program	0.4592834436	492.188934	0.7261135323
70	Animal Nutrition and FeedTechnology Program	0.5034602914	184.9903997	0.646122724
71	Aquaculture Study Program	0.475145446	404.6190054	0.6790236179
72	Doctoral Program in AgriculturalSciences	0.9449627517	573.7857003	0.05849474941
73	Forestry Study Program	0.4739062901	4254.633823	0.5891520577
74	Master of Agribusiness Program	0.9132297154	315.0793425	0.06293418965
75	Master Program in Agroecotechnology	0.5921873364	280.9192014	0.3916220903
76	Master Program in Agronomy	0.521430067	118.2019792	0.5802355652
77	Geophysical Engineering Program	0.4905781189	5105.186886	0.5259008141
78	Geodetic Engineering StudyProgram	0.4623979648	412.1179729	0.6099492043
79	D3 Civil Engineering StudyProgram	0.6506877125	111.747278	0.3114347
80	D3 Survey and Mapping StudyProgram	0.7639043985	817.3145449	0.3094277958
81	D3 Mechanical Engineering StudyProgram	0.7024992952	122.6001396	0.3374450702
82	Master Program in CivilEngineering	0.6325944622	254.5081069	0.4579902817
83	Master's Program in MechanicalEngineering	0.6359497177	186.8342982	0.333275908
84	Civil Engineering Program	0.5629114663	1678.186894	0.5664636958
85	Electrical Engineering Program	0.4790047524	773.3184727	0.6724525976
86	Chemical Engineering Program	0.6170604178	737.7720447	0.613540253
87	Mechanical Engineering Program	0.4853931879	1343.989841	0.5498384654
88	Informatics Engineering Program	0.6397052667	1076.106831	0.3900090882
89	Architecture Program	0.560260335	339.4545567	0.4816008777

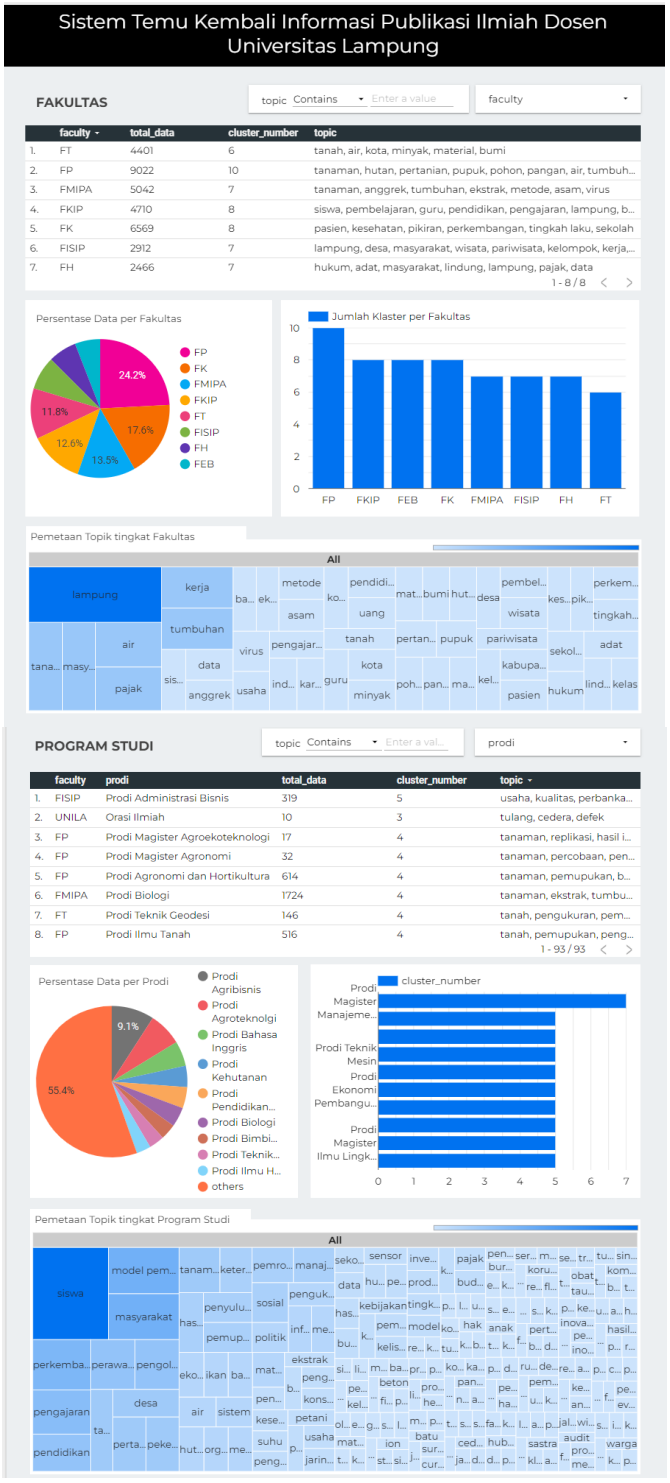


Fig. 6. Dashboard of information retrieval development.

E. Data Distribution

The last stage in the development of this information retrieval system is system distribution. The purpose of this stage is to make it easier for users to access the information generated by the developed information retrieval system.

This system distribution is in the form of a dashboard developed using Google Looker. The information contained in this dashboard consists of general information about this information retrieval system, clustering and classification of topics in the form of detailed data and visualizations. In addition, users can control the menu available on this dashboard and select specific information based on the faculty and study program level. The appearance of this system can be seen in Fig. 6.

F. Analysis Results

From the results of the development of this information retrieval, clustering, topic identification and model evaluation values for each clustering at the University, Faculty and Study Program levels were obtained. For the university level, 7 clusters were obtained with topic classifications including agriculture, soil, education, plants, learning, society, and Lampung. For the faculty level, 6 clusters were obtained for FT, 7 clusters for FMIPA, FISIP, FH, 8 clusters for FKIP, FEB, FK, and 10 clusters for FP.

For the study program level, the most clusters were obtained, namely 5 clusters, this cluster was owned by Master of Government Science study program, Bachelor of Education in Indonesian and Regional Languages and Literature, Bachelor of Education in Drama, Dance and Music, Bachelor of State Administration, Bachelor of Agricultural Engineering, Bachelor of Communication, Bachelor of Agribusiness, Bachelor of Mechanical Engineering, Master of Agricultural Extension and Communication, Diploma in Civil Engineering, Master of Environmental Science, Bachelor of Physical Education, Health and Recreation, Bachelor of Computer Science, Bachelor of Physics. Most of the clusters in this study program belong to study programs with faculties FISIP, FKIP, FT, FMIPA and FP. The amount of data is not always directly proportional to the number of clusters, for example at the faculty level it can be seen that the smallest number of clusters is FT, which is 6, but the amount of data is greater than FEB, FISIP and FT, as well as at the study program level.

Furthermore, the evaluation results of the clustering model obtained for the university level are 0.53 for the Silhouette score, 107298.36 for the Calinski Harabasz score, and 0.60 for the Davies Boulding score. As for the faculty level, the optimal value for Silhouette score is 0.514 which is obtained by FISIP, the optimal value for Calinski Harabasz score is 19386.733 which is obtained by FMIPA, the optimal value of Davies Bouldin score is 0.518 which is obtained by FT. As for the study program level, the optimal value for the Silhouette score is 0.945 which is obtained by the Doctor of Agricultural Science Study Program, the optimal value for the Calinski Harabasz score is 7450.063 which is obtained by the Agribusiness Study Program, the optimal value of the Davies Bouldin score is 0.058 which is obtained by the Doctor of Agricultural Science study program. All study programs obtain the optimal value are come from FP.

In addition, based on the optimal scores obtained from the three clustering evaluation methods used, there are variations in results between faculties and study programs, which indicate the formation of study relationships. Therefore, this study also examined the study relationships between faculties and/or study programs, the results of which can be seen in Table V. FP is the faculty with the highest number of study

relationships, namely 11 relationships. The relationships owned by the Faculty of Agriculture include relationships between study programs in FP, FH, FT, FISIP, and FK.

TABLE V. STUDY RELATIONSHIPS BETWEEN FACULTIES AND STUDY PROGRAMS

Faculty	Faculty relationship	Number of Relationships
FKIP (Faculty of Teacher Training and Education)	[['FKIP1', 'FKIP14'], ['FKIP19', 'FKIP20', 'FKIP21'], ['FKIP23', 'FMIPA3'], ['FK1','FKIP1']]	4
FMIPA (Faculty of Mathematics and Natural Sciences)	[['FMIPA9', 'FT5', 'FT8'], ['FMIPA1', 'FMIPA4'], ['FMIPA6', 'FT6'], ['FISIP1', 'FMIPA5', 'FT5', 'FT8'], ['FKIP23', 'FMIPA3']]]	5
FEB (Faculty of Economics and Business)	[['FEB5', 'FEB6'], ['FEB4', 'FEB6'], ['FEB4', 'FEB5'], ['FEB5', 'FEB2'], ['FEB3', 'FH1','FH3'], ['FEB3', 'FEB6']]]	6
FT (Faculty of Engineering)	[['FMIPA9', 'FT5', 'FT8'], ['FMIPA6', 'FT6'], ['FISIP1', 'FP5', 'FT5', 'FT8', 'FT7'], ['FT9', 'FT12'], ['FISIP1', 'FMIPA5', 'FT5', 'FT8'], ['FP19', 'FP22', 'FT9', 'FT13', 'FT14', 'FT12', 'FT3', 'FT2'], ['FT11', 'FT1'], ['FT11', 'FT4'], ['FT14', 'FP19']]	9
FP (Faculty of Agriculture)	[['FP5', 'FP6', 'FP7', 'FP8', 'FP10', 'FP11', 'FP12', 'FP13', 'FP14', 'FP15', 'FP16', 'FP27', 'FP26', 'FP17', 'FP24', 'FP20', 'FP18'], ['FH1', 'FP1'], ['FP15', 'FP19'], ['FP19', 'FP22'], ['FISIP1', 'FP5', 'FT5', 'FT8', 'FT7'], ['FP1', 'FP25'], ['FP19', 'FP22', 'FT9', 'FT13', 'FT14', 'FT12', 'FT3', 'FT2'], ['FK1', 'FP19', 'FP22'], ['FP1', 'FP25', 'FP2', 'FP24', 'FP19', 'FP22'], ['FP2', 'FP3'], ['FT14', 'FP19'])	11
FISIP (faculty of Social Sciences and Political Sciences)	[['FISIP3', 'FISIP4'],['FISIP1', 'FP5', 'FT5', 'FT8', 'FT7'], ['FISIP1', 'FISIP2', 'FISIP3'], ['FISIP7', 'FK1'], ['FISIP1', 'FMIPA5', 'FT5', 'FT8']]]	5
FK (Faculty of Medical)	[['FKIP1', 'FKIP14'], ['FKIP19', 'FKIP20', 'FKIP21'], ['FISIP7', 'FK1'], ['FKIP23', 'FMIPA3'], ['FK1', 'FP19', 'FP22'], ['FK1', 'FKIP1']]].	6
FH (Faculty of Law)	[['FH1', 'FP1'], ['FEB3', 'FH1', 'FH3']]	2

IV. CONCLUSION

This research aims to develop an information retrieval system that generates clustering and expertise in study fields at the program, faculty, and university levels of Lampung University.

The development stages of this information retrieval system include data collection, data preprocessing, data modeling using VSM, K-Means, and LSA, data evaluation, and system distribution. To address the first research question, VSM is used to obtain text similarity in vector form, K-Means is used for data clustering, and LSA is used to identify study expertise based on the obtained clustering. The analysis results of this information retrieval system development include the number of clusters, the scores from the model evaluation, and the topics identified according to the number of clusters. The obtained cluster numbers for the university level are 7 clusters, while for the faculty level there are 6, 7, 8, and 10 clusters, and for the program level there are 3 to 5 clusters. These cluster numbers answer the second research question. Furthermore, based on the number of clusters, study relationships, and cluster model evaluation scores, study clustering occurs predominantly in programs and faculties of FP, FISIP, FMIPA, FT, and FKIP. Research question number three is addressed by identifying topics based on the number of clusters identified as the most dominant study expertise at Lampung University, including agriculture, soil, education, plants, learning, society, and region Lampung.

Although the topic identification in this research uses Latent Semantic Analysis, the number of identified topics with the number of clusters is still manually selected. Therefore, further research is needed to develop information retrieval systems that can automatically expertise topics based on a set of identification topics. Nevertheless, the development of information retrieval system in this research addresses the needs related to the excellence of study fields at Lampung University.

REFERENCES

[1] Hienert, D., Sawitzki, F., & Mayr, P. (2015). Digital library study in action-supporting information retrieval in sowiport. D-Lib Magazine, 21(3/4), 2015.

[2] Zibani, P., Rajkoomar, M., & Naicker, N. (2022). A systematic review of faculty study repositories at higher education institutions. Digital Library Perspectives, 38(2), 237-248.

[3] Piazzini, T. (2022). Bibliographic control and institutional repositories: welcome to the jungle. Bibliographic control and institutional repositories: welcome to the jungle, 132-142.

[4] Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., & Lin, J. (2022). Making a MIRACL: Multilingual information retrieval across a continuum of languages. arXiv preprint arXiv:2210.09984.

[5] Leticia, T., & Elvis, F. (2014, June). Data mining as a tool for information retrieval in digital institutional repositories. In 3rd International Conference on Computer Science and Service System (pp. 180-183). Atlantis Press.

[6] Angelova, M., Vishnu Manasa, D., Boeva, V., Linde, P., & Lavesson, N. (2018). An Expertise Recommender SystemBased on Data from an Institutional Repository (DiVA). In 22nd edition of the International Conference on ELectronic PUBlishing-Connecting the Knowledge Commons: From Projects to Sustainable Infrastructure, Toronto.

[7] Suma, D. V. (2020). A novel information retrieval system for distributed cloud using hybrid deep fuzzy hashing algorithm. Journal of Information Technology and Digital World, 2(3), 151-160.

[8] Saha, K. K., Ray, S., & Sadhukhan, D. (2022, June). A Lightweight and Precise Information Retrieval System for Organisational Wiki. In International Conference on Frontiers of Intelligent Computing: Theory and Applications (pp. 495-507). Singapore: Springer Nature Singapore.

[9] Vallejo-Huanga, D., Jaime, J., & Andrade, C. (2023, March). Similarity Visualizer Using Natural Language Processing in Academic Documents of the DSpace in Ecuador. In International Conference on Information (pp. 343-359). Cham: Springer Nature Switzerland.

[10] Boukhari, K., & Omri, M. N. (2023). DL-VSM based document indexing approach for information retrieval. Journal of Ambient Intelligence and Humanized Computing, 14(5), 5383-5394.

[11] Malkawi, R., Daradkeh, M., El-Hassan, A., & Petrov, P. (2022). A Semantic Similarity-Based Identification Method for Implicit Citation Functions and Sentiments Information. Information, 13(11), 546.

[12] Abasi, A. K., Khader, A. T., Al-Betar, M. A., Naim, S., Alyasseri, Z. A. A., & Makhadmeh, S. N. (2021). An ensemble topic extraction approach based on optimization clusters using hybrid multi- verse optimizer for scientific publications. Journal of Ambient Intelligence and Humanized Computing, 12, 2765-2801.

[13] Lund, B., & Ma, J. (2021). A review of cluster analysis techniques and their uses in library and information science study: k-means and k-medoids clustering. Performance Measurement and Metrics, 22(3), 161-173.

[14] Joby, D. P. (2020). Expedient information retrieval system for web pages using the natural language modeling. Journal of Artificial Intelligence and Capsule Networks, 2(2), 100-110.

[15] Shahapure, K. R., & Nicholas, C. (2020, October). Cluster quality analysis using silhouette score. In 2020 IEEE 7th international conference on data science and advanced analytics (DSAA) (pp. 747-748). IEEE.

[16] Ashari, I. F., Nugroho, E. D., Baraku, R., Yanda, I. N., & Liwardana, R. (2023). Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta. Journal of Applied Informatics and Computing, 7(1), 95-103.

[17] Ashari, I. F., Banjarnahor, R., Farida, D. R., Aisyah, S. P., Dewi, A. P., & Humaya, N. (2022). Application of data mining with the K-means clustering method and Davies Bouldin index for grouping IMDB movies. Journal of Applied Informatics and Computing, 6(1), 07-15.

[18] Pradana, A. W., & Hayaty, M. (2019). The effect of stemming and removal of stopwords on the accuracy of sentiment analysis on indonesian-language texts. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 375-380.

Adaptive Gray Wolf Optimization Algorithm based on Gompertz Inertia Weight Strategy

Qiuhua Pan
Institute of Mathematical Sciences, Shanghaitech University, Shanghai 201210, China

***Abstract*—To solve the problems that the Gray Wolf Optimizer (GWO) convergence speed is not fast enough and the solution accuracy is not high enough, this paper proposes an Adaptive Gray Wolf Optimizer based on Gompertz inertia weighting strategy (GGWO). GGWO uses the characteristics of the Gompertz function to achieve nonlinear adjustment of the inertia weight, which better balances the speed of global search and accuracy of local search of the GWO algorithm. At the same time, the Gompertz function is used to realize the adaptive adjustment of the individual gray wolf's position and to better update the gray wolves' position according to the fitness values of different gray wolf individuals. Use 6 classic test functions to compare the performance of GGWO in optimization and 10 other classic or improved swarm intelligence algorithms. Results show that GGWO has better solution accuracy, stability, and faster convergence than all other 10 swarm intelligence algorithms.**

***Keywords*—Gray wolf optimization algorithm; inertia weight; adaptive; Gompertz function; swarm intelligence algorithm**

I. INTRODUCTION

The world is full of unknowns, and there are many uncertain problems in the unknown world. There is a lot of uncertain information that needs to be processed. To represent and process uncertain information, many optimization problems arise. Many optimization algorithms have emerged as the times require. In the fields of applied mathematics and engineering, there are a large number of optimization problems, and the computational solutions of these problems are located in a complex solution space. Therefore, finding new optimization methods with fast computation speed and strong convergence ability is of great practical significance. With the development of digitization and informatization, more and more meta-heuristic algorithms are being applied in the fields of science and engineering. Metaheuristic algorithms have the characteristics of self-organization, compatibility, parallelism, holism, and coordination. Their working principle is to initialize a set of random solutions and then perform repeated feedback iterations to approach the expected goal. In this search mechanism, the algorithm only needs to know the objective function and search range and can obtain the target solution regardless of whether the search range is continuous and differentiable. Metaheuristic algorithms are mainly divided into three types: biological evolution, natural phenomena, and species' living habits. The swarm intelligence optimization algorithm can specifically represent unknown information. The swarm intelligence algorithm is a model with optimization as its task. The swarm intelligence algorithm is generally based on imitating the living methods and behavioral habits of natural organisms and can optimize problems and methods with special methods. The iterative process of swarm intelligence algorithms is generally based on feedback from individuals within the population, such as particle swarm optimization, where all individuals in the population refer to the position information of the globally optimal individual to move. The cuckoo algorithm simulates the behavioral habits of cuckoo chicks and updates position information based on Levy flight. The grey wolf optimization algorithm is based on the hunting behavior habits of the grey wolf population, and the position updates of all individuals in the population are based on the globally optimal positions of three grey wolves. Researchers have proposed many solutions and drawn many important conclusions and achievements in exploring the performance improvement of swarm intelligence algorithms.

In the expansive domain of optimization algorithms, methodologies that diligently seek optimal or near-optimal solutions have demonstrated their indispensable value across a spectrum of disciplines and fields. These algorithms are not merely instrumental in enhancing the efficiency and performance of systems, designs, or models but also play a pivotal role in decision-making, resource allocation, and quality improvement. In various domains, such as machine learning, engineering, economics, logistics, biology, and computer science, optimization algorithms facilitate optimal parameter adjustments, design, resource distribution, and experimental design, thereby crafting a robust platform for interdisciplinary communication and collaboration. While these instances merely skim the surface regarding the application of optimization algorithms across disciplines, their profound impact and extensive connections undeniably propel the continuous progression and development of scientific technology. Transitioning from the general landscape of optimization algorithms, the Gray Wolf Optimizer (GWO) warrants specific attention, presenting its unique methodology in the rich field of optimization.

Researchers have been inspired by long-term observations of the social interactions, lifestyles, and biological behaviors of organisms such as fish, ants, elephants, wolves, and bees in nature, and have developed a series of related optimization algorithms to solve many practical problems such as engineering optimization and power dispatch. The GWO algorithm is implemented by utilizing the intelligence of gray wolves and their collective hunting characteristics. The GWO algorithm simulates a group of wolves following a specific hierarchical pattern, with different categories of wolves (named alpha, beta, delta, and omega) playing different roles

in the hunting mechanism to achieve search and hunting purposes.

II. Literature Review

GWO is a swarm intelligence optimization algorithm proposed by Mirjalili. This algorithm is derived from the hunting mechanics and leadership levels of gray wolves in the natural world. Mirjalili [1] used GWO to train the multi-layer perceptron (MLP). It's found that GWO has higher accuracy and is competitive in avoiding local optimality. Mohanty [2] completed the maximum power point tracking design of photovoltaic PV systems by GWO. Mirjalili [3] proposed the multi-objective gray wolf optimizer (MOGWO) in 2016, which is very effective in solving multi-objective optimization problems. Emary [4] proposed a binary version called bGWO for better feature selection. Heidari [5] integrated Levy flight and greedy algorithms to obtain a new LGWO algorithm to improve optimization performance. Faris [6] also discussed different versions of GWO optimization in detail, divided them into modified versions, hybrid versions, and parallel versions, and analyzed their role in the main application fields. Kohli [7] introduced chaos theory into GWO, which greatly improved the global convergence speed. Gupta [8] proposed an improved algorithm called RWGWO rooted in random walks to solve optimization problems in life. Nadimi [9] proposed an efficient gray wolf optimizer (I-GWO) based on the dimensional learning hunting (DLH) search strategy. Al [10] combined GWO with PSO to implement a new BPSOGWO binary algorithm to find the best feature subset.

GWO transcends its foundational application in mathematical function optimization, demonstrating utility across diverse domains. Altan [11] used GWO to optimize the intrinsic model function output and efficiently utilize wind energy. Zhao [12] used GWO to search feature sets to improve the diagnostic accuracy of patients with paraquat poisoning. Jayabarathi [13] also used GWO to solve economic dispatch problems. Sulaiman [14] applied GWO to solve the power deployment problem (ORPD) in the power system. Shariati [15] created a model combining a hybrid extreme learning machine with GWO to forecast the strength of concrete if replaced partially by cement. Jino [16] applied optimization algorithms like GWO in advanced image processing fields. Ramakrishnan [17] uses MRG-GWO for segmentation in an estimate of the CT brain tumor images. The GWO algorithm can help find optimal or suboptimal scheduling solutions, Jiang [18] used GWO to solve cases of scheduling Job Shop and Flexible Job Shop. Wei [19] improved SVM by using GWO and applied it in predicting the second major. Yang [20] used grouped GWO to optimize the parameters of wind turbines and improve the maximum power and obstacle-breaking capability.

GWO has more application scenarios including machine learning, image processing, and engineering design.

Within the machine learning sphere, SVM's (Support Vector Machines) potential is often bound by the intricacies of parameter optimization. Zhou's research [21] in 2021 elucidated this challenge by introducing two models. While both models aimed at earthquake forecasting, the latter, harnessing the capabilities of GWO, showcased commendable performance. This reinforces GWO's capability for effective exploration and exploitation in parameter optimization.

In the domain of image processing, particularly image segmentation, achieving a synergy of quality and efficiency remains pivotal. Khairuzzaman's work [22] in 2017 offers a paradigm in this regard. By leveraging GWO for multilevel thresholding, the research underscored GWO's adaptability in delivering quality segmentation with computational expediency.

Transitioning to engineering design, particularly in wind energy optimization, the nuances of turbine efficiency stand paramount. Yang's 2017 study [20] on the optimization of Maximum power points for wind turbines, specifically those operating on doubly-fed induction generators, provides a testament to GWO's efficacy. The introduction of Grouped GWO in this context exemplifies the optimizer's finesse in adaptive parameter adjustments for enhanced energy outputs.

Beyond its established domains, GWO has shown remarkable adaptability in addressing various complex optimization problems, encompassing both classical combinatorial issues and cutting-edge applications.

A quintessential example is the Traveling Salesman Problem (TSP). Given the complexity of determining the shortest possible route that visits each city exactly once and returns to the origin, Panwar's contribution is notable. In 2021, Panwar [23] employed a discrete GWO approach, paving the way for efficient solutions to symmetric TSP instances. This application not only accentuates GWO's versatility but also underscores its potential in combinatorial optimization.

Furthermore, in the domain of software engineering, predicting software defects based on metrics data is a crucial task. GWO's application in Software Defect Prediction (SDP) focuses on optimizing both feature selection and classifier parameters. One intriguing application by Kermadi [24] highlighted GWO's efficacy in designing an efficient photovoltaic array hybrid maximum power point tracker, specifically tailored for intricate local shading conditions in SDP.

Multi-objective optimization problems (MOP) present another challenging arena, where the goal is to find non-inferior solutions across multiple objectives. Wu's research in 2020 [25] exemplifies this by integrating GWO with other objective optimizations for wind speed forecasting. This innovative approach, leveraging GWO's capabilities, orchestrates harmony among various objectives, generating superior solutions.

It further reveals novel applications of GWO in electrical and control systems. Lakum [26] employed GWO for optimally placing and sizing active power filters in radial systems, especially amidst nonlinear-distributed generation. Arora's work [27] ventured into algorithmic hybridization, combining GWO with the Crow Search Algorithm (CSA) for enhanced function optimization and feature selection. Sun's research [28], on the other hand, utilized GWO for the intricate task of feedback control optimization in PM hub motors. Moreover, Eltamaly's study [29] stands out in

harnessing GWO-FLC to track dynamic maximum power points (MPP) for PV systems under variable shading conditions. Jamal employs the Improved Grey Wolf Optimization (IGWO) algorithm to optimize overcurrent relay coordination, demonstrating its enhanced efficiency and reliability over conventional methods [30]. Telugu utilizes the Chaos-enhanced Grey Wolf Optimization Algorithm (CGWO) for designing a two-stage CMOS Differential Amplifier, achieving significant improvements in reducing circuit size and power dissipation compared to traditional optimization methods [31].

The diverse applications of GWO demonstrate its adaptability and robustness across various research arenas, addressing distinct challenges and expanding its applicability in optimization landscapes. Its multifaceted use across sectors showcases its versatility in delivering optimal solutions. However, current improvement ideas often blend multiple algorithms, facing issues like low accuracy, slow convergence speed, and poor stability. This article suggests a streamlined and efficient enhancement plan solely based on the GWO algorithm, aiming to tackle these issues.

Section III introduces the background and basic principles of the GWO algorithm; Section IV introduces the two strategies of inertia weight and adaptive weight respectively, and uses them for GWO position update, thereby obtaining an adaptive algorithm based on Gompertz inertia weight(GGWO); Section V uses simulation experiments to compare and analyze the convergence performance, convergence speed, stability and time complexity of 11 algorithms including GWO and GGWO from three dimensions and six test functions; finally Section VI summarizes the full text and looks forward to the diverse application scenarios of GGWO in the future.

III. Gray Wolf Optimization Algorithm

GWO simulates the hunting mechanism and leadership levels of gray wolves by dividing these wolves into four layers according to the characteristics of gray wolves. The first layer is the α layer, the leaders of the population, in charge of leading the rest wolves to hunt prey, which is interpreted as the optimal solution in the algorithm. The second layer is the β layer, responsible for assisting the α layer wolf pack, and this layer interprets the sub-optimal solution in GWO. The third layer is the δ layer, which should obey any orders or decisions made by the previous two α and β and they have to investigate more etc. The grading mechanism of gray wolf packs is not static. Some of the α and β with poor fitness will degenerate to δ. The fourth layer is called the ω layer, which updates its position according to the previous three α, β, or δ. Furthermore, these four layers of wolves α, β, δ, and ω cooperate in the hunting. There are in total three main stages of gray wolf hunting, and they are simulated specifically as surrounding prey, chasing prey, and attacking prey.

In the first stage of surrounding the prey, the gray wolf will update its position based on the position information of the prey and gradually surround the prey, as shown in Eq. (1):

$$\overrightarrow{X}(t+1) = \overrightarrow{X_p}(t) - \vec{A} \cdot \vec{D} \quad (1)$$

where X stands for the position vector of the gray wolf, t represents the number of iterations, X_p

is the position vector of the prey, and the parameters A and D as shown in Eq. (2), (3):

$$\vec{A} = 2\vec{a}\vec{r_1} - \vec{a} \quad (2)$$

$$\vec{D} = |\vec{C} \cdot \overrightarrow{X_p}(t) - \vec{X}(t)| \quad (3)$$

Among them, $\vec{a}$ linearly decreases from 2 to 0 during iteration, $\vec{r_1}$ is a random vector on $[0,1]$, and

the parameter C is as follows in Eq. (4):

$$\vec{C} = 2\vec{r_2} \quad (4)$$

where is a random vector.

While hunting prey, the behaviors of all the gray wolves are guided by α, β and δ gray wolves and these three layers of wolves may also cooperate with ω gray wolves in hunting. To better simulate and reproduce the hunting strategy of gray wolves, we suppose that α, β, and δ have a better idea of the potential position of prey. Via sorting the fitness values of all the wolves, the best three layers of wolves are chosen as α, β, and δ respectively. The specific gray wolf location update steps are as follows:

First, the corresponding $\vec{D}_\beta$ and $\vec{D}_\delta$ are as shown in Eq. (5):

$$\vec{D}_\alpha = |\overrightarrow{C_1} \cdot \overrightarrow{X_\alpha} - \vec{X}| \ \vec{D}_\beta = |\overrightarrow{C_2} \cdot \overrightarrow{X_\beta} - \vec{X}| \ \vec{D}_\delta = |\overrightarrow{C_3} \cdot \overrightarrow{X_\delta} - \vec{X}| \quad (5)$$

Then, solve the position vector $\vec{X}(t+1)$ of the current gray wolf in the next iteration, as shown in Eq. (6) and (7):

$$\overrightarrow{X_1} = \overrightarrow{X_\alpha} - \overrightarrow{A_1} \cdot \overrightarrow{D_\alpha} \quad \overrightarrow{X_2} = \overrightarrow{X_\beta} - \overrightarrow{A_2} \cdot \overrightarrow{D_\beta} \quad \overrightarrow{X_3} = \overrightarrow{X_\delta} - \overrightarrow{A_3} \cdot \overrightarrow{D_\delta} \quad (6)$$

$$\vec{X}(t+1) = \frac{\overrightarrow{X_1} + \overrightarrow{X_2} + \overrightarrow{X_3}}{3} \quad (7)$$

In summary, this algorithm outlines a structured process for GWO, iterating through defined steps to ascertain optimal solutions by emulating the hierarchical and behavioral dynamics of gray wolves.

IV. Adaptive Gray Wolf Optimization Algorithm based on Gompertz Inertia Weight Strategy

A. Gompertz Inertia Weight Strategy

The Gompertz function [32] is monotonic, and its function expression is as shown in Eq. (8):

$$y = -e^{-e^{-x}} + 1 \quad (8)$$

Draw the graph of the Gompertz function as shown in Fig. 1:

As shown in Fig. 1, the Gompertz function curve is characterized by slow growth in the initial and final stages and rapid growth in the middle section. The image of the

Gompertz function tends to decrease as the abscissa increases, which is related to the iterative process of swarm intelligence algorithms. In the early stages of swarm intelligence algorithm iteration, the population is prone to falling into local optima, so it is necessary to give a larger step size initially. Giving a larger inertia weight can increase the step size of individual gray wolf movements, thereby helping the population to better conduct global search. As the algorithm iterates, the needs of individual populations gradually shift from global optimization to local optimization. In the later stages of the algorithm, some individuals need to strive to explore global optima within a small range, so giving a smaller step size in the later stages of the algorithm can help the algorithm perform local optimization. The Gompertz function has this feature, as its value decreases as the abscissa increases, which helps our algorithm balance global optimization and local optimization. This article uses it to improve the inertia weight of the GWO. The Gompertz inertia weight ω used in this article is as follows in Eq. (9):

$$w = -e^{-e^{\frac{t}{M}}} + 1 \tag{9}$$

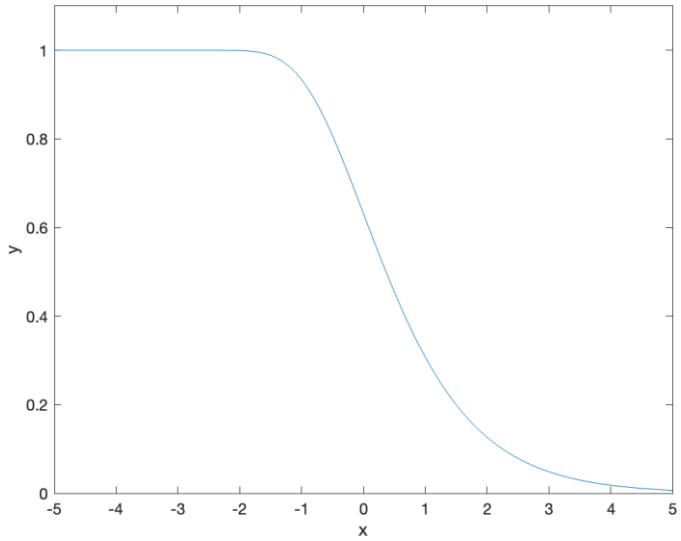


Fig. 1. Gompertz function.

The graph of selecting the right side of the y-axis as the inertia weight ω of the GWO algorithm is shown in Fig. 2:

As shown in Fig. 2, the Gompertz inertia weight remains large in the early stages of iteration, which is conducive to maintaining a large search range of the algorithm, making the algorithm less likely to fall into the local optimal solution; as the number of iterations of the algorithm increases, the curve in the middle section will decrease rapidly and eventually stabilize at a smaller value, which will help the algorithm have a smaller inertia weight in the later stages of the iteration, which will help find the optimal solution more thoroughly in local area and during long periods.

B. Gompertz Adaptive Position Update Strategy

Gompertz function is used to construct the adaptive weight Φ_i strategy as shown in Eq. (10):

$$\Phi_i = -e^{\frac{f_i}{f_{avg}}+1} - e^{\frac{f_i}{f_{avg}}} + 1, (i = 1, 2, 3) \tag{10}$$

Among them, f_1, f_2, and f_3 represent the fitness of the α, β and δ layer wolves respectively. f_{avg} is as shown in Eq. (11):

$$f_{avg} = \frac{f_1 + f_2 + f_3}{3} \tag{11}$$

Gompertz adaptive weight Φ_i is shown in Fig. 3:

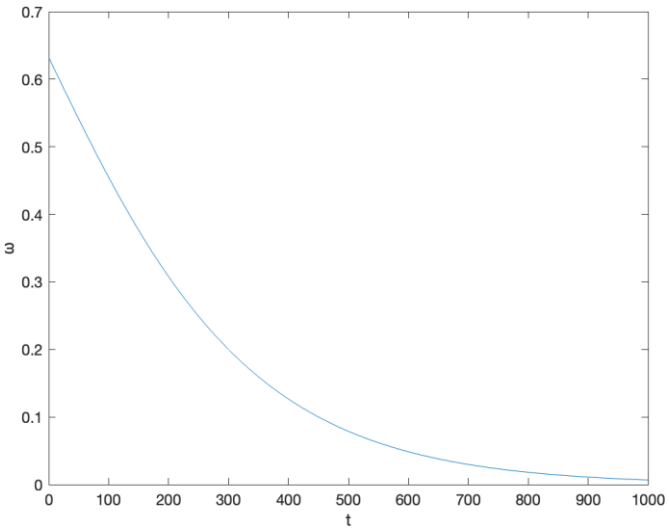


Fig. 2. Gompertz inertia weight.

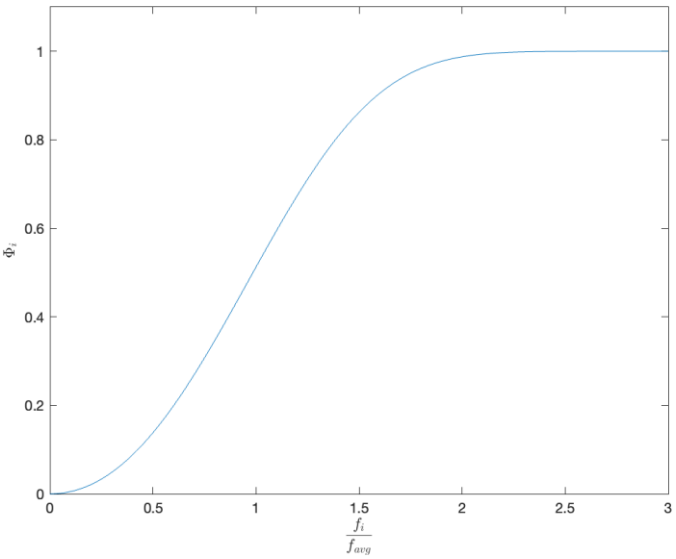


Fig. 3. Gompertz adaptive weight strategy.

As shown in Fig. 3, the Gompertz adaptive weight is close to 0 when the corresponding wolf fitness value is relatively small, indicating that the wolf is close to the prey. This step control is extremely small, which is conducive to more thoroughly finding the optimal value locally; As the corresponding wolf fitness value ratio increases, it indicates that the wolf is far away from the prey, so the Gompertz adaptive weight increases rapidly to prevent falling into the

local optimum, and the step size increases, which is conducive to searching for the optimum in the global scope.

C. Adaptive Gray Wolf Optimization Algorithm based on Gompertz Inertia Weight Strategy

Based on the Gompertz inertia weight strategy, the position update formula is modified as Eq. (12):

$$\overrightarrow{X_1} = w \cdot \overrightarrow{X_\alpha} - \overrightarrow{A_1} \cdot \overrightarrow{D_\alpha} \quad \overrightarrow{X_2} = w \cdot \overrightarrow{X_\beta} - \overrightarrow{A_2} \cdot \overrightarrow{D_\beta}$$
$$\overrightarrow{X_3} = w \cdot \overrightarrow{X_\delta} - \overrightarrow{A_3} \cdot \overrightarrow{D_\delta} \tag{12}$$

Based on the Gompertz adaptive weight strategy, the new gray wolf position $\overrightarrow{X}$ is obtained as shown in Eq. (7), (13):

$$\overrightarrow{X_1} = \Phi_1 \cdot \overrightarrow{X_1} \quad \overrightarrow{X_2} = \Phi_2 \cdot \overrightarrow{X_2} \quad \overrightarrow{X_3} = \Phi_3 \cdot \overrightarrow{X_3} \tag{13}$$

The steps of the GGWO algorithm are as follows:

1) Set the relevant parameters $\vec{a}$, $\overrightarrow{r_1}$, $\overrightarrow{r_2}$, $\overrightarrow{A}$, $\overrightarrow{C}$ according to Eq. (2), (3) and (4);

2) Define α, β and δ wolves;

3) Initialize the position of the population;

4) Calculate the fitness value according to Eq. (5) sort the fitness value from large to small, and filter out the top three D_α, D_β and D_δ corresponding to α, β and δ wolves respectively;

5) Update the positions of α, β and δ wolves according to Eq. (12) by adding inertia weight ω;

6) Then update the positions according to Eq. (7), (13) by adding an adaptive weight Φ_i and performing boundary check;

7) If the maximum number of iterations is reached, the algorithm stops and the optimal value is output; otherwise, return to step 4.

The flow chart of GGWO is shown in Fig. 4:

As shown in Fig. 4, the basic parameters of the GGWO algorithm are set to initialize the wolves, and then start iteration. In the process of continuous iteration, the position of the wolves is updated through various strategies like inertial weight, adaptive weight, and bounds check. Finally, after reaching the maximum number of iterations, the optimal value is output. The algorithm in this article uses the Gompertz inertia weight strategy and adaptive position update strategy to balance the global search and local search of the gray wolf population, which can effectively consider all information during the iteration process. Gompertz inertia weight strategy calculates the iteration of the algorithm, giving the same inertia weight value to all gray wolf individuals. This is to balance the global and local search performance of all gray wolf individuals from a global perspective. The adaptive position update strategy allows all gray wolf individuals to give different position update schemes to different gray wolf individuals when the number of iterations is fixed. This reflects the different fitness values of different particles and should be given different inertia weights, which is very scientific and necessary. The unity of the Gompertz inertia weight strategy and adaptive position update strategy can use all particles in the population to have targeted inertia weights and position update adjustment strategies, which is a non-multiplicative and scientifically reasonable solution.

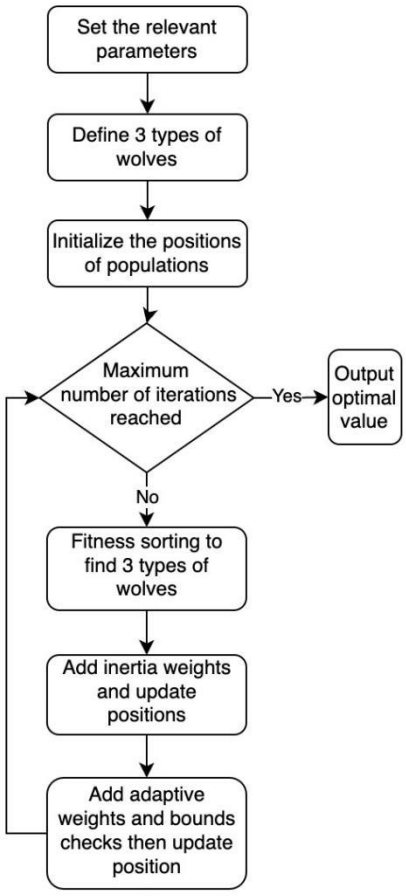


Fig. 4. The flow chart of GGWO.

V. Simulation Experiment and Result Analysis

Simulation environment: MacOS, memory: 256GB, machine frequency 3.49GHz, MATLAB R2022a.

A. Test Function and Parameter Settings

The six test functions used in the simulation experiments of this article are shown in Table I:

Table I introduces the expressions, upper and lower limits, and optimal values of the six test functions used in this simulation.

As shown in Table I, $f_1(x)$ is a simple sum of squares function, which is smooth and convex. It is typically used to assess the basic performance of an optimization algorithm. $f_2(x)$ combines a linear sum component and a multiplicative component, introducing both global structure and local minima. It tests an algorithm's ability to handle non-separable and multimodal functions. $f_3(x)$ is a nested sum of squares, adding complexity by testing the algorithm's performance on hierarchical problems where optimization at one level depends on the optimization at another. $f_4(x)$ is a maximization

function that tests the algorithm's ability to find the largest element in a vector, which can be useful for problems that require selection from a set of alternatives. $f_5(x)$ is reminiscent of the Rastrigin function, which introduces a large number of local minima, making it a challenge for algorithms to find the global minimum. $f_6(x)$ resembles a modified Schwefel function with a sinusoidal component, which is very challenging due to its complex landscape with many local optima.

TABLE I. TABLE OF TEST FUNCTIONS

Test Functions	Expressions of Functions	Domian	Optimal
F1	$f_1(x)=\sum_{i=1}^{n} x_i^2$	[-100,100]	0
F2	$f_2(x)=\sum_{i=1}^{n}\lvert x_i\rvert+\prod_{i=1}^{n}\lvert x_i\rvert$	[-10 ,10]	0
F3	$f_3(x)=\sum_{i=1}^{n}(\sum_{j=1}^{i} x_j)^2$	[-100,100]	0
F4	$f_4(x)=\max_i\{\lvert x_i\rvert, 1\le i\le n\}$	[-10,10]	0
F5	$f_5(x)=\sum_{i=1}^{n}[x_i^2-10\cos(2\pi x_i)+10]$	[-5.12,5.12]	0
F6	$f_6(x)=\frac{1}{4000}\sum_{i=1}^{n} x_i^2-\prod_{i=1}^{n}\cos(\frac{x_i}{\sqrt{i}})+1$	[-600,600]	0

B. Experimental Results Analysis

1) Comparison of average convergence curves of 11 algorithms

Select 8 classic swarm intelligence optimization algorithms [33-40], and then add the three algorithms improved in this article, which are Gray Wolf Optimization Algorithm Adding Inertia Weight (GIGWO), Gray Wolf Optimization Algorithm Adding Self-adaptive Weights (GSGWO) and Adaptive gray wolf optimization algorithm based on Gompertz inertia weight strategy (GGWO) for a total of 11 optimization methods. The average convergence curves of these 11 algorithms in three dimensions on 6 test functions are shown in Fig. 5 to 10:

In Fig. 5 to10, the abscissa reflects the number of current iterations, and the ordinate represents the logarithm of the fitness value. In low dimension (D=30) and high dimension (D=200, 300), As the iteration proceeds, the other eight algorithms except GIGWO, GSGWO and GGWO converge slowly and easily fall into local optimal, the evolution curve of the GGWO algorithm is the unique algorithm with the most obvious decline, the highest solution accuracy, and the fastest convergence speed, and will not fall into the local optimal.

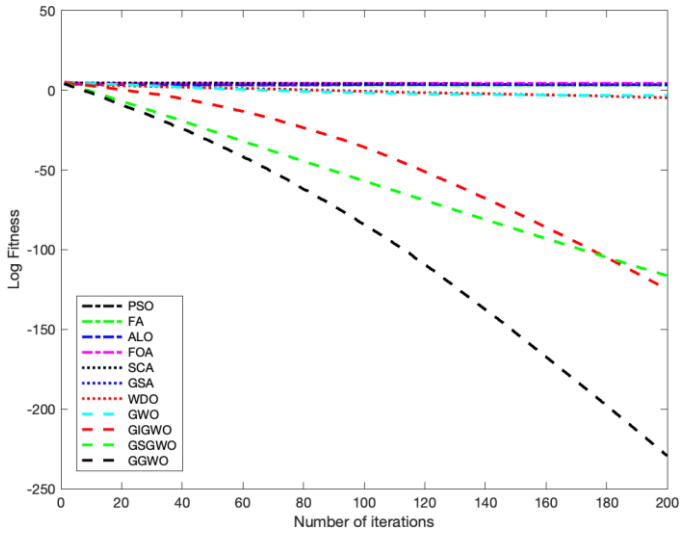
Except for the initial convergence speed of the F2 test function in high dimensions, the other eight swarm optimization algorithms are close to the optimized algorithm GIGWO, GSGWO, and GGWO. However, they will easily fall into local optimality when the number of iterations grows, and the convergence speed and accuracy are also far inferior to those of the optimized ones. In addition, the standard GWO algorithm based on GIGWO, GSGWO, and GGWO, such as the high-dimensional F6 test function, has slightly lower accuracy and convergence speed than WDO. However, after the optimization, the convergence speed of the three algorithms GIGWO, GSGWO, and GGWO are all much higher than that of WDO, which shows that the optimization strategy in this article is quite effective.

And the convergence speed and solution accuracy of GIGWO and GSGWO are far better than GWO. The convergence speed of the GGWO algorithm on F1-F4 test functions is much higher than that of the GIGWO and GSGWO algorithms, indicating that both Gompertz inertia weight and adaptive optimization strategies are effective. On F4 and F5, the convergence speed of GGWO is far better than that of GIGWO and slightly better than GSGWO. The solution accuracy of the three algorithms is close and far better than the other eight algorithms, reflecting that the superposition of the two optimization strategies is still effective because the convergence results of GGWO are better in more cases.

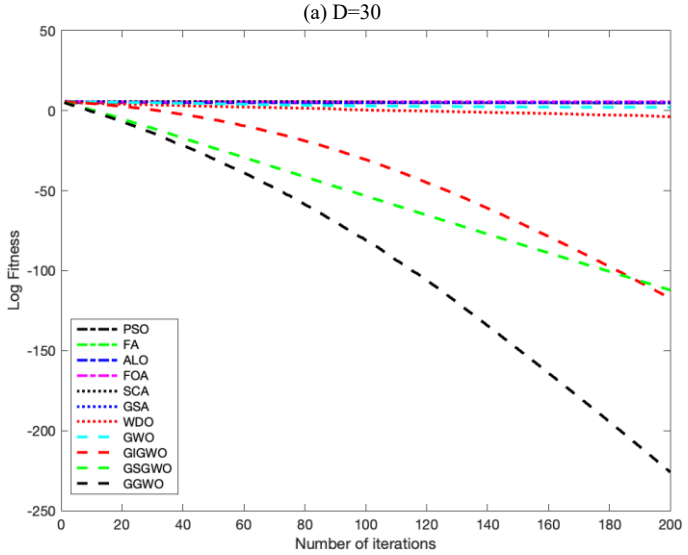
As can be seen from Fig. 5 to 10, the GGWO has the fastest decline rate and the smallest final fitness value. In the high-dimensional case of F4, although GIGWO is not completely stuck in the local optimum, the curve is stable at first, indicating that it is still stuck in the local optimum at the beginning of the iteration, which makes it impossible to search for the global optimal solution as quickly as possible. Similarly, although GSGWO has not completely stuck into the local optimal solution in the F2 high dimension, the curve gradually stabilizes as the iterations proceed, indicating that it has fallen into the local optimal, which is also not conducive to the search for the global optimal.

The optimization performance and stability of 11 algorithms in low dimension (D=30) and high dimension (D=200, 300) are shown in Tables II to IV:

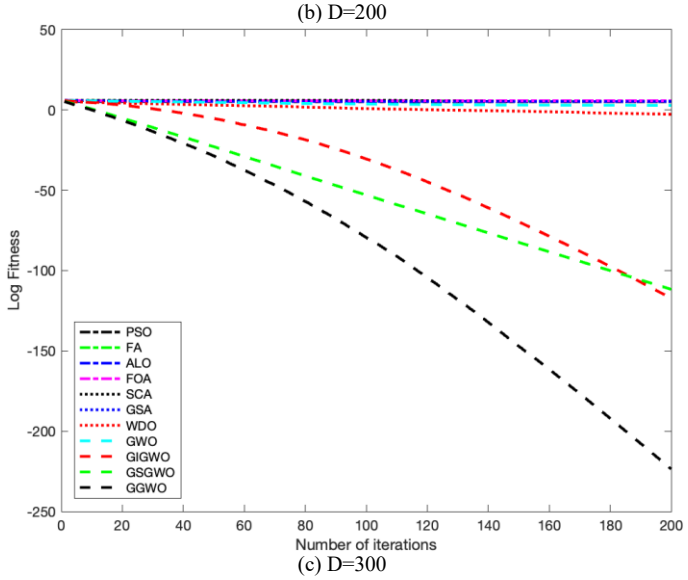
In Tables II to IV, the optimization performance and stability of 11 algorithms in low dimension (D=30) and high dimension (D=200, 300) are reflected by calculating the mean and variance. Among them, the bold data is the minimum value of the mean or standard deviation among the 11 algorithms.



(a) D=30

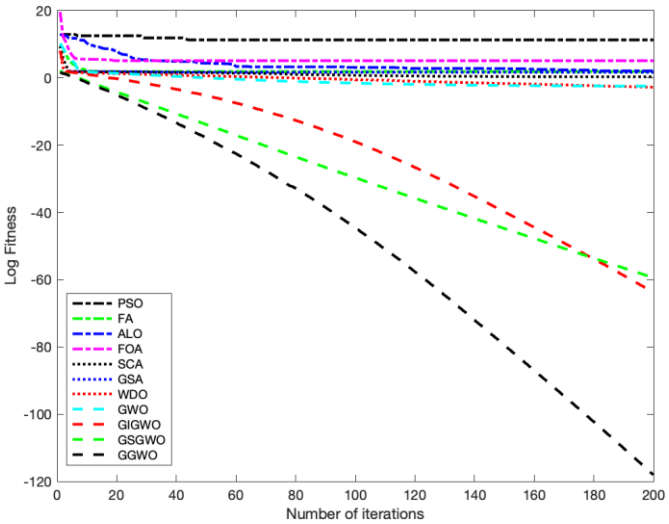


(b) D=200

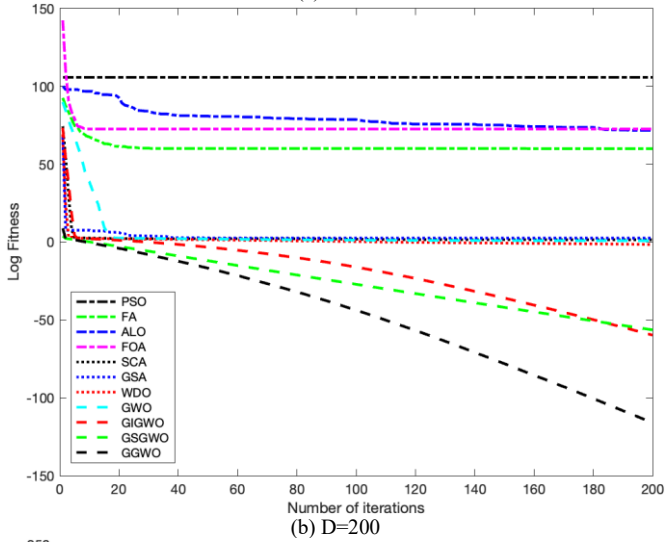


(c) D=300

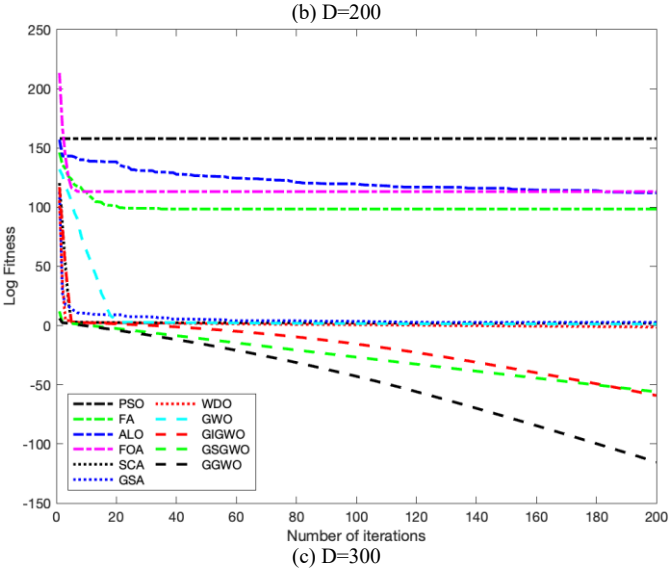
Fig. 5. Comparison chart of average convergence curve of F1.



(a) D=30

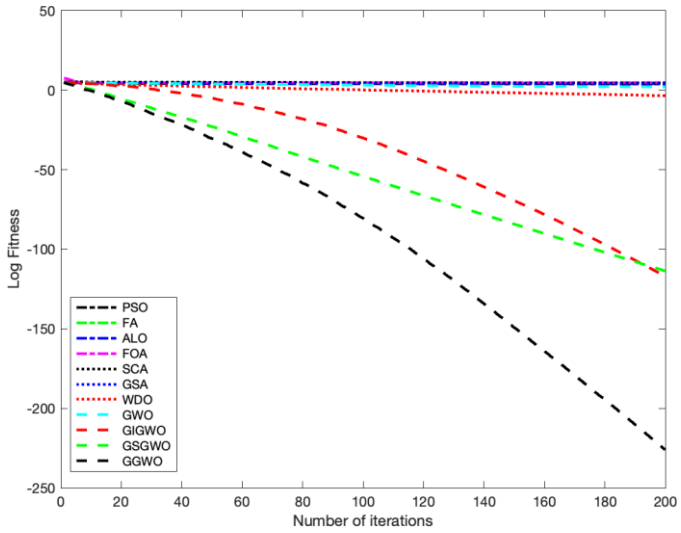


(b) D=200

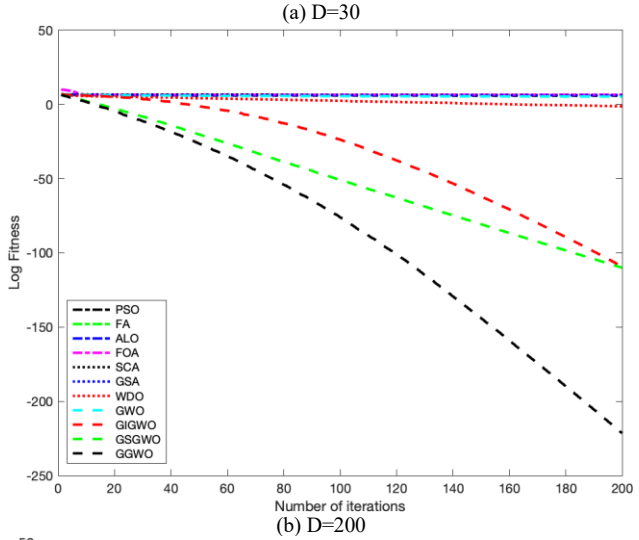


(c) D=300

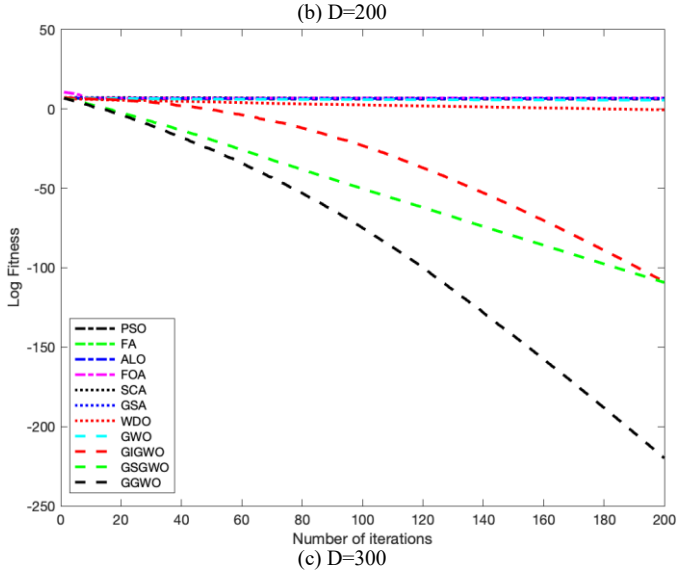
Fig. 6. Comparison chart of average convergence curve of F2.



(a) D=30

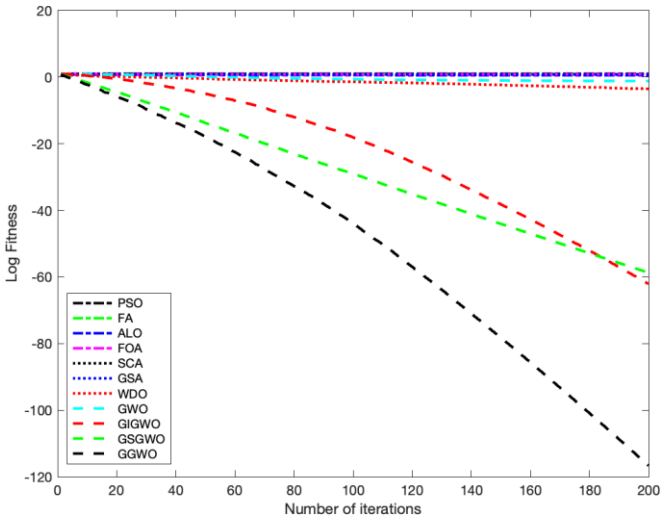


(b) D=200

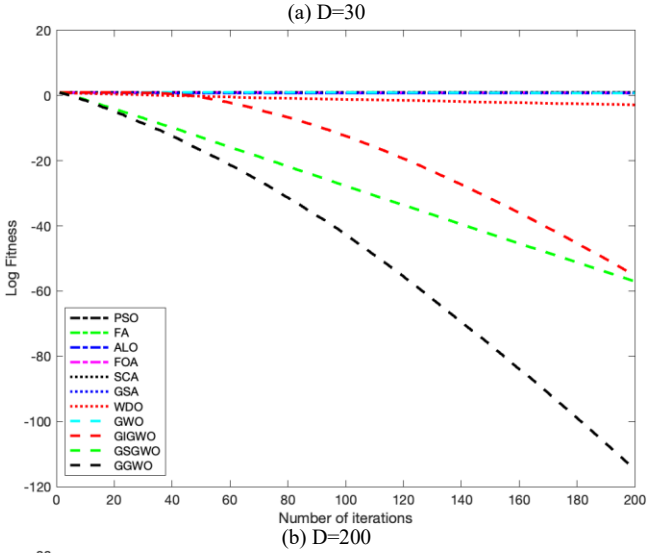


(c) D=300

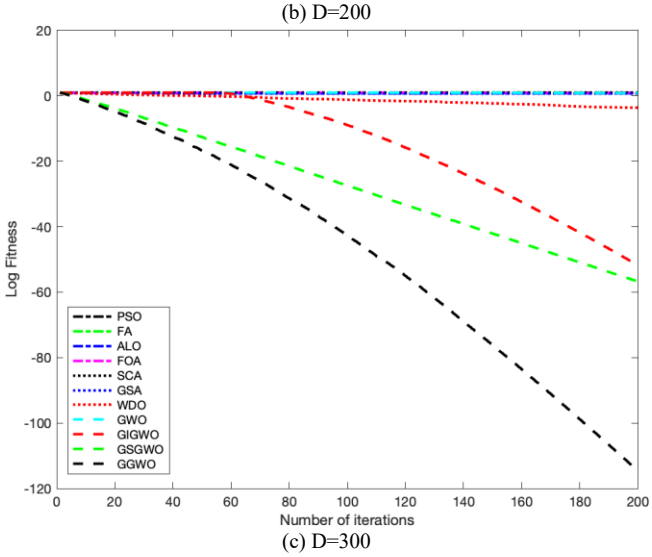
Fig. 7. Comparison chart of average convergence curve of F3.



(a) D=30

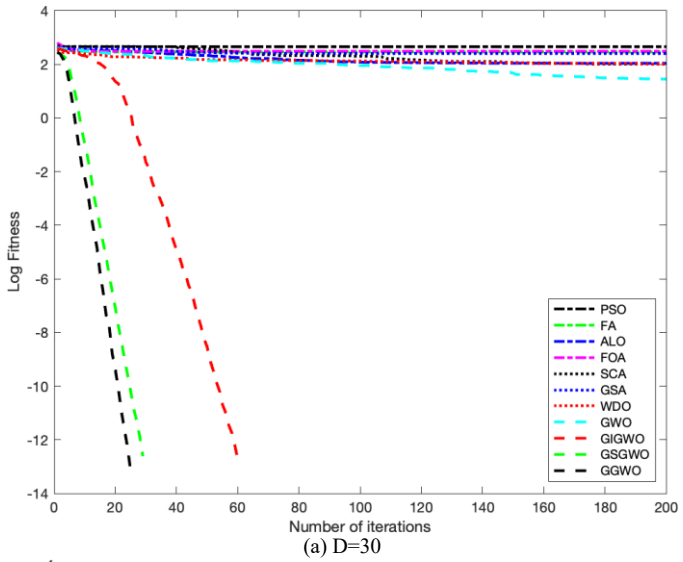


(b) D=200

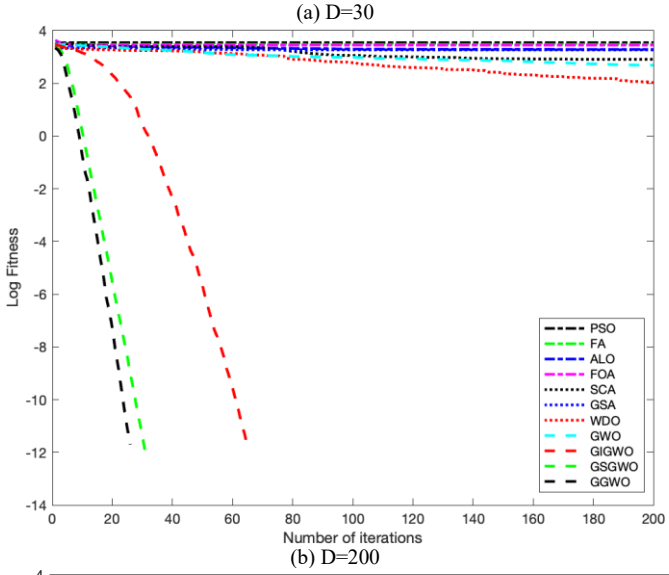


(c) D=300

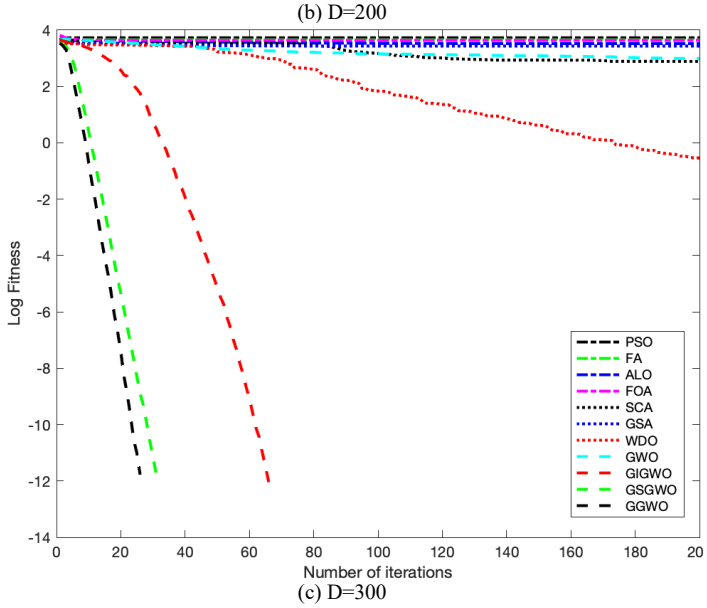
Fig. 8. Comparison chart of average convergence curve of F4.



(a) D=30

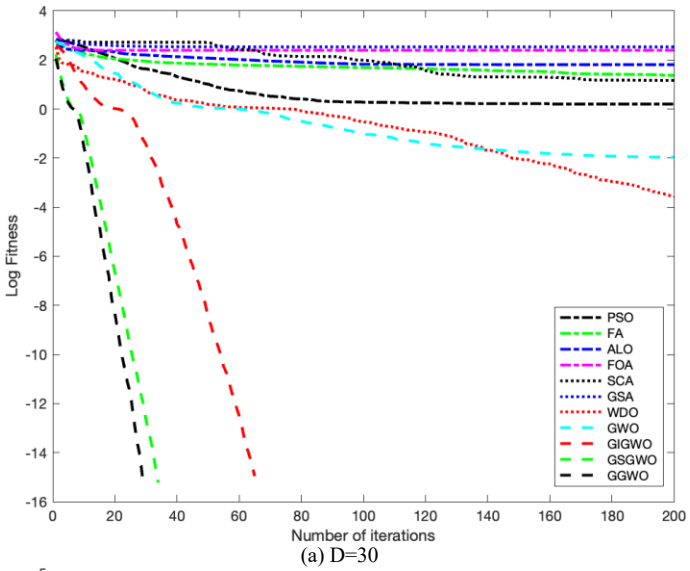


(b) D=200

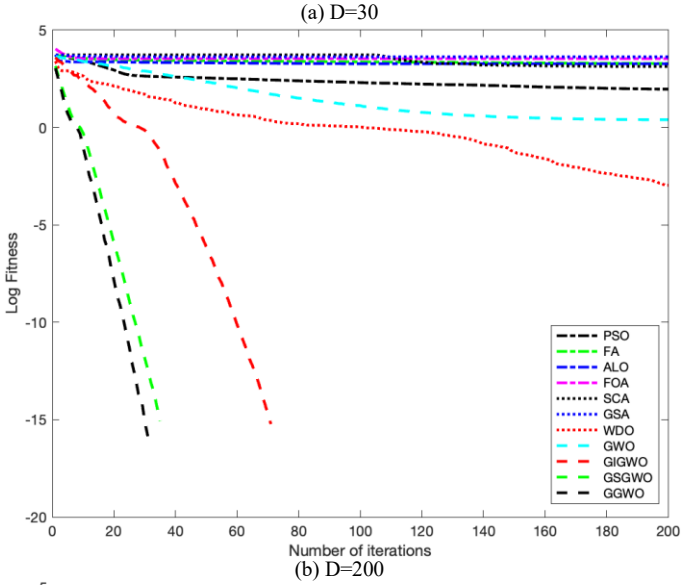


(c) D=300

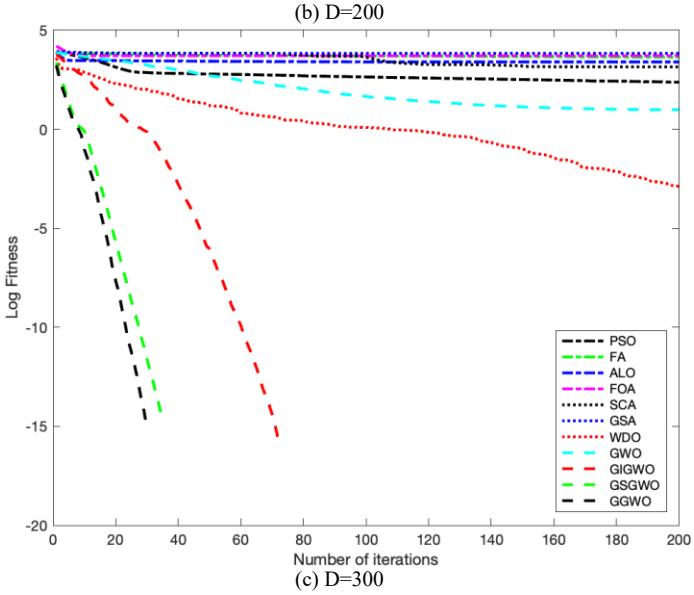
Fig. 9. Comparison chart of average convergence curve of F5.



(a) D=30



(b) D=200



(c) D=300

Fig. 10. Comparison chart of average convergence curve of F6.

2) Comparison of global optimal values of 11 algorithms

Whether it is high-dimensional or low-dimensional, GGWO has the smallest mean or standard deviation, which shows that GGWO has extremely strong optimization performance and stability. On the two test functions F5 and F6, GIGWO, GSGWO, and GGWO found the global optimal solution 0 in every experiment. The performance of GWO is lower than that of WDO, but after adding Gompertz inertia weight or adaptive weight, the performance and stability are much higher than that of WDO, which shows that the optimization strategy for GWO is quite effective.

TABLE II. TEST RESULTS OF 11 ALGORITHMS ON 6 TEST FUNCTIONS (D=30)

Functions	F1		F2		F3		F4		F5		F6	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
PSO	2.29E+03	2.79E+02	1.47E+12	2.62E+12	8.42E+03	3.20E+03	9.38E+00	2.09E-01	4.41E+02	3.31E+01	1.62E+00	1.98E-02
FA	1.30E+03	5.56E+02	5.52E+01	2.35E+01	2.96E+04	5.64E+03	7.73E+00	8.52E-01	3.11E+02	1.69E+01	1.08E+01	1.03E+01
ALO	1.03E+04	5.63E+03	1.49E+04	3.20E+04	4.49E+04	3.28E+04	4.09E+00	5.69E-01	1.49E+02	2.93E+01	7.88E+01	2.64E+01
FOA	2.87E+04	4.69E+03	1.43E+06	2.55E+06	6.11E+04	1.29E+04	6.43E+00	5.36E-01	3.39E+02	1.20E+01	2.85E+02	3.95E+01
SCA	1.35E+03	8.05E+02	3.85E+00	2.59E+00	3.63E+04	1.41E+04	6.65E+00	5.74E-01	8.93E+01	8.48E+01	6.48E+00	5.65E+00
GSA	5.14E+03	1.69E+03	5.05E+01	1.74E+01	1.19E+04	6.54E+03	5.61E+00	1.16E+00	2.32E+02	3.19E+01	3.19E+02	2.70E+01
WDO	3.90E-05	4.32E-05	4.19E-03	3.63E-03	6.23E-04	5.15E-04	4.14E-04	2.95E-04	1.46E+02	2.17E+01	5.54E-02	1.24E-01
GWO	1.49E-03	2.16E-03	2.05E-03	5.31E-04	9.82E+01	1.76E+01	9.74E-02	4.75E-02	2.54E+01	1.00E+01	5.02E-02	4.58E-02
GIGWO	1.19E-124	1.59E-124	1.40E-64	8.69E-65	3.36E-118	2.36E-118	1.35E-62	1.05E-62	0.00E+00	0.00E+00	0.00E+00	0.00E+00
GSGWO	5.73E-117	6.47E-117	7.42E-60	3.01E-60	2.80E-114	4.14E-114	3.70E-59	2.96E-59	0.00E+00	0.00E+00	0.00E+00	0.00E+00
GGWO	**8.33E-227**	**0.00E+00**	**1.20E-118**	**2.35E-118**	**3.45E-225**	**0.00E+00**	**6.19E-117**	**1.12E-116**	**0.00E+00**	**0.00E+00**	**0.00E+00**	**0.00E+00**

TABLE III. TEST RESULTS OF 11 ALGORITHMS ON 6 TEST FUNCTIONS (D=200)

Functions	F1		F2		F3		F4		F5		F6	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
PSO	1.03E+05	4.46E+03	2.10E+103	2.46E+103	6.95E+05	1.73E+05	9.87E+00	6.04E-02	3.46E+03	1.11E+02	7.72E+01	1.11E+01
FA	1.97E+05	2.18E+04	5.74E+72	1.28E+73	1.57E+06	2.18E+05	9.75E+00	8.21E-02	2.97E+03	1.08E+02	2.30E+03	2.60E+02
ALO	2.06E+05	5.68E+04	4.64E+86	1.04E+87	1.46E+06	7.02E+05	6.40E+00	5.04E-01	1.82E+03	1.19E+02	1.90E+03	2.36E+02
FOA	3.73E+05	7.12E+03	7.58E+77	1.69E+78	3.53E+06	1.59E+06	8.75E+00	4.70E-02	2.86E+03	5.08E+01	3.32E+03	1.24E+02
SCA	1.14E+05	1.95E+04	5.22E+01	3.03E+01	1.30E+06	4.11E+05	9.79E+00	4.73E-02	5.53E+02	2.85E+02	9.72E+02	3.95E+02
GSA	9.04E+04	6.60E+03	3.56E+02	2.20E+01	1.77E+06	8.79E+05	5.49E+00	3.72E-01	1.86E+03	7.78E+01	4.28E+03	1.13E+02
WDO	1.30E-02	2.34E-02	7.84E-02	7.50E-02	1.67E-01	2.18E-01	4.24E-04	3.49E-04	9.07E+02	8.30E+02	2.95E-03	3.05E-03
GWO	2.10E+02	2.49E+01	6.50E+00	1.03E+00	1.98E+05	4.72E+04	5.79E+00	4.61E-01	5.25E+02	7.61E+01	2.75E+00	6.56E-01
GIGWO	5.69E-118	9.94E-118	2.08E-60	1.43E-60	2.17E-110	1.26E-110	3.19E-54	3.89E-54	0.00E+00	0.00E+00	0.00E+00	0.00E+00
GSGWO	1.23E-112	1.11E-112	4.16E-57	1.44E-57	3.15E-110	3.71E-110	8.39E-58	2.33E-58	0.00E+00	0.00E+00	0.00E+00	0.00E+00
GGWO	**1.13E-225**	**0.00E+00**	**1.50E-116**	**1.04E-116**	**1.58E-221**	**0.00E+00**	**2.75E-115**	**1.56E-115**	**0.00E+00**	**0.00E+00**	**0.00E+00**	**0.00E+00**

TABLE V. TEST RESULTS OF 11 ALGORITHMS ON 6 TEST FUNCTIONS (D=300)

Functions	F1		F2		F3		F4		F5		F6	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
PSO	1.93E+05	2.47E+04	4.66E+159	Inf	1.44E+06	3.70E+05	9.86E+00	6.02E-02	5.33E+03	1.14E+02	2.55E+02	1.59E+01
FA	4.60E+05	5.14E+04	1.50E+118	3.15E+118	3.15E+06	8.46E+05	9.84E+00	2.44E-02	4.66E+03	1.17E+02	4.18E+03	4.56E+02
ALO	3.26E+05	4.78E+04	2.46E+147	5.51E+147	2.74E+06	1.14E+06	6.57E+00	6.14E-01	3.15E+03	1.69E+02	2.55E+03	3.25E+02
FOA	6.14E+05	2.51E+04	1.32E+120	2.95E+120	2.41E+07	1.02E+07	8.92E+00	9.37E-02	4.38E+03	8.33E+01	5.39E+03	3.12E+02
SCA	1.84E+05	2.49E+04	1.23E+02	4.22E+01	3.85E+06	5.83E+05	9.90E+00	4.23E-02	1.33E+03	6.70E+02	1.70E+03	3.63E+02
GSA	2.10E+05	1.08E+04	5.23E+02	3.14E+01	4.49E+06	1.88E+06	5.29E+00	3.65E-01	2.74E+03	6.60E+01	6.95E+03	2.51E+02
WDO	1.07E-03	1.02E-03	9.01E-02	5.76E-02	9.67E-02	5.88E-02	9.56E-04	4.91E-04	9.90E+02	1.36E+03	4.88E-04	4.13E-04
GWO	9.48E+02	2.23E+02	2.17E+01	3.17E+00	4.72E+05	8.27E+04	7.04E+00	5.97E-01	8.33E+02	2.47E+01	9.61E+00	7.45E-01
GIGWO	9.35E-118	5.23E-118	1.05E-59	6.71E-60	6.09E-109	7.32E-109	1.81E-52	2.53E-52	0.00E+00	0.00E+00	0.00E+00	0.00E+00
GSGWO	3.53E-112	3.51E-112	1.05E-56	3.26E-57	4.20E-109	9.09E-109	2.62E-57	1.30E-57	0.00E+00	0.00E+00	0.00E+00	0.00E+00
GGWO	**3.53E-224**	**0.00E+00**	**7.52E-115**	**1.02E-114**	**1.01E-219**	**0.00E+00**	**4.41E-115**	**4.04E-115**	**0.00E+00**	**0.00E+00**	**0.00E+00**	**0.00E+00**

In conclusion, the stability and convergence performance of GGWO is also the best among the 11 algorithms.

C. GGWO Time Complexity Analysis

The time complexity of GWO is $O(nmD)$, where n is the gray wolves' total number of in populations, m is the maximum number of iterations, and D is the dimension of the corresponding optimization problem. Moreover, GWO has one of the smallest time complexity among the eight algorithms in this article because the time complexity of other algorithms such as FA and GSA is as high as $O(n^2mD)$. GGWO uses Gompertz inertia weights and adaptive weights to update the position in the algorithm, which is essentially equivalent to linearly multiplying a constant in the formula during each iteration of the standard gray wolf optimization algorithm. Therefore, GGWO does not increase the time complexity of the original algorithm GWO, which means the time complexity of the improved algorithm GGWO in this article is also the smallest, $O(nmD)$.

GGWO greatly improves the algorithm's convergence speed, ability to jump out of local optima, and stability without additional increase in time complexity. In comparison with the other 10 population intelligent optimization algorithms, it clearly shows that the optimization performance far exceeds that of other algorithms.

VI. CONCLUSION

An adaptive gray wolf optimization algorithm based on the Gompertz inertia weight strategy is proposed, which uses the Gompertz function to improve the inertia weight and position update formulas. By comparing the simulation experiments, 11 different swarm intelligence algorithms were used on 6 test functions to draw the average convergence curve, and the average value of 10 runs was taken as the final display result. Experimental results show that GGWO has the smallest variance in the test functions, proving that it has the best stability. In addition, the average convergence curve of GGWO decreases the fastest, indicating that it has the fastest convergence speed. Moreover, the time complexity of GGWO is $O(nmD)$, which has a certain application potential. From the comparative analysis of standard deviation, convergence curve, time complexity, and other angles, the experimental results show that the improved algorithm GGWO has the characteristics of good stability, fast convergence speed, and high solution accuracy.

Although GGWO has made certain improvements in solution accuracy, speed, and stability, there are still some areas for future improvements. GGWO's work mainly focuses on adjusting inertial weights; we can consider other position update formulas in the future. In addition, the population initialization of GGWO is too random. Due to this, Latin hypercube sampling will be considered to initialize the population operation. Besides, while GGWO improved sharply on the simple or unimodal test functions, for processing some complex test functions or data, the effect might not greatly improved. Based on various industrial applications, GGWO can be an efficient optimization tool that can be used to deal with many practical optimization problems. In the future, GGWO will be used in some practical problems, such as medical image recognition, fault detection, UAV path planning, quantum neural network optimization, and other issues.

DATA AVAILABILITY

The data supporting the findings of this study are available from the GGWO repository at this link: GGWO

CONFLICTS OF INTEREST

The author declares that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] Seyedali Mirjalili. How effective is the grey wolf optimizer in training multi-layer perceptrons. Applied Intelligence, 43:150–161, 2015.

[2] Satyajit Mohanty, Bidyadhar Subudhi, and Pravat Kumar Ray. A new mppt design using grey wolf optimization technique for photovoltaic system under partial shading conditions. IEEE Transactions on Sustainable Energy, 7(1):181–188, 2015.

[3] Seyedali Mirjalili, Shahrzad Saremi, Seyed Mohammad Mirjalili, and Leandro dos S Coelho. Multi-objective grey wolf optimizer: a novel

algorithm for multi-criterion optimization. Expert systems with applications, 47:106–119, 2016.

[4] Eid Emary, Hossam M Zawbaa, and Aboul Ella Hassanien. Binary grey wolf optimization approaches for feature selection. Neurocomputing, 172:371–381, 2016.

[5] Ali Asghar Heidari and Parham Pahlavani. An efficient modified grey wolf optimizer with levy flight for optimization tasks. Applied Soft Computing, 60:115–134, 2017.

[6] Hossam Faris, Ibrahim Aljarah, Mohammed Azmi Al-Betar, and Seyedali Mirjalili. Grey wolf optimizer: a review of recent variants and applications. Neural computing and applications, 30:413–435, 2018.

[7] Mehak Kohli and Sankalap Arora. Chaotic grey wolf optimization algorithm for constrained optimization problems. Journal of computational design and engineering, 5(4):458–472, 2018.

[8] Shubham Gupta and Kusum Deep. A novel random walk grey wolf optimizer. Swarm and evolutionary computation, 44:101–112, 2019.

[9] Mohammad H Nadimi-Shahraki, Shokooh Taghian, and Seyedali Mirjalili. An improved grey wolf optimizer for solving engineering problems. Expert Systems with Applications, 166:113917, 2021.

[10] Qasem Al-Tashi, Said Jadid Abdul Kadir, Helmi Md Rais, Seyedali Mirjalili, and Hitham Alhussian. Binary optimization using hybrid grey wolf optimization for feature selection. Ieee Access, 7:39496–39508, 2019.

[11] Ayta¸c Altan, Se¸ckin Karasu, and Enrico Zio. A new hybrid model for wind speed forecasting combining long short-term memory neural network, decomposition methods, and grey wolf optimizer. Applied Soft Computing, 100:106996, 2021.

[12] Xuehua Zhao, Xiang Zhang, Zhennao Cai, Xin Tian, Xianqin Wang, Ying Huang, Huiling Chen, and Lufeng Hu. Chaos enhanced grey wolf optimization wrapped elm for diagnosis of paraquat-poisoned patients. Computational biology and chemistry, 78:481–490, 2019.

[13] T Jayabarathi, T Raghunathan, BR Adarsh, and Ponnuthurai Nagaratnam Suganthan. Economic dispatch using hybrid grey wolf optimizer. Energy, 111:630–641, 2016.

[14] Mohd Herwan Sulaiman, Zuriani Mustaffa, Mohd Rusllim Mohamed, and Omar Aliman. Using the gray wolf optimizer for solving optimal reactive power dispatch problem. Applied Soft Computing, 32:286–292, 2015.

[15] Mahdi Shariati, Mohammad Saeed Mafipour, Behzad Ghahremani, Fazel Azarhomayun, Masoud Ahmadi, Nguyen Thoi Trung, and Ali Shariati. A novel hybrid extreme learning machine– grey wolf optimizer (elm-gwo) model to predict compressive strength of concrete with partial replacements for cement. Engineering with Computers, pages 1–23, 2022.

[16] SR Jino Ramson, K Lova Raju, S Vishnu, and Theodoros Anagnostopoulos. Nature inspired optimization techniques for image processing—a short review. Nature inspired optimization techniques for image processing applications, pages 113–145, 2019.

[17] T Ramakrishnan and B Sankaragomathi. A professional estimate on the computed tomography brain tumor images using svm-smo for classification and mrg-gwo for segmentation. Pattern Recognition Letters, 94:163–171, 2017.

[18] Tianhua Jiang and Chao Zhang. Application of grey wolf optimization for solving combinatorial problems: Job shop and flexible job shop scheduling cases. Ieee Access, 6:26231–26240, 2018.

[19] Yan Wei, Ni Ni, Dayou Liu, Huiling Chen, Mingjing Wang, Qiang Li, Xiaojun Cui, and Haipeng Ye. An improved grey wolf optimization strategy enhanced svm and its application in predicting the second major. Mathematical Problems in Engineering, 2017:1–12, 2017.

[20] Bo Yang, Xiaoshun Zhang, Tao Yu, Hongchun Shu, and Zihao Fang. Grouped grey wolf optimizer for maximum power point tracking of doubly-fed induction generator based wind turbine. Energy conversion and management, 133:427–443, 2017.

[21] Jian Zhou, Shuai Huang, Mingzheng Wang, and Yingui Qiu. Performance evaluation of hybrid ga–svm and gwo–svm models to predict earthquake-induced liquefaction potential of soil: a multi-dataset investigation. Engineering with Computers, pages 1–19, 2021.

[22] Abdul Kayom Md Khairuzzaman and Saurabh Chaudhury. Multilevel thresholding using grey wolf optimizer for image segmentation. Expert Systems with Applications, 86:64–76, 2017.

[23] Karuna Panwar and Kusum Deep. Discrete grey wolf optimizer for symmetric travelling salesman problem. Applied Soft Computing, 105:107298, 2021.

[24] Mostefa Kermadi, Zainal Salam, Jubaer Ahmed, and El Madjid Berkouk. An effective hybrid maximum power point tracker of photovoltaic arrays for complex partial shading conditions. IEEE Transactions on Industrial Electronics, 66(9):6990–7000, 2018.

[25] Chunying Wu, Jianzhou Wang, Xuejun Chen, Pei Du, and Wendong Yang. A novel hybrid system based on multi-objective optimization for wind speed forecasting. Renewable energy, 146:149–165, 2020.

[26] Ashokkumar Lakum and Vasundhara Mahajan. Optimal placement and sizing of multiple active power filters in radial distribution system using grey wolf optimizer in presence of nonlinear distributed generation. Electric Power Systems Research, 173:281–290, 2019.

[27] Sankalap Arora, Harpreet Singh, Manik Sharma, Sanjeev Sharma, and Priyanka Anand. A new hybrid algorithm based on grey wolf optimization and crow search algorithm for unconstrained function optimization and feature selection. Ieee Access, 7:26343–26361, 2019.

[28] Xiaodong Sun, Changchang Hu, Gang Lei, Youguang Guo, and Jianguo Zhu. State feedback control for a pm hub motor based on gray wolf optimization algorithm. IEEE Transactions on Power Electronics, 35(1):1136–1146, 2019.

[29] Ali M Eltamaly and Hassan MH Farh. Dynamic global maximum power point tracking of the pv systems under variant partial shading using hybrid gwo-flc. Solar Energy, 177:306–316, 2019.

[30] Noor Zaihah Jamal, Mohd Herwan Sulaiman, Omar Aliman and Zuriani Mustaffa, “Optimal Overcurrent Relays Coordination using an Improved Grey Wolf Optimizer” International Journal of Advanced Computer Science and Applications(ijacsa), 9(11), 2018.

[31] Telugu Maddileti, Govindarajulu Salendra and Chandra Mohan Reddy Sivappagari, “Design Optimization of Power and Area of Two-Stage CMOS Operational Amplifier Utilizing Chaos Grey Wolf Technique” International Journal of Advanced Computer Science and Applications(IJACSA), 11(7), 2020.

[32] Kathleen MC Tjørve and Even Tjørve. The use of gompertz models in growth analyses, and new gompertz-model approach: An addition to the unified-richards family. PloS one, 12(6):e0178691, 2017.

[33] James Kennedy and Russell Eberhart. Particle swarm optimization. In Proceedings of ICNN’95-international conference on neural networks, volume 4, pages 1942–1948. IEEE, 1995.

[34] Xin-She Yang. Firefly algorithms for multimodal optimization. In International symposium on stochastic algorithms, pages 169–178. Springer, 2009.

[35] Seyedali Mirjalili. The ant lion optimizer. Advances in engineering software, 83:80–98, 2015.

[36] Wen-Tsao Pan. A new fruit fly optimization algorithm: taking the financial distress model as an example. Knowledge-Based Systems, 26:69–74, 2012.

[37] Seyedali Mirjalili. Sca: a sine cosine algorithm for solving optimization problems. Knowledgebased systems, 96:120–133, 2016.

[38] Esmat Rashedi, Hossein Nezamabadi-Pour, and Saeid Saryazdi. Gsa: a gravitational search algorithm. Information sciences, 179(13):2232–2248, 2009.

[39] Zikri Bayraktar, Muge Komurcu, Jeremy A Bossard, and Douglas H Werner. The wind driven optimization technique and its application in electromagnetics. IEEE transactions on antennas and propagation, 61(5):2745–2757, 2013.

[40] Seyedali Mirjalili, Seyed Mohammad Mirjalili, and Andrew Lewis. Grey wolf optimizer. Advances in engineering software, 69:46–61, 2014.

Optimizing Shuttle-Bus Systems in Mega-Events using Computer Modeling: A Case Study of Pilgrims' Transportation System

Mohamed S. Yasein[1], Esam Ali Khan[2]
The Custodian of the Two Holy Mosques Institute of Hajj and Umrah Research, Umm Al-Qura University, Makkah, KSA[1, 2]
Electrical Engineering Department - Faculty of Engineering, Port-Said University, Port-Said, Egypt[1]

***Abstract*—Mega-events are held in a city, or more, during a limited time, which requires special attention to the infrastructure and the offered services. The Hajj event, hosted in Makkah - Saudi Arabia, is considered as an excellent example of religious mega-events. The field of computer modeling and simulation is one of the main technical tools that help in developing and understanding the risks of crowds and studying the safety means during organizing of many major events in the world. This paper focuses on using computer simulation to optimize the pilgrims' shuttle-bus transportation system in Holy Sites (Mashaaer), as a case study of optimizing shuttle-bus Systems in Mega-Events using computer modeling. The objective of paper is to develop a model of the shuttle-bus transport system to give insights of the advantages of the use as an alternative for transporting pilgrims as well as to provide decision makers with a tool that could be used to select the best parameters of the system for the most efficient operation. For this purpose, pilgrims' evacuation time, traffic congestion and average trip time, from Arafat to Muzdalifa, are identified as the performance measures for evaluating the proposed transport system. The conducted simulation can be used to assess the current systems, recommend changes to the systems, and offer indicators and readings to assist decision makers.**

Keywords*—Computer modeling; simulation; optimization; shuttle-bus systems; mega-events; hajj

I. INTRODUCTION

Mega-events are considered as large-scale events attended by huge numbers of visitors. In addition, they have been getting much attention as key ingredients in tourism for many destinations [1]. Organizing a mega-event has a great impact on the hosting cities, which requires special attention to the infrastructure and the offered services [2].

In order to classify an event as a mega event, different criteria can be considered, such as duration, scale, number of visitors and importance of the event. Among the famous events that can be considered as mega-events are the Olympic Games, the World Expo and the Football World Cup [3], in addition to some religious festivities.

The Islamic pilgrimage (Hajj) event, which occurs once a year and is hosted in Mecca KSA, is an excellent example of religious mega-events.

The Hajj organizers aim at allowing the best possible number of Muslims to perform Hajj to the fullest; and providing the best services before, during and after their visit to Mecca. Transportation is an important service that is offered throughout the entire year and especially during Hajj days. Improving the quality of transportation services requires the use of modern techniques to raise the efficiency of operations and improve services.

Hajj is a set of rites unchanged for 14 centuries that takes place from the 8^{th} to 12^{th} (or in some cases 13^{th}) of Dhul al-Hijjah, the 12^{th} month of the lunar year. On the 8^{th} of Dhul al-Hijjah, pilgrims confirm their intention to make the pilgrimage. After Fajr prayer on the 8^{th} of Dhul al-Hijjah, the pilgrims proceed to Mina where they spend the whole day. The next morning (The 9^{th} of Dhul al-Hijjah), they leave Mina to go to Arafat, where they supplicate and seek mercy of Allah. Pilgrims leave Arafat for Muzdalifa after sunset. After returning from Muzdalifa, the pilgrims perform symbolic stoning of the pillars (Ramy al-Jamarat). On the same or the following day, the pilgrims revisit the Haram Mosque in Makkah for Tawaf al-Ifadah. The night of the 10^{th}, 11^{th}, and 12^{th} of Dhul al-Hijjah are spent back at Mina, where the same process of stoning of the pillars takes place. Finally, before leaving Mecca, pilgrims perform a farewell tawaf called Tawaf al-Wadaa. Fig. 1 illustrates the route taken by pilgrims on their journey and the relative proximity between the Holy sites.

The constraints of time (from the 8^{th} to 12^{th}/13^{th} of Dhul al-Hijjah) and area (Haram, Mina, Muzdalifa, and Arafat) pose different challenges for Hajj authorities responsible for controlling traffic and crowd movements of pilgrims. Such challenges further increase with the steady increase in the number of pilgrims each year.

The restricted times for pilgrim movements between Holy sites suggest that carefully planned pilgrim services are required to be aligned in accordance with such movements. In particular, one of the most important pilgrims' services is transportation.

During the past Hajj seasons, many transportations operational strategies, including regular buses, shuttle-bus, train, or on foot (walking), were successfully experimented between the Holy sites in Makkah.

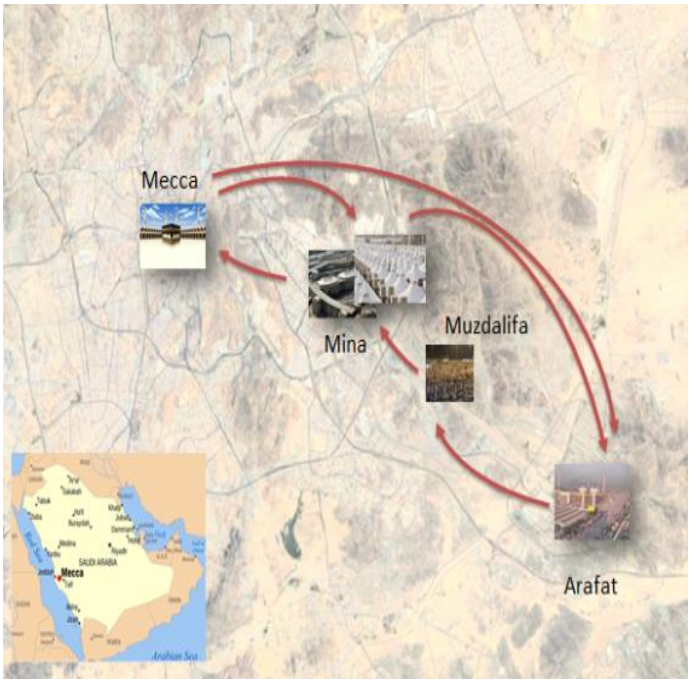


Fig. 1. The route of Hajj journey.

The Saudi Ministry of Transportation first reported on the idea of using shuttle-buses to transport pilgrims between Arafat and Muzdalifa. In the following years, the shuttle-bus project was extended to include the Muzdalifa–Mina roadway and presently continues to operate on the two segments of Arafat–Muzdalifa and Muzdalifa–Mina [4].

Due to the steady increase of number of pilgrims, which may reach five million by 2030, an optimized transportation system is of a high importance. This paper aims at utilizing computer modeling and simulation to study the effectiveness of the pilgrims' shuttle-bus transportation system in Mashaaer Holy Sites. It focuses on using computer modeling to optimize pilgrims' shuttle-bus transportation system for the path from Arafat towards Muzdalifa (Nafrah). More specifically, the paper focuses on the path of pilgrims of Turkey and Muslims of Europe, America, and Australia. This model can be used for the planning and analyzing all Hajj transportation operations. The results of the simulation may be used to assess the effectiveness of the current system and to suggest changes that might be made. Hence, the designed model can be used for optimizing the pilgrim's transportation system with respect to different parameters.

The paper is organized as follows. Section II discusses the literature review. Section III presents the proposed computer modeling of the shuttle-bus system, while results and discussions are presented in Section IV and V respectively. Conclusions are drawn in Section VI.

II. Literature Review

Several research studies investigated the research issues associated with pilgrims' transportation planning and the deployed shuttle-bus project [5], [6].

In [7], an integrated solution to the problem of pilgrimage transportation control while tracking the shuttle-bus from its starting point till its final destination was proposed. The system is designed to work in an environment where vehicle and passenger identification is required, namely at check points, in order to speed up the checking process with best quality services. In the designed system, buses are identified using RFID tags, while passengers boarding or getting off the bus are identified on the basis of RFID cards they have and finger identification.

In study [8, 9], a tool was created to assist in planning the shuttle-bus service operation configuration during Hajj using an analytical approach. With the use of this tool, the shuttle service between Arafat and Muzdalifa can be run with the fewest possible buses and cycles. It uses GPS data to gather information about the shuttle-bus service, including pick-up and drop-off timings, travel times along the path, and the number of pilgrims.

The research reported in [11] presents a simulation study of the 1422H Hajj season shuttle-bus transportation system using the Arena simulation system, with the goal of understanding of the characteristics and limitations of the system, by examining bus routes, dispatching mechanisms, and loading/unloading scenarios on system performance. During this Hajj season, approximately 160,000 pilgrims used the shuttle-buses, and 542 buses were used to transport pilgrims from Arafat to Muzdalifa. The paper suggested that in order to maintain a reasonable evacuation time, the number of buses should not be decreased below a limit of about 500 buses.

An approach in [12] is based on a stochastic simulation model to design a shuttle-bus system to transport pilgrims along the entire Arafat and Muzdalifa segment roadway, using logistics features of ProModel simulator. The model is used to estimate the number of buses needed to transport pilgrims at the shortest possible evacuation time with the least amount of congestion, as well as the number of bus stops needed to accommodate pilgrims. To evacuate Arafat in six hours, the study offered an expanded plan with 3160 buses and 740 bus stations.

A number of research papers addressed the optimization of shuttle-bus systems. The research in [14] examined the optimization of a local shuttle route. It creates a solution to the optimal routing design problem in order to minimize overall cost, which includes user and supplier costs, while taking budgetary restrictions and demand for passenger traffic into account. Both a genetic algorithm (GA) and a depth-first search algorithm (DFS) are used to accomplish this.

A hybrid genetic algorithm is used in [15] to solve a shuttle-bus route optimization model. The travel time reliability estimation approach is based on back propagation (BP) neural networks, and the optimization model is built with operational reliability maximization as its primary goal.

A structure for an optimization model that maximizes the general satisfaction of users and public transit businesses was presented in [16]. A hybrid intelligent optimization technique is utilized to solve the optimization model, which is based on

meeting the passenger journey time requirements and reducing the cost-of-service operations.

On a case study of a university shuttle-bus, the research in [17] employs the Artificial Neural Network and Support Vector Machine algorithms for the optimum journey time prediction with a lower error rate. It is also used to suggest appropriate pathways for the selected scenario.

In study [18], an optimization modeling approach based on simulation to assist airport shuttle operators in deploying electric vehicles efficiently was suggested. In order to achieve predefined objectives, the suggested approach uses an event-driven simulation model, Airport Shuttle Planning and Improved Routing Event-driven Simulation (AS-PIRES) [19], to propose an optimization model that determine the battery capacity, charging power, and number of chargers.

In study [10], a transport planning approach that is adopted for the Summer Olympic Games (SOG) and identifies lessons learned for planning of the Hajj/Umrah was explored. It described the context for each event and the nature of transport demand and supply and outlines the transport planning approaches used.

When planning a community shuttle service, the optimal stop location and route can help to reduce the walking distance of passengers and the route length. In study [13], a discrete optimization problem was proposed to make a trade-off between the walking distance of passengers and route length.

III. Modeling the Shuttle-Bus Transportation System in Arafat-Muzdalifa Area

A. Problem Formulation

The issue addressed in this study focuses on computer modeling of the shuttle-bus transportation system for pilgrims on the route from Arafat to Muzdalifa (Nafrah), in particular the route taken by Muslims from Turkey and other parts of Europe, North America, and Australia. Fig. 2 depicts the shuttle-bus route from Arafat to Muzdalifa that is taken into account in the proposed model.

The shuttle-bus transportation system in Arafat-Muzdalifa area is responsible for transporting the pilgrims from loading stations in Arafat (L_{SN} loading bus-stations) to unloading stations in Muzdalifa (U_{SN} unloading bus-stations). In Arafat, the pilgrims are moved to the bus stops in groups, each of fifty pilgrims, and loaded into the bus at that stop. When the Nafrah starts, the first group of buses departs Arafat taking pilgrims towards Muzdalifa and the following group is released to the bus stops. In Muzdalifa, buses unload the pilgrims at the pre-assigned bus stops. The shuttle-bus transportation network utilizes a dedicated two-lane bus road, where bus stops are constructed on the roads' shoulders.

The goal is to complete the evacuation of the designated number of Turkish pilgrims - roughly 240000 pilgrims - within the Nafrah's allocated time frame, maintaining the bus-stations capacity.

In order to complete evacuating the designated number of pilgrims, there is a need for a buses batch size (N), which is the overall number of buses dedicated for the transportation of pilgrims. The number of buses in a batch should be kept to a minimum both economically and to prevent excessive traffic congestion on the roads. In addition, the total time for completing the evacuation (T) should be kept within the time allotted for the Nafrah (T_A).

The different parameters that affect the shuttle-bus transportation system are as follows:

- The interarrival time for buses entering the system (I_T), which is the time interval between bus arrivals in the system. A small interarrival time may appear to speed up the evacuation operation for pilgrims because the entire bus batch size (N) begins doing their job more quickly, however, it may actually lead to traffic jams and causes overload in Arafat bus-stations, as the number of buses exceeds their capacity. Therefore, it's important to maintain a balance between the level of congestion and the rate at which buses enter the route. Using less bus interarrival time causes overload in Arafat bus-stations, as the number of buses exceeds their capacity.

Fig. 2. The path of shuttle-bus from Arafat to Muzdalifa.

- The bus trip time during Nafrah (B_T), which depends on the following parameters:
  - The pilgrims' loading time in Arafat (L_T).
  - The pilgrims' unloading time in Muzdalifa (U_T).
  - The buses' speed (B_S).
- The maximum number of cycles (loops) that each bus completes across the system before departing for the parking area (C_N). Apparently, a higher number of cycles require less bus batch size (N), but bus drivers may be more tired as they spend more time driving.

The following performance metrics can be used to assess the efficiency of shuttle-bus transportation system:

- The number of bus trips reached Muzdalifa (M_N). As mentioned earlier, for each bus trip, a group of fifty pilgrims are evacuated. Therefore, the number of bus trips is considered an indicator of the number of pilgrims that are evacuated. Moreover, another type of city-bus can be utilized, which allows evacuating a group of eighty pilgrims per bus trip.
- Total number of buses exists in the system at a certain time (N_{CB}), which is considered an indicator of the total time for completing the evacuation (T).

Moreover, traffic congestion is taken into consideration for evaluating the transportation system efficiency.

Table I summarizes the parameters used in this work to model the shuttle-bus pilgrims transportation system.

TABLE I. PARAMETERS OF THE SIMULATION MODEL

Parameter	Description
Buses batch size (N)	The overall number of buses dedicated for the transportation of pilgrims.
Nafrah time (T_A)	Time allotted for the Nafrah (the available time to finish the evacuate pilgrims from Arafat to Muzdalifah)
Total evacuation time (T)	Total time for completing the evacuation process
Interarrival time (I_T)	The time interval between bus arrivals in the system.
Bus trip time (B_T)	Total time for a bus to travel from Arafat to Muzdalifah
Loading time (L_T).	Pilgrims' loading time in Arafat
Unloading time (U_T)	Pilgrims' unloading time in Muzdalifa
Bus speed (B_S).	Average speed of a bus during the trip from Arafat to Muzdalifah
Number of cycles (C_N)	The maximum number of cycles (loops) that a bus completes across the system before departing for the parking area
Number of bus trips (M_N)	Total number of bus trips reached Muzdalifa
number of buses (N_{CB})	Total number of buses exists in the system at a certain time

B. Setting up the Model of the Shuttle-Bus System

The first step in the computer modeling operation is the data collection of system specifications, input variables, as well as performance of the existing shuttle-bus system. This is done by observing the actual transportation system performance during Hajj seasons. Based on the actual data collected from the field during Hajj seasons, the values of the following parameters were determined as:

- The interarrival time for buses entering the system (I_T) is 10 seconds. This means that six buses enter the system every minute.
- The number of loading stations in Arafat (L_{SN}) is 80 stations.
- The number of unloading stations in Muzdalifa (U_{SN}) is 32 stations.
- The pilgrims' loading time in Arafat (L_T) was taken as a triangular distribution function as triangular (6, 8, 10), with the min value of 6 min., max value of 10 min., and most likely value of 8 min.
- The pilgrims' unloading time in Muzdalifa (U_T) was taken as a triangular distribution function as triangular (4, 6, 8), with the min value of 4 min., max value of 8 min., and most likely value of 6 min.
- The buses' speed (B_S) was taken as 70 km/h, unless other values are tested.
- The maximum number of loops (C_N) was taken as 6 loops, unless other values are tested.
- The typical time of Nafrah (T_A) is about seven hours.

C. Building the Model of the Shuttle-Bus System

The next step is model construction by developing schematics and network diagrams of the shuttle-bus system, abstracting the essential features of the system, and programming the system operations that characterize the system.

This model architecture was built using Anylogic, which is a multi-method simulation tool with an intuitive GUI. It is completely written in Java and enables combining process modeling, system dynamics and agent-based modeling in one model. Moreover, it has a built-in Road Traffic Library. On one hand, Anylogic seems to be a perfect tool for our research since a microsimulation model of traffic flow can be created directly in Anylogic. On the other hand, several difficulties had to be overcome due to the fact that the tool is not primarily intended for shuttle-bus transportation systems. In order to make sure that the designed model is an accurate representation of the real system, a model validation is performed. This is done by using actual data collected from the field during Hajj seasons. The typical duration of Nafrah is about seven hours.

The model is carried out using an exclusive two-lane bus road. The processes' general structure of the model is shown in Fig. 3. In this modeled environment of roads in Mashaaer Holy Sites (especially Arafat-Muzdalifa area), concurrent groups of simulated agents interact with potential to enable direct acquisition of statistics and indications.

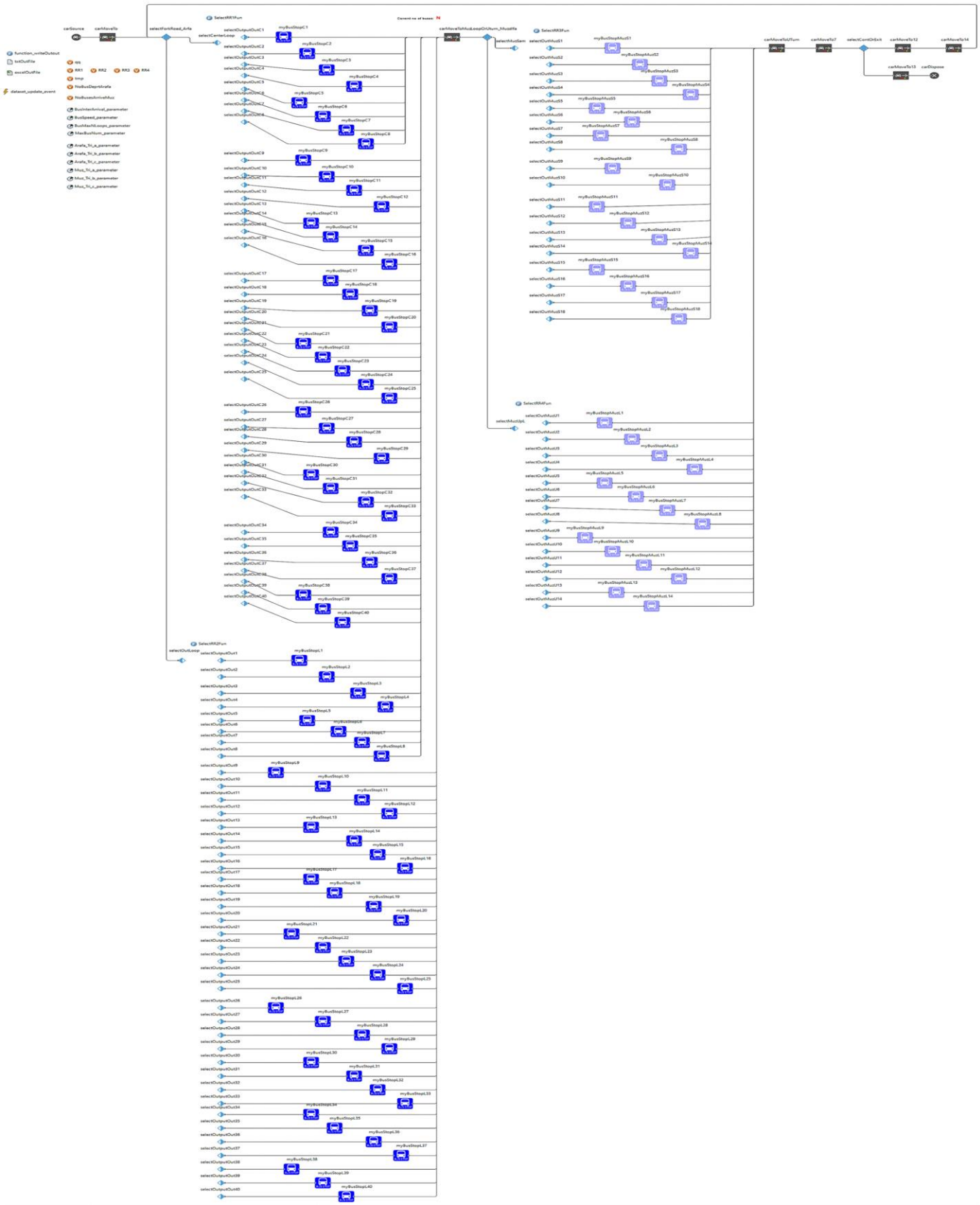


Fig. 3. The processes structure of the model.

The main processes of the model include:

- CarSource: It generates cars (buses) and tries to put them into the specified location inside a road network.
- CarMoveTo: It controls the car movement by calculating the way from its current location to the specified destination.
- SelectOutputIn and SelectOutputOut: For routing agents (buses) to different flowchart branches (bus stations).
- Delay: Delays agents (buses) for a given amount of time.
- CarDispose: Removes a car (bus) from the model.

The system parameters can be controlled through some input-fields in the starting screen of the proposed simulation system. These model parameters include:

- BusInterArrival_parameter: It controls the interarrival time for buses entering the system (in min.).
- Arafa_Tri_a_parameter, Arafa_Tri_b_parameter, and Arafa_Tri_c_parameter: It controls the loading time in Arafat.
- Muz_Tri_a_parameter, Muz_Tri_b_parameter, and Muz_Tri_c_parameter: It controls the unloading time in Muzdalifa.
- BusSpeed_parameter: It controls the speed of a bus.
- BusMaxNLoops_parameter: It controls the maximum number of cycles (loops) that a bus travels before leaving the system.

The simulation screen has some active camera viewports. Examples of the cameras that the system has include:

- Arafat area Cam: A camera that follows the loading bus stations in Arafat site.
- Muzdalifa area Cam: A camera that follows the unloading bus stations in Muzdalifa site.

A screenshot of the simulation screen (at time = 60 min.) is shown in Fig. 4.

D. Validation of the Model of the Shuttle-Bus System

Comes next an important operation, which is model verification. This is to ensure the correctness of the logical structure of the model and the correctness of the represented form in the computer.

In order to make sure that the designed model is an accurate representation of the real system, a model validation is performed. This is done by comparing the results of the simulation model with actual data collected from the field during Hajj seasons.

In the following section, an example of a simulation scenario is shown, which has been used to prove that our model gives realistic results compared to the actual data already collected.

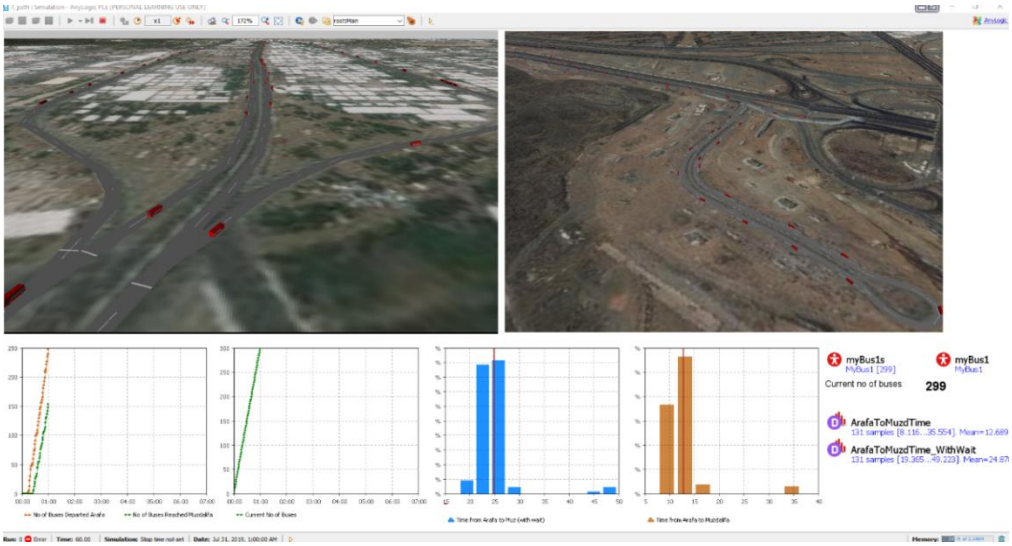


Fig. 4. A screenshot of the simulation screen.

IV. Simulation Results

Through a series of simulations, the model's behavior, and the effects of changing various parameters on the shuttle-bus transportation system are examined. An example of a simulation scenario is provided in this section.

After running the simulation experiment using the aforementioned parameters for a long duration of time, the simulation results were observed. It was found that the average trip time B_T, from Arafat to Muzdalifa, is about 21 min. The total number of buses that existed in the system and the number of trips departed Arafa and arrived Muzdalifa were observed as can be shown in Fig. 5. It can be seen in the figure that using an interarrival time of 10 seconds, the buses batch size (N) increases gradually and reach a number of 700 buses during around 114 min. (point B in the figure). In addition, using this buses batch size, and with a maximum number of loops (C_N) of 6 loops, buses start to depart to the parking area (point C in the figure) after 4 hours and 15 min., hence, the

number of buses existed in the system starts to decrease until all buses finish their job. By the time of around eight hours, the number of trips departed Arafa and arrived Muzdalifa was 4200 (point D in the figure), which means that around 210000 pilgrims were successfully evacuated to Muzdalifa, or around 336000 pilgrims were successfully evacuated, when the type of 80-passenger city-bus is used.

In Fig. 6, the results of a simulation experiment that examines the effect of tuning the maximum number of cycles (C_N) are shown. The figure shows the number of trips arriving in Muzdalifa and the total number of buses existing in the system for different number of loops, using a bus batch size N = 600. It can be depicted from the figure that:

- For C_N = 6, a total of 3600 trips (180000 pilgrims, or 288000 when the type of 80-passenger city-bus is used) were able to reach Muzdalifa within 7 hours and 20 minutes. In addition, the total duration for completing the evacuation (whole buses batch were able to leave the system to the parking area) was 7 hours and 42 minutes.
- For C_N = 7, a total of 4200 trips (210000 pilgrims, or 336000 pilgrims when the type of 80-passenger city-bus is used) were able to reach Muzdalifa within 8 hours and 26 minutes. In addition, the total duration for completing the evacuation was 8 hours and 48 minutes.
- For C_N = 8, a total of 4800 trips were able to reach Muzdalifa within 9 hours and 20 minutes. In addition, the total duration for completing the evacuation was 9 hours and 54 minutes.

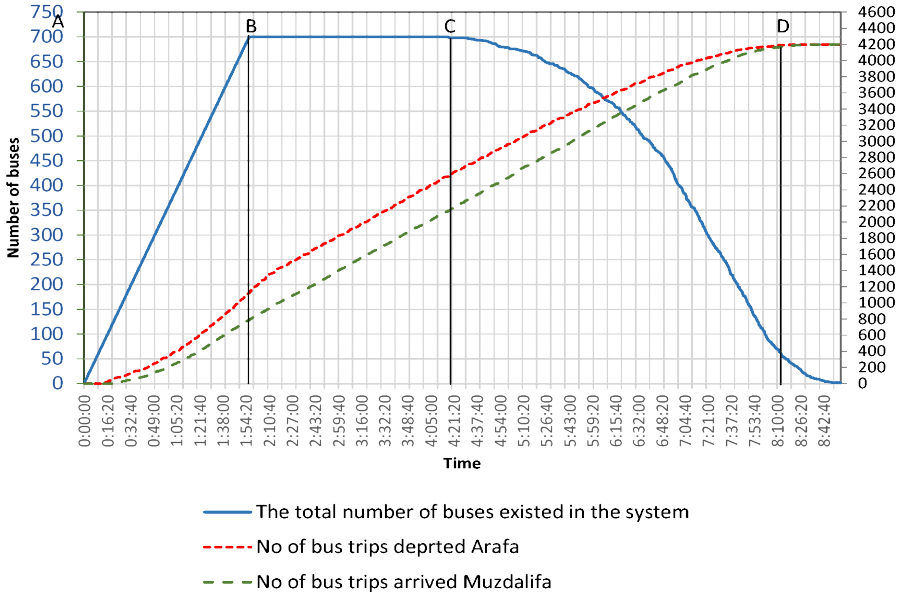
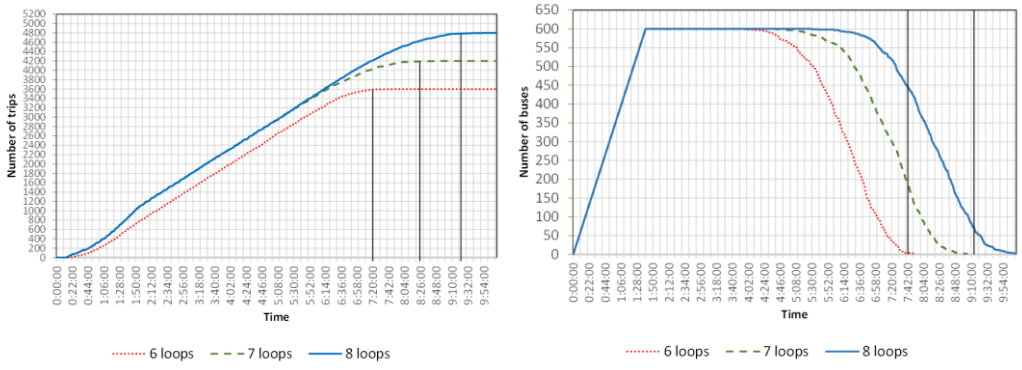


Fig. 5. The total number of buses existed in the system and the number of trips departed Arafa and arrived Muzdalifa for (700 batch of buses and 6 loops).



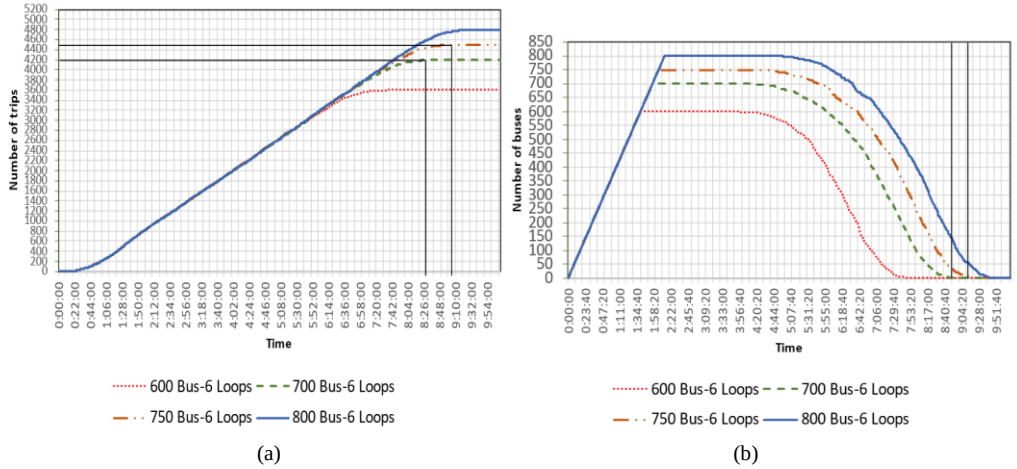
(a) (b)

Fig. 6. The effect of tuning the maximum number of cycles (C_N): (a) The number of trips arrived Muzdalifa for different number of loops; (b) The total number of buses existed in the system for different number of loops.

The results of another simulation experiment that examines the effect of tuning different bus batch sizes (N) are shown in Fig. 7. The figure shows the number of trips arriving in Muzdalifa and the total number of buses existing in the system for different N. It can be depicted from the figure that:

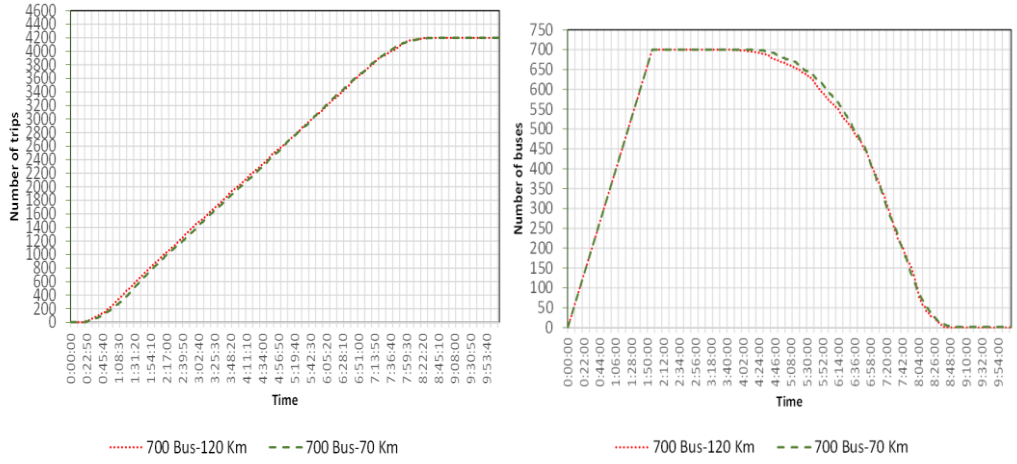
- for N = 600, a total of 3600 trips (180000 pilgrims, or 288000 pilgrims when the type of 80-passenger city-bus is used) were able to reach Muzdalifa within 7 hours and 20 minutes. In addition, the total duration for completing the evacuation (whole buses batch were able to leave the system to the parking area) was 7 hours and 42 minutes.
- for N = 700, a total of 4200 trips (210000 pilgrims, or 336000 pilgrims when the type of 80-passenger city-bus is used) were able to reach Muzdalifa within 8 hours and 26 minutes. In addition, the total duration for completing the evacuation was 8 hours and 48 minutes.
- for N = 750, a total of 4500 trips (225000 pilgrims, or 360000 pilgrims when the type of 80-passenger city-bus is used) were able to reach Muzdalifa within 9 hours. In addition, the total duration for completing the evacuation was 9 hours and 10 minutes.
- for N = 800, a total of 4800 trips (240000 pilgrims, or 384000 pilgrims when the type of 80-passenger city-bus is used) were able to reach Muzdalifa within 9 hours and 32 minutes. In addition, the total duration for completing the evacuation was 9 hours and 40 minutes.

While not always applicable and for the sake of experimenting, buses with faster speed (120 km/h) were tested. The calculated statistics, after running the simulation, are shown in Fig. 8. It can be seen in the figure that increasing the bus speed does not help in increasing the arrival rate in Muzdalifa, as the results are similar to the results when using bus speed of 70 km/h.



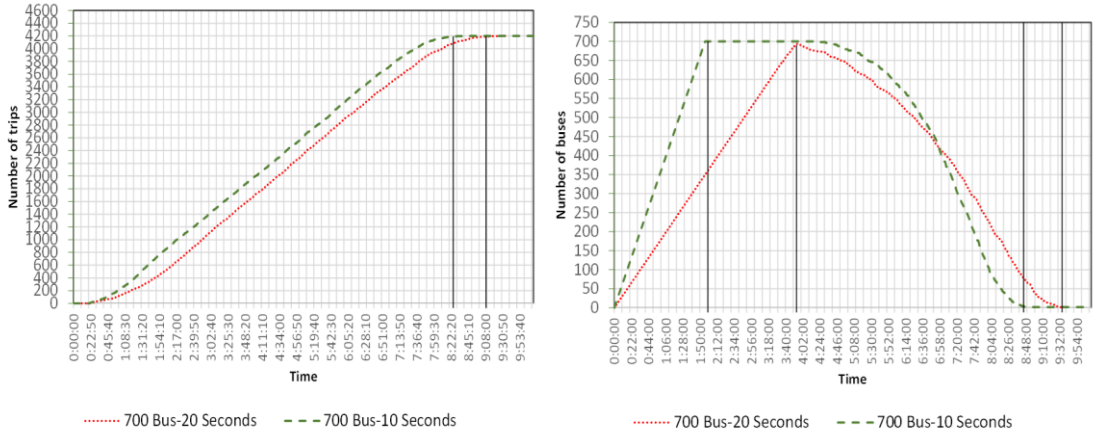
(a) (b)

Fig. 7. The effect of tuning different bus batch sizes (N): (a) The number of trips arrived Muzdalifa for different bus batch sizes; (b) The total number of buses existed in the system for different bus batch sizes.



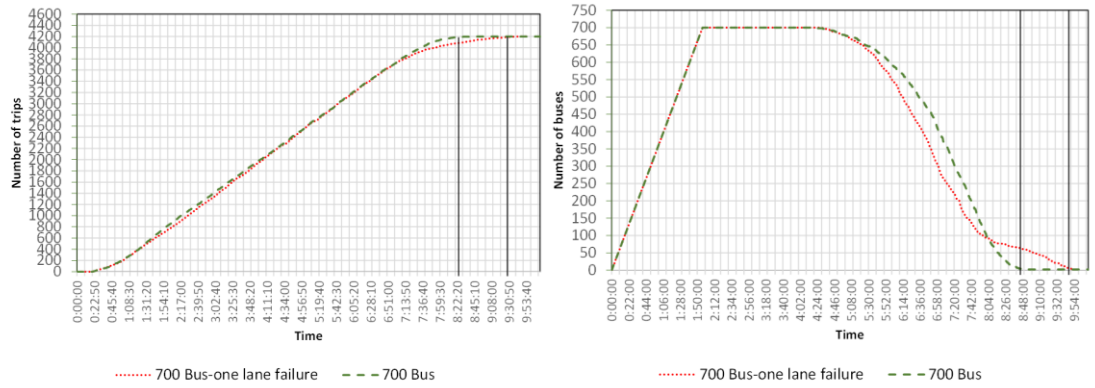
(a) (b)

Fig. 8. The effect of tuning different bus speed (Bs): (a) The number of trips arrived Muzdalifa for different bus speed; (b) The total number of buses existed in the system for different bus speed.



(a) (b)

Fig. 9. The effect of tuning the interarrival time for buses entering the system (IT): (a) The number of trips arrived Muzdalifa for different Bus interarrival times; (b) The total number of buses existed in the system for different Bus interarrival times.



(a) (b)

Fig. 10. The impact of a failure that might occur in one of the road lanes: (a) The number of trips arriving in Muzdalifa; (b) The total number of buses existing in the system.

The results of another simulation experiment that examines the effect of tuning the interarrival time for buses entering the system (I_T), are shown in Fig. 9. It can be seen in this figure that using I_T = 10 seconds, the buses batch size (700) increases gradually and reach a number of 700 buses during around 114 min, and after 4 hours and 15 min., buses start to depart to the parking area. On the other hand, using I_T = 20 seconds, the buses batch size (700) increases gradually and reach a number of 700 buses during around 3 hours and 55 min. and at that time buses start to depart to the parking area, as several buses have already reached the maximum number of cycles. The results of the simulation showed that using I_T = 10 seconds was a suitable choice as it maintains a good balance between the level of roads congestion and the rate at which buses enter the route.

Fig. 10 depicts the results of another simulation experiment that investigates the impact of a failure that might occur in one of the road lanes. The figure depicts the number of trips arriving Muzdalifa and the total number of buses in the system in two scenarios: a scenario with no failure and another scenario with one-lane failure. The figure shows that the presence of one-lane failure causes a slight delay in the movement of the buses. However, the delay is minor because using dedicated roads for traffic helps to keep the level of congestion low.

V. Optimization of the Shuttle-Bus Transportation System in Arafat-Muzdalifa Area

In the optimization process, the results of simulation experiments that include all possible combinations of parameters are registered. As mentioned earlier, the parameters include:

- the overall number of buses dedicated for the transportation of pilgrims (N),
- the maximum number of cycles (loops) that each bus completes across the system before departing for the parking area (C_N),
- the total amount of time for completing the evacuation that all buses spend in the system before departing for the parking area (T), and
- The total number of bus trips that reach Muzdalifa (N_{Trips}).

The objective of our optimization problem is to determine the optimal solution with respect to one or more of the above parameters. This could be the minimum N, the minimum C_N, the minimum T, or the maximum (N_{Trips}). In general, N_{Trips} = N * C_N; hence, maximizing N_{Trips} while minimizing N indicates that C_N must be maximized.

The objective function can be formulated as $\min_{(N,C_N,T,N_{Trips})} S_\omega$, where S_ω is a weighted function that includes the system parameters. This function can be re-written as a weighted function of the parameters as:

$$S_\omega = \omega_1 N + \omega_2 C_N + \omega_3 T + \omega_4 \frac{1}{N_{Trips}} \quad (1)$$

However, the values of the system parameters in the above equation are not within close ranges. Therefore, to have a consistent formula for S_ω , the system parameters in the equation are normalized by dividing each one by its possible maximum value. Such maximum values are clearly determined from the set of all possible combination of parameters that are used in the simulation experiments. Hence, a normalized parameter is obtained as

$$\hat{p} = p/p_{Max}\,, \quad (2)$$

where, p is a parameter value, $\hat{p}$ is a normalized parameter value, and p_{Max} is the maximum possible value of the parameter. The weighted summation function in Eq. (1) is, therefore, reformulated as,

$$S_\omega = \omega_1 \hat{N} + \omega_2 \widehat{C_N} + \omega_3 \hat{T} + \omega_4 \widehat{\frac{1}{N_{Trips}}} \quad (3)$$

To be accepted as a possible solution, the obtained parameters shouldn't exceed predefined limits (the available number of buses, the maximum number of cycles, and the target time for completing the evacuation). In addition, the target number of bus trips that reach Muzdalifa should be satisfied. In the considered system (the shuttle-bus transportation system for pilgrims on the route from Arafat to Muzdalifa (Nafrah)), the parameters are determined using a range of (500 to 850) for N, a range of (3 to 8) for CN, and a range of (1500 to 6800) for NTrips. Table II illustrates the targeted values of the system parameters.

An excerpt of the outcomes from simulation experiments that looked at every possible combination of parameters is shown in Table III.

Eq. (3) can be used either to optimize one parameter or to find an optimal solution with respect to more than one parameter. In the former case, the weight of the targeted parameter is set to 1, while the weights of the other parameters are set to 0. In the latter case, the weights of the selected parameters are set to 1, while the other weights are set to 0. The following experiments illustrate how different sets of weights are used for the optimization process:

- For the maximum number of buses (N), the weights are set to the values of ($\omega_1 = 1, \omega_2 = 0, \omega_3 = 0, \omega_4 = 0$), hence, only the effect of $\hat{N}$ is considered in S_ω. The optimized result of S_ω in this case is 0.59. This optimized result gives a number of possible solutions. The best of them in terms of total evacuation time gives the following values: N=500, C_N = 4, N_{Trips} = 2000, T = 294.92 min. However, this solution is not an acceptable as it doesn't satisfy the target number of bus trips that reach Muzdalifa. The best solution that complies with the predefined limits has the following values: N=500, C_N = 6, N_{Trips} = 3000, which results in a total time (T) of 395.87 min.
- When the optimization process is performed to fine the minimum number of cycles (C_N), the weights are set to the values of ($\omega_1 = 0$, $\omega_2 = 1$, $\omega_3 = 0$, $\omega_4 = 0$), where the effect of $\widehat{C_N}$ is only considered in S_ω. The optimized parameters, in this case, were determined as N=500, C_N = 3, N_{Trips} = 1500, and the total amount of time for completing the evacuation (T) was found to be 240.27 min. Although this solution minimizes the number of cycles, this is not an acceptable solution as it doesn't satisfy the target number of bus trips that reach Muzdalifa. The first solution that complies with the predefined limits has the following values: N=720, C_N = 4, N_{Trips} = 2880, which results in a total time T of 378.88 min.
- To minimize the total amount of time for the evacuation process (T), the weights are set to the values of ($\omega_1 = 0, \omega_2 = 0, \omega_3 = 1, \omega_4 = 0$), which means only the effect of $\hat{T}$ is considered in S_ω.

TABLE II. THE TARGETED VALUES OF THE SYSTEM PARAMETERS

N	≤ 800 buses
C_N	≤ 8 cycles
T	≤ 480 min.
N_{Trips}	≥ 2820 (around 225600 pilgrims, when 80-passenger city-bus is used)

TABLE III. AN EXCERPT OF THE SIMULATION EXPERIMENTS RESULTS

N	C_N	T	N_{Trips}	$1/N_{Trips}$	$\hat{N}$	$\widehat{C_N}$	$\hat{T}$	$\widehat{1/N_{Trips}}$
500	4	294.92	2000	0.000500	0.59	0.50	0.37	0.75
510	4	298.44	2040	0.000490	0.60	0.50	0.37	0.74
530	4	302.33	2120	0.000472	0.62	0.50	0.38	0.71
520	4	314.95	2080	0.000481	0.61	0.50	0.39	0.72
500	5	356.44	2500	0.000400	0.59	0.63	0.44	0.60
500	3	240.28	1500	0.000667	0.59	0.38	0.30	1.00
510	3	242.13	1530	0.000654	0.60	0.38	0.30	0.98
540	4	319.03	2160	0.000463	0.64	0.50	0.40	0.69
520	3	248.80	1560	0.000641	0.61	0.38	0.31	0.96
510	5	366.51	2550	0.000392	0.60	0.63	0.46	0.59
530	3	251.00	1590	0.000629	0.62	0.38	0.31	0.94
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

- In this case, the optimal solution gives the values of N=500, C_N = 3, N_{Trips} = 1500. The total amount of time for completing the evacuation (T) in this case was

found to be 240.27 min. Although this solution resulted in the minimum time, the solution is not acceptable because it doesn't satisfy the target number of bus trips that reach Muzdalifa. The first solution that complies with the predefined limits has the following values: N=720, C_N = 4, N_{Trips} = 2880, which results in a total time T of 378.88 min.

- To satisfy the target number of bus trips (N_{Trips}), a possible scenario is to find the maximum N_{Trips}. Here, the weights are set to the values of ($\omega_1 = 0, \omega_2 = 0, \omega_3 = 0, \omega_4 = 1$), to allow the effect of $\widehat{\frac{1}{N_{Trips}}}$ to be only considered in S_ω. The results of this case were determined as N=850, C_N = 8, N_{Trips} = 6800, in which the total amount of time for completing the evacuation (T) was found to be 801.30 min. Although this solution achieves a large number of bus trips that reach Muzdalifa, it consumes lots of resources, in terms of the number of buses dedicated for the transportation of pilgrims (N), in addition to the maximum number of cycles (loops) that each bus completes across the system. Moreover, the required time for completing the evacuation is long. Hence, this is not an acceptable solution. The first solution that complies with predefined limits has the following values: N=750, C_N = 5, N_{Trips} = 3750, which results in a total time T of 476.75 min.

In order to maintain the effects of all parameters, the weights are set to the values of ($\omega_1 = 1, \omega_2 = 0, \omega_3 = 1, \omega_4 = 1$). The reason for having the value $\omega_2 = 0$ is to solve the contradiction in (4), by omitting the effect of C_N. The optimized parameters, in this case were determined as N=500, C_N = 6, N_{Trips} = 3000, and with these parameters the total amount of time for completing the evacuation (T) was found to be 395.87 min. This can be considered as an accepted solution that achieves the targets while maintaining the system resources within the available limits.

Table IV summarizes the results of the above experiments.

In fact, the values of weights used in Eq. (3) can be chosen to define the significance of each parameter in S_ω. In our experiments, we used equal weights to indicate equal importance of the parameters. However, unequal values of weights will give some parameters more importance than others.

TABLE IV. A SUMMARY OF THE EXPERIMENTS ILLUSTRATING EFFECTS OF VARYING THE WEIGHTS IN THE OPTIMIZATION PROCESS

weights	N	C_N	T	N_{Trips}
$\omega_1 = 1, \omega_2 = 0, \omega_3 = 0, \omega_4 = 0$	500	6	395.87	3000
$\omega_1 = 0, \omega_2 = 1, \omega_3 = 0, \omega_4 = 0$	720	4	378.88	2880
$\omega_1 = 0, \omega_2 = 0, \omega_3 = 1, \omega_4 = 0$	720	4	378.88	2880
$\omega_1 = 0, \omega_2 = 0, \omega_3 = 0, \omega_4 = 1$	750	5	476.75	3750
$\omega_1 = 1, \omega_2 = 0, \omega_3 = 1, \omega_4 = 1$	500	6	395.87	3000

VI. CONCLUSIONS

In this paper, a computer modeling of pilgrims' shuttle-bus transportation system in Mashaaer Holy Sites, were presented. The model was designed using a multimodal modeling and simulation tool developed by AnyLogic to include operations during the pilgrims' transport from Arafat to Muzdalifa via shuttle-buses. The proposed model can be used for optimizing the pilgrim's transportation system with respect to different parameters since the results of the simulation may be used to assess the effectiveness of the current system and to suggest changes that might be made. In our experiments, we used equal weights to indicate equal importance of the parameters. In future work, unequal values of weights of parameters can be set to give more importance of some parameters than others. In addition, this paper focused on one of the paths of the hajj shuttle bus system. As a future work, the simulation model may be extended to include other paths.

REFERENCES

[1] L. Jago, L. Dwyer, G. Lipman, D. van Lill, and S. Vorster, "Optimising the potential of mega-events: an overview," International Journal of Event and Festival Management, vol. 1, no. 3, pp. 220–237, Oct. 2010, doi: https://doi.org/10.1108/17852951011078023.

[2] B. S. Taha and A. Allan, "Hosting Mega Event - Drive towards Sustainable Planning for Public Transport - Case Study: Metro Line Route 2020," Transportation Research Procedia, vol. 48, pp. 2176–2186, 2020, doi: https://doi.org/10.1016/j.trpro.2020.08.274.

[3] F. A. Girgin and O. E. Tasci, "Mega-Event Organization Considering Safety, Security and Resilience," TeMA - Journal of Land Use, Mobility and Environment, vol. 12, no. 3, pp. 249–264, Dec. 2019, doi: https://doi.org/10.6092/1970-9870/6269.

[4] O. Tayan, A. M. Al BinAli, and M. N. Kabir, "Analytical and Computer Modelling of Transportation Systems for Traffic Bottleneck Resolution: A Hajj Case Study," Arabian Journal for Science and Engineering, vol. 39, no. 10, pp. 7013–7037, Jun. 2014, doi: https://doi.org/10.1007/s13369-014-1231-3.

[5] Y. Cao and J. Wang, "The Key Contributing Factors of Customized Shuttle Bus in Rush Hour: A Case Study in Harbin City," Procedia Engineering, vol. 137, pp. 478–486, 2016, doi: https://doi.org/10.1016/j.proeng.2016.01.283.

[6] X. Kong, M. Li, T. Tang, K. Tian, L. Moreira-Matias, and F. Xia, "Shared Subway Shuttle Bus Route Planning Based on Transport Data Analytics," IEEE Transactions on Automation Science and Engineering, vol. 15, no. 4, pp. 1507–1520, Oct. 2018, doi: https://doi.org/10.1109/TASE.2018.2865494.

[7] F. Abdessemed, "An integrated system for tracking and control pilgrims shuttle buses," in 2011 14th Int. IEEE Conf. Intell. Transp. Syst. - (ITSC 2011), Washington, DC, USA, Oct. 5–7, 2011. IEEE, 2011. Accessed: May 22, 2023. [Online]. Available: https://doi.org/10.1109/itsc.2011.6083024

[8] O. Hussain, E. Felemban, and F. U. Rehman, "Optimization of the Mashaer Shuttle-Bus Service in Hajj: Arafat-Muzdalifah Case Study," Information, vol. 12, no. 12, p. 496, Nov. 2021, doi: https://doi.org/10.3390/info12120496.

[9] E. Felemban, F. U. Rehman, A. A. Biabani, A. Naseer, O. Hussain, E. U. Warriach, "An Interactive System for Analyzing Movement of Buses in Hajj," J. Theor. Appl. Inf. Technol., 98, 3468–3481, 2020.

[10] G. Currie and A. Shalaby, "Synthesis of Transport Planning Approaches for the World's Largest Events," Transport Rev., vol. 32, no. 1, pp. 113–136, Jan. 2012. Accessed: May 22, 2023. [Online]. Available: https://doi.org/10.1080/01441647.2011.601352

[11] S. Al-Sabban and H. Ramadan, "A Simulation Study of the Shuttle-Bus Pilgrim Transportation System between the Holy Sites for the 1422H Hajj Season," J. King Abdulaziz University-Eng. Sci., vol. 16, no. 2, pp.

71–93, 2005. Accessed: May 22, 2023. [Online]. Available: https://doi.org/10.4197/eng.16-2.5.

[12] M. Seliaman, S. Duffuaa, A. Andijani, "Stochastic Simulation Model for the Design of a Shuttle Bus System to Transport Pilgrims in Hajj, " Researchgate, 2013.

[13] X. Guo, R. Song, S. He, M. Bi, and G. Jin, "Integrated Optimization of Stop Location and Route Design for Community Shuttle Service," Symmetry, vol. 10, no. 12, p. 678, Nov. 2018. Accessed: May 22, 2023. [Online]. Available: https://doi.org/10.3390/sym10120678.

[14] J. Xiong, W. Guan, L. Song, A. Huang, and C. Shao, "Optimal Routing Design of a Community Shuttle for Metro Stations," J. Transp. Eng., vol. 139, no. 12, pp. 1211–1223, Dec. 2013. Accessed: May 22, 2023. [Online]. Available: https://doi.org/10.1061/(asce)te.1943-5436.0000608.

[15] D. Bao, J. Gu, Z. Di, and T. Zhang, "Optimization of Airport Shuttle Bus Routes Based on Travel Time Reliability," Math. Problems Eng., vol. 2018, pp. 1–12, 2018. Accessed: May 22, 2023. [Online]. Available: https://doi.org/10.1155/2018/2369350.

[16] Z. Sun, K. Zhou, X. Yang, X. Peng, and R. Song, "Optimization Method of Customized Shuttle Bus Lines under Random Condition," Algorithms, vol. 14, no. 2, p. 52, Feb. 2021. Accessed: May 22, 2023. [Online]. Available: https://doi.org/10.3390/a14020052.

[17] R. M. Noor, N. B. G. Rasyidi, T. Nandy, and R. Kolandaisamy, "Campus Shuttle Bus Route Optimization Using Machine Learning Predictive Analysis: A Case Study," Sustainability, vol. 13, no. 1, p. 225, Dec. 2020. Accessed: May 22, 2023. [Online]. Available: https://doi.org/10.3390/su13010225.

[18] Z. Liu et al., "Data-driven simulation-based planning for electric airport shuttle systems: A real-world case study," Appl. Energy, vol. 332, p. 120483, Feb. 2023. Accessed: May 22, 2023. [Online]. Available: https://doi.org/10.1016/j.apenergy.2022.120483.

[19] Q. Wang, D. Sigler, Z. Liu, A. Kotz, K. Kelly, and C. Phillips, "ASPIRES: Airport Shuttle Planning and Improved Routing Event-driven Simulation," Transp. Res. Record: J. Transp. Res. Board, p. 036119812210957, Jun. 2022. Accessed: May 22, 2023. [Online]. Available: https://doi.org/10.1177/03611981221095744.

Development of Nursing Process Expert System for Android-based Nursing Student Learning

Aristoteles[1]*, Abie Perdana Kusuma[2], Anie Rose Irawati[3],
Dwi Sakethi[4], Lisa Suarni[5], Dedy Miswar[6], Rika Ningtias Azhari[7]
Department of Computer Science - Faculty of Mathematics and Natural Sciences, Lampung University, Indonesia[1, 2, 3, 4]
Department of Nursing, Polytechnic of Tanjung Karang, Bandar Lampung, Indonesia[5]
Department of Geography Education - Faculty of Teacher Training and Education, Lampung University, Indonesia[6]
Master's Program in Computer Science - Faculty of Mathematics and Natural Sciences, Lampung University, Indonesia[7]

***Abstract*—Nurses are professionals who provide health services using a scientific process called nursing. In nursing, problem-solving uses the nursing process which is a critical thinking method, nurses must analyze the data found in patients to diagnose and determine the results and appropriate intervention plans. Prospective nursing students are required to be able to apply the nursing process in carrying out nursing care according to existing nursing standards, of course, with supervision by nursing experts to improve the quality of medical services. This study aims to develop an android application with the help of an expert system as a nursing diagnosis tool, which helps nursing students learn the nursing process and helps lecturers monitor the nursing process carried out by nursing students. This research uses 116 symptom data, 22 diagnosis data, 60 intervention data, 8 type data, and 864 description data. The results of this research are in the form of an expert system with an android-based forward chaining method that has been tested using the black box testing method.**

***Keywords*—Classification; expert system; forward chaining; blackbox testing; android; flutter; nursing process**

I. INTRODUCTION

Nursing service is a way for nurses to provide professional services, both emotional and otherwise, to individuals, families, groups, and communities, for the healthy and the sick. [1]. Nursing services are part of health services, and the quality of nursing services in general determines the quality of health services [2]. Nurses are professionals who provide medical services using a scientific process called nursing, problem-solving using the nursing process. The nursing process is used as a tool for nurses to practice nursing systematically in solving nursing problems [3]. Nurses provide care according to established standards. This standard was developed by the Indonesian National Nurses Association (PPNI). RI Law No. 36 of 2009 concerning Health has mandated that health workers are obliged to meet professional standards and respect patient rights [4]. The relationship between quality and standards are two things that are very closely related, because through these standards it can be measured that service is improving or even getting worse [5].

The standards ensure that nurses make appropriate and rational decisions and perform interventions that are safe and legally responsible [6]. Nursing students are prospective nurses according to their expertise, they must be able to apply the nursing process in the performance of nursing care, therefore vocational nurses (D3) and nurses are equipped with knowledge of the nursing process in education. Students can utilize expert system applications to support the learning process of nursing diagnoses and appropriate interventions by the Indonesian Nursing Diagnosis Standards (SDKI). In the diagnostic process, experienced experts are needed to provide the correct conclusions [7].

An expert system is a computer-based system that uses knowledge, facts, and reasoning techniques to solve a particular problem which normally can only be solved by experts in their field [8]. The nursing process Expert System is designed to be able to assist nurses in analyzing patient data so that interventions and outputs can be determined according to the conditions and circumstances of the patient.

There are several studies on diagnostic expert systems but for expert systems that produce nursing diagnoses to help nursing students there is still no. Research in [9] discusses the creation of an expert system to diagnose eye disease with the android-based certainty factor method. This Android-based expert system has 75% accuracy results with details of 15 diseases and 52 symptoms of eye disease.

In the study [10] built an android-based expert system to detect dental and oral diseases. The data used is 13 diseases and 44 symptoms. This dental and oral disease detection expert system has an accuracy of 100% if tested with black box testing, while the test results of the User Acceptance test have an accuracy of 93.03%.

Then in the study [11] built an android-based expert system to detect liver disease. This study used 64 test data. The method used in this study is Fuzzy Tsukamoto to diagnose liver disease. The test data was then tested and has an accuracy of 96.87%.

Furthermore, the study [12] built an expert system to diagnose diseases in chili plants. The data used is in the form of 37 symptom data, 10 chili disease data, and 10 rules. The method used is forward chaining, testing carried out using black box testing gives the expected results from each test class. As for the test results from the User Acceptance Test obtained an average of 84%.

In research [13], the researcher developed an expert system that can identify stroke symptoms using the naive

*Corresponding Author

bayes method. In this study using data from patient consultation results at Dr. M Hatta Bukittinggi Brain Hospital for seven months in 2021. Drs. M Hatta Bukittinggi for seven months in 2021. The research has an accuracy of 100% in identifying the type of stroke disease from 10 data samples used.

Based on previous related research, the expert system to improve the quality of nursing quality still does not exist. Meanwhile, in the era of the Industrial Revolution like now, nursing students must be able to adapt to existing technological advances. In research [14], it is said that technological developments really need further development to get better results, in order to get accurate results. Therefore, this research aims to develop an information system for information system for monitoring the nursing process with the addition of an expert system that uses the forward chaining method as a nursing diagnosis tool based on android application, to help nursing students in learning and improving the ability of the nursing process as well as assisting lecturers in monitoring the process carried out by prospective students. Students will enter input first by analyzing the symptoms that have occurred, and the expert system that will be developed will save the results of the student's work. Furthermore, lecturers will be able to monitor and see whether students have correctly analyzed or not.

II. Materials and Methods

The workflow of this research is divided into two stages, namely the 1st stage of building an expert system for nursing diagnoses, and the 2nd stage of implementing the android application using the waterfall system development method. This research workflow can be seen in Fig. 1.

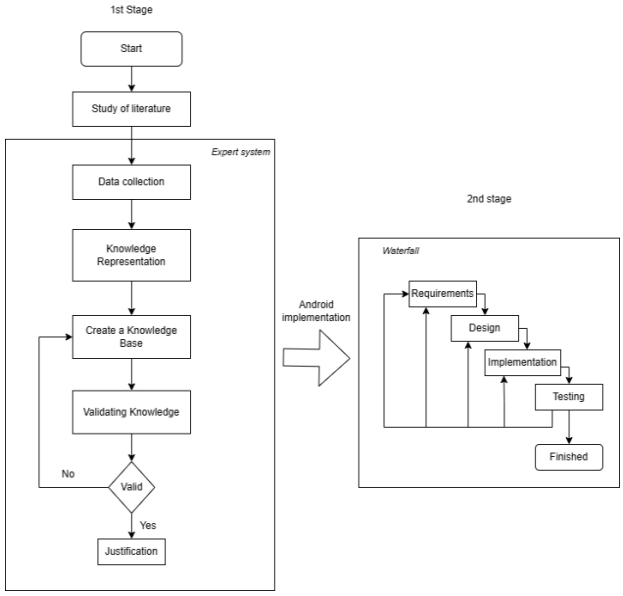


Fig. 1. Research workflow.

A. Building an Expert System

At the expert system development stage, several steps are carried out as follows.

1) Data collection: It was carried out by conducting literature studies and interviewing experts in the field of nursing. The results of the collection are in the form of data needed by expert systems such as symptom data, results of nursing diagnoses, interventions, and descriptions of interventions. In addition, an overview of the system requirements to be built is also obtained.

2) Representing knowledge: The goal at this stage is to represent the knowledge that has been collected to build a knowledge base. Examples of knowledge in this study are symptoms. Each of these symptoms has a different code and has its score as well. Scores on symptoms ranging from a score of 1 to 5 are obtained from interview data with nursing experts, which will later be used in the diagnosis process, where the diagnosis will only be made if the total number of selected overall scores is more than 4 scores. The symptom data can be seen in Table I.

Table I is part of the symptom data collection needed in this study. Symptom data is denoted by the letter "G", and this symptom data has 116 data. After collecting symptom data, then the next step is to collect diagnosis data to enter the next stage. The following are some of the diagnosis data used in this study.

Table II is some data from the results of diagnosis data collection. Diagnosis data has a total of 22 data. Diagnosis data is symbolized by "D" to simplify the running of the system. From the diagnosis data above, researchers also need intervention data. Here is some intervention data used in this study.

TABLE I. Symptom Data

Code	Symptoms	Score
G1	Cough is not effective	5
G2	Can't cough	4
G3	Excess sputum	3
G4	Wheezing, and or dry rhonchi	2
G5	Meconium in the airway (in neonates)	1
G6	Use of accessory muscles of respiration	5
G7	Abnormal breathing pattern	4
G8	The expiratory phase is prolonged	3
G9	Pursed lip breathing	2
G10	Vital capacity decreases	1
G11	Complain of nausea	5
G12	Feeling like vomiting	4
G13	Not interested in eating	3
G14	Sour taste in the mouth	2
G15	Increased saliva	1
G16	Complain of pain	5
G17	Looked grimaced	4
G18	Being protective (positioning to avoid pain)	3
G19	Nervous	2
G20	Focus on yourself	1

TABLE III. DIAGNOSIS DATA

Code	Diagnosis
D1	Ineffective airway clearance
D2	Ineffective breathing pattern
D3	Nausea
D4	Acute pain
D5	Childbirth pain

TABLE IV. INTERVENTION DATA

Code	Intervention
A1	Effective coughing exercises (A)
A2	Airway management (B)
A3	Respiration monitoring
A4	Airway management
A5	Respiration monitoring

Table III above is some data from the intervention in this study. Intervention data has a total of 80 data. The diagnosis data will be coded with the letter "A" to facilitate the running of the system.

3) Creating a knowledge base: The knowledge representation from an expert that the system needs to solve certain problems. The knowledge that has been obtained is made into rules; these rules consist of relationships between existing data such as symptoms, diagnosis results, type interventions, and intervention descriptions.

4) Validate the knowledge: Validate the knowledge base by experts so that there is no mission of interpretation between an expert's knowledge and the knowledge base created.

5) Justification: The final stage, namely justification, the system can already provide the results of nursing diagnoses based on existing symptoms and descriptions of interventions that are suitable for the results of these diagnoses.

B. Implementation of Android Applications

The development of an Android application is carried out using the Flutter framework. The method used in the manufacture of the system is the "waterfall" method, which takes a systematic approach starting from the stage: Requirement, Design, Implementation, until Testing. The following describes each stage.

1) Requirement: The requirement is the stage of identifying the needs of the system to be built and being a reference in determining what functions need to be developed. Here are the system requirements in research that can be seen in Fig. 2.

In the use case diagram, it has been shown that there are two levels of users, namely students, and lecturers. In the use case diagram, limitations have also been given on what each type of user can do on the system. Students can perform the nursing process on patients and lecturers can monitor the results of the nursing process carried out by their students.

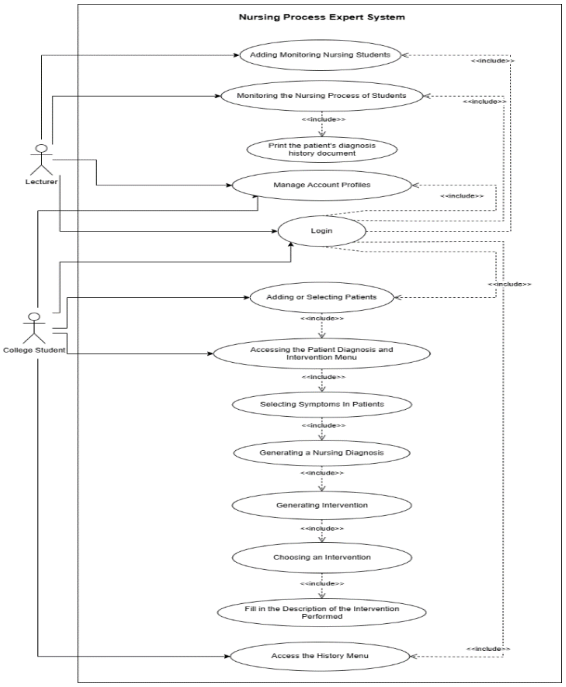


Fig. 2. Use case diagram.

2) Design: The next stage is to make a system development plan into design forms such as ERD designs and activity diagram designs. The following is the ERD system which will be shown in Fig. 3.

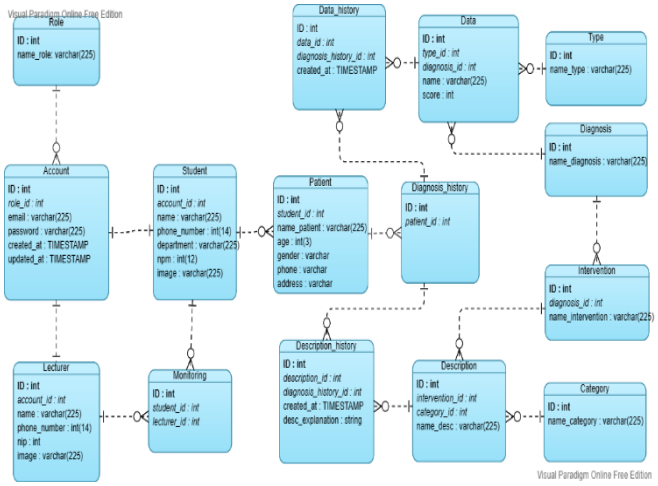


Fig. 3. Entity relationship diagram.

Entity Relation Diagram explains the relationship between tables in the database that the system uses. In the tables, some data is used in expert systems. These data consist of 116 symptom data, 22 diagnosis data, 60 intervention data, 8 type data, and 864 description data. The activity design for students explains how the system flows if the logged-in user is a student, students can carry out the patient nursing process from determining the type of diagnosis, selecting symptoms, generating diagnoses, selecting interventions, and filling out appropriate intervention descriptions for patients. The following activity design for nursing students can be seen in Fig. 4.

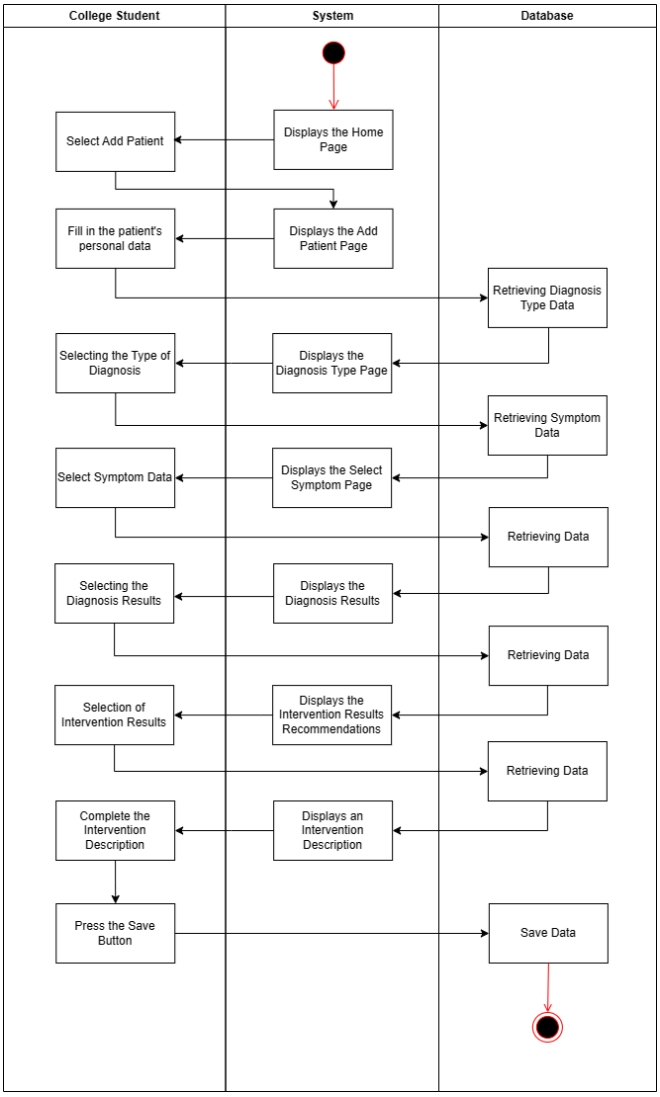


Fig. 4. Nursing process activity by nursing students.

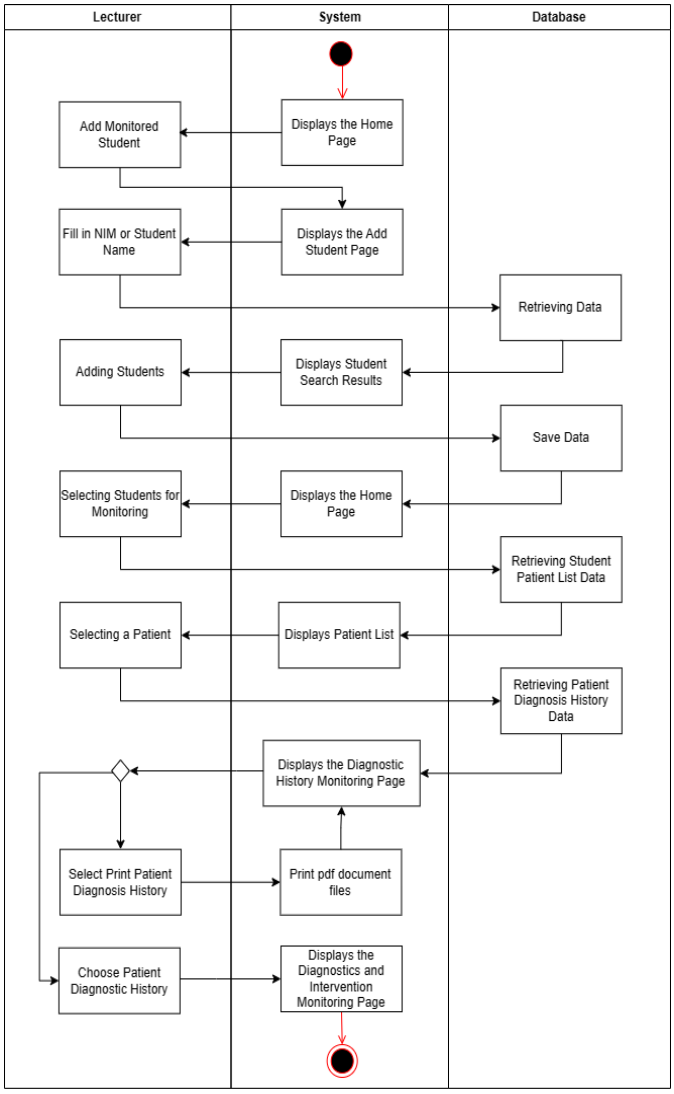


Fig. 5. Activity monitoring nursing process.

Fig. 4 explains the flow of how students diagnose patients, starting from students adding patients first by filling in the data of the patient to be diagnosed, followed by students starting to diagnose from symptoms to the end until the system saves patient data that has been diagnosed by students. The activity plan for the lecturer explains the flow of how the system runs if the user who login is a lecturer, lecturers can monitor the results of the nursing process carried out by their students, starting from a nursing process carried out by their students starting from adding a student to being monitored. Students, who are monitored, view all patient lists who have been treated by students, to see all the history of the nursing process that has been carried out. The following activity plan for lecturers can be seen in Fig. 5.

Fig. 5 explained the flow of how lecturers monitor the results of their student nursing process, starting from lecturers adding students to the system of displaying the history of the patient's nursing process by students.

3) Implementing: All system designs that have been created including expert system functions to help students learn in the nursing process into the form of program code built using dart language, with the help of Flutter frameworks to build android applications and Laravel8 frameworks to build system database APIs.

4) Testing: After implementation, the next stage is testing. Testing will be done using the Black-Box Testing method. In addition to testing the system using black-box, user acceptance testing of the built system is also carried out, the target of system users, namely students and nursing science lecturers.

IV. Result and Discussion

A. Knowledge Base

The knowledge base is the core of the expert system, which is a representation of knowledge from experts. The knowledge base shown in Table IV is a description of the relationship between symptom data and other data contained in the database. Here is the knowledge base that can be seen in Table IV.

TABLE V. Expert System Knowledge Base

Type	Symptom	Diagnosis Results	Intervention	Description
T1	G1 G2 G3 G4 G5	D1	A1	U1, U2, U3, U4, U6, U10, U5, U7, U8, U9
			A2	U12, U13, U14, U 15, U16, U17, U18, U19, U20, U21, U22, U23, U24, U25
			A3	U26, U27, U28, U29, U30, U31, U32, U33, U34, U35, U36, U37, U38, U39, U40
	G6 G7 G8 G9 G10	D2	A4	U41, U42, U43, U44, U45, U46, U47, U48, U49, U50, U51, U52, U53, U54
T2	G11 G12 G13 G14 G15	D3	A6	U55, U56, U57, U58, U59, U60, U61, U62, U63, U64, U65, U66, U67, U68, U69, U70
			A7	U71, U72, U73, U74, U75, U76, U77, U78, U79, U80, U81, U82, U83, U84, U85, U86, U87, U88, U89
	G16 G17 G18 G19 G20	D4	A8	U90, U91, U91, U93, U94, U95, U96, U97, U98, U99, U100, U101, U102, U103, U104, U105, U106, U107, U108
			A9	U109, U110, U111, U112, U113, U114, U115, U116, U117, U118, U119

Table IV above is part of the expert system knowledge base consisting of the data used, for symptoms coded with the letter "G", for diagnoses coded "D", interventions coded "A", types coded "T" and descriptions coded "U". The results of the system diagnosis will appear if the main requirements are met, namely if the total score of the selected symptoms is more than equal to 5 scores (∑Symptom Score ≥ 5). After the results of the system diagnosis appear and are selected, the results of the intervention and description of the intervention will appear following the options.

1) User interface: The user interface is part of an information system that requires user interaction to create input and output for the system [15]. The User Interface is used to provide an overview of the appearance of the application that will be used by the user.

B. System Testing Results

System testing is carried out by testing the system functionally using the black-box testing method and non-functionally, the system aims to determine the response and assessment of users of the system created.

1) Black-box testing: Black-box testing is a software testing technique that focuses on the functional specifications of software [16]. This test is carried out by examiners as well as application users consisting of lecturers and nursing students. All test results obtained from the examiner are successful for each test case carried out.

2) User acceptance testing: This test is carried out with a focus on the non-functional attributes of the system. The main goal is to develop software that can meet user requirements. Not only meets system specifications and can be used but also to find out whether the system is acceptable to users or not [17]. Technical testing was carried out by the way respondents downloaded and ran the application, after using the application respondents were directed to fill out the questionnaire. Here are the results of the questionnaire in this test.

TABLE VI. User Acceptance Test Results

No.	Question	Index (%)
1	The application is easy to understand how to use it	**87,83%**
2	The application interface display is good and easy to understand	**85,22%**
3	The images and icons used in the application display are easy to understand	**86,96%**
4	The instructions in the application are clear and easy to understand	**86,1%**
5	I feel comfortable while using the app	**81,74%**
6	The application helps nursing students in making nursing diagnoses	**86,96%**
7	The diagnostic results given are by the facts	**81,74%**
8	The application helps nursing students in nursing learning	**86,96%**
9	This application can assist lecturers in monitoring the nursing process of students	**73,33%**
Average Index		**84%**

Table V is a summary of the test results conducted by each user. Testing was carried out by students and lecturers, 22 students and three lecturers. Testing is done with nine questions that must be filled in by each user whether the features in the system are running properly; the functionality of each feature will be tested by students and lecturers. The results of the test show that this research has an average index value of 84% which means that the system built has a "Very Good" value.

V. Conclusion

In this study, the expert system for the learning process of nursing students is considered capable of helping nursing students to properly analyze the diagnosed patients and can assist lecturers in monitoring every nursing process carried out by students. The expert system application that was successfully built using the Flutter framework has a test value of "Very Good" with an average value obtained of 84%.

Recommendations for future researchers are that future researchers can add and complete data that is in accordance with SDKI (Indonesian Nursing Diagnosis Standards) and SIKI (Indonesian Nursing Intervention Standards), add assessment features, corrections and notes for lecturers on each treatment result carried out by students, add features to replace lecturers for students, add features to remove students from the monitoring list for lecturers, add features for students to be able to add patient progress notes and add features for lecturers to evaluate the patient care process carried out by students.

ACKNOWLEDGMENT

We would like to express our deepest thanks to LPPM Lampung University for providing research funding for this project. The financial support provided has enabled us to obtain the data and resources necessary to complete this research.

REFERENCES

[1] Hadjam, M. N. R. (2001). Efektivitas Pelayanan Prima Di Rumah Sakit. *Jurnal Psikologi*, *1*(2), 105–115.

[2] Pakpahan, H. M., Sigalingging, G., and Simbolon, R. (2022). Hubungan mutu pelayanan keperawatan dengan kepuasan pasien rawat inap di RSIA Stella Maris Medan. *Jurnal Darma Agung Husada*, *9*(1), 14–23.

[3] Syaifullah, R. M. (2018). Peran Perawat Pada Proses Keperawatan Dalam Memberikan Asuhan Keperawatan. *Jurnal Keperawatan*, *1*(1), 1–11.

[4] PPNI. (2009). *Standar profesi dan kode etik perawat Indonesia*. Pengurus Pusat Persatuan Perawat Nasional Indonesia (PP-PPNI).

[5] Sitorus, H. (2013). Pengaruh Model Supervisi Klinik Terhadap Kinerja Perawat Dalam Asuhan Keperawatan di Instalasi Rawat Inap 2 RS TNI Jakarta. *Jurnal Keperawatan*, *1*(1), 1–10.

[6] Lestari, T. R. P. (2014). Harapan Atas Profesi Keperawatan di indonesia. *Jurnal Keperawatan*, *19*(1), 51–67.

[7] Aristoteles, Adhianto, K., Andrian, R., and Sari, Y. N. (2019). Comparative analysis of cow disease diagnosis expert system using Bayesian network and Dempster-Shafer method. *International Journal of Advanced Computer Science and Applications*, *10*(4), 227–235.

[8] Sulistyohati, A., and Hidayat, T. (2008). Aplikasi Sistem Pakar Diagnosa Penyakit Ginjal Dengan Metode Dempster-Shafer. *Seminar Nasional Aplikasi Teknologi Informasi*, *1*(1), 1–6.

[9] Permana, Y., Wijaya, I. G. P. S., and Bimantoro, F. (2017). Sistem Pakar Diagnosa Penyakit Mata Menggunakan Metode Certainty Factor Berbasis Android (Android Based Expert System for Eye Diseases Diagnosis using Certainty Factor). *Journal of Computer Science and Informatics Engineering (J-Cosine)*, *1*(1), 1–10.

[10] Arfajsyah, H. S., Permana, I., and Salisah, F. N. (2018). Sistem Pakar Berbasis Android Untuk Diagnosa Penyakit Gigi Dan Mulut. *Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi*, *4*(2), 110–117.

[11] Falatehan, A. I., Hidayat, N., and Brata, K. C. (2018). Sistem Pakar Diagnosis Penyakit Hati Menggunakan Metode Fuzzy Tsukamoto Berbasis Android. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *2*(8), 2373–2381.

[12] Aristoteles, Fuljana, M., Prasetyo, J., and Muludi, K. (2017). Expert System of Chili Plant Disease Diagnosis using Forward Chaining Method on Android. *International Journal of Advanced Computer Science and Applications*, *8*(11), 164–168.

[13] Karim, F., Nurcahyo, G. W., & Sumijan, S. (2021). Sistem Pakar dalam Mengidentifikasi Gejala Stroke Menggunakan Metode Naive Bayes. *Jurnal Sistim Informasi dan Teknologi*, *3*, 221–226.

[14] Aristoteles, A., Syarif, A., Sutyarso, S., & Lumbanraja, F. R. (2022). Identification of human sperm based on morphology using the you only look once version 4 algorithm. International Journal of Advanced Computer Science and Applications, 13(7), 424-431.,

[15] Elfida, M., and Nasution, M. K. M. (2005). Perancangan Antarmuka Sistem Informasi. *Al-Khawarizmi: Journal of Computer Science*, *1*(1), 11–17.

[16] Jan, S. R., Shah, S. T. U., Johr, Z. U., Shah, Y., and Khan, F. (2016). An Innovative Approach to Investigate Various Software Testing Techniques and Strategies. *International Journal of Scientific Research in Science, Engineering, and Technology*, *2*(2), 682–689.

[17] Hady, E. L., Haryono, K., and Rahayu, N. W. (2020). User Acceptance Testing (UAT) of the Prototype of Students ' Savings Information System (Case Study : Al-Mawaddah Islamic Boarding School). *Jurnal Ilmiah Multimedia dan Komunikasi*, *5*(1), 1–10.

Emotional State Prediction Based on EEG Signals using Ensemble Methods

Norah Alrebdi[1], Amal A. Al-Shargabi[2]
Department of Cybersecurity, College of Engineering & Information Technology,
Buraydah Private Colleges, Buraydah, 51418, Kingdom of Saudi Arabia[1]
Department of Information Technology, College of Computer, Qassim University, Buraydah, 51452, Saudi Arabia[2]

***Abstract*—The emotional state is an essential factor that affects mental health. Electroencephalography (EEG) signal analysis is a promising method for detecting emotional states. Although multiple studies exist on EEG emotional signals classification, they have rarely considered processing time as a metric for classification model evaluation. Instead, they used either model accuracy and/or the number of features for evaluation. Processing time is an important factor to be considered in the context of mental health. Many people commonly use smart devices, such as smartwatches to monitor their emotional state and such devices require a short processing time. This research proposes an EEG-based model that detects emotional signals based on three factors: accuracy, number of features, and processing time. Two feature extraction algorithms were applied to EEG emotional signals: principal components analysis (PCA) and fast independent components analysis (FastICA). In the classification process, ensemble method classifiers were adopted due to their powerful performance. Three ensemble classifiers were used: random forest (RF), extreme gradient boosting (XGBoost), and adaptive boost (AdaBoost). The experimental results showed that the RF and XGBoost achieved the best accuracy, i.e., 95%, for both methods. However, XGBoost outperformed RF in terms of the number of features; it used 33 components extracted by PCA within 14 seconds, while RF used 36 within 4 seconds. AdaBoost was the worst in terms of both accuracy and processing time in the two experiments.**

Keywords*—*Electroencephalograph; mental health; feature extraction; random forest; extreme gradient boosting; adaptive boost

I. Introduction

Mental health is a source of concern due to its significant impact on an individual's quality of life and on society, where poor mental health can lead to multiple health and financial losses, in addition to suicide in critical cases. Mental health issues are increasing due to different factors such as social media, social state, and financial state. Additionally, natural disasters and global epidemics affect an individual's mental health, such as what is caused by the Covid-19 pandemic [1]. In 2011, the World Health Organization predicted that depression would be the leading cause of the global illness burden by 2030 [2]. Twenty-five percent of people worldwide have mental health problems [3]. Hence, it is essential to pay attention to this field and find solutions to mitigate the expected impacts of poor mental health.

Furthermore, personal emotions primarily affect people's mental health [4]. EEG signals contain brain electrical activity information gathered from the scalp by electrodes [5]. Thus, several studies have been concerned with conducting experiments classifying emotions using EEG signals. Multiple previous studies have focused on classification accuracy and the number of features. The previous experiments achieved high accuracy with an appropriate number of components in the EEG signal classification process [6], but ignored processing time. On the other hand, limited studies have considered time to be an essential factor for the model evaluation. However, their results have shown that the processing took a relatively long time, impairing the model's efficiency. A short processing time is a required factor in mental health data processing, enabling smart devices to adopt these models and allowing their users to monitor their emotional states through them directly. Moreover, fast processing to detect personal emotions contributes to preventing critical cases, such as suicide due to depression.

This gap encouraged us to propose a model that classifies emotional signals considering accuracy, number of features, and processing time. PCA and FastICA were used in the extraction feature stage due to their speed in the extraction, while RF, AdaBoost, and XGBoost were implemented to classify the emotions.

The contributions of this study are summarized as follows:

- Implementing three ensemble classification algorithms on EEG signals using two feature extraction algorithms.
- Defining the best combination of feature extraction and classification algorithms based on three factors: accuracy, number of features, and processing time.
- Comparison of the proposed model results with several studies that used the same dataset.

II. Literature Review

Table I shows a summary of related works in terms of the dataset, features extraction algorithm, number of features, best accuracy, and processing times. As shown in the table, few studies have been concerned with the models' processing times. On the other hand, various datasets and algorithms were used.

TABLE I. SUMMARIZATION OF RELATED WORK

Ref	Dataset	Best Feature Algorithm	Number of features/channels	Best Accuracy	Time (s)
[7]	Emotional state	Biologically inspired computing	500 features	AdaBoost LSTM: 97.06%	595
[8]	Emotional state	Info-Gain	63 features	RF: 97.89%	-
[9]	Emotional state	-	-	3D CNN: 98.43%	-
[10]	SEED	-	62 channels	STRNN: 89.5%	-
[11]	SEED	PCA and t-statistics	5 channels	ANN: 86.57%	-
[12]	DEAP	PCA	-	F1-score SVM: 84.73% in Arousal class	-
[13]	DEAP	-	-	KohonenNN I: 87%	-
[14]	DEAP	differential entropy	-	ECLGCNN: 90.60% of Arousal class	-
[15]	DEAP	EMD, SODP	-	MLP: 100% in high and low of arousal	-
[16]	Preprocessed version of DEAP	Statistical	14 channels	SVM: 78.96% of arousal class	-
[17]	DEAP and DREAMER	spatial and temporal information	-	Deep Forest: 97.69% of valence class 90.41% of arousal class	693 1307
[18]	SEED and DEAP	FAWT, IP	-	SVM: SEED: 83.33% DEAP: 59.06%	-
[19]	SEED and DEAP	variance, DWT, FFT	SEED: 62 input neurons DEAP: 32 input neurons	SNNs SEED: 96.67%. DEAP: 86.27% in liking class.	-
[20]	Clips of Indian films	EMD, VMD	-	MC-LS-SVM with MD: 93.13% of relax	-
[21]	sad, happy, fear, and neutral emotions dataset	sub bands	10 features for each channel	ELM: 94.72%	-

The emotions-state dataset was used in several studies; it contained 2548 features and classified the emotions into three classes: negative, positive, and nature [7]. For instance, Bird et al. [7] used the dataset to propose a work that included biological inspiration used in all implementation steps rather than being limited to the classification stage. Additionally, they explored deep learning and tuning using long short-term memory (LSTM). Moreover, they have tested AdaBoost using two different models.

The system implements an evolutionary optimization of a multilayer perceptron (MLP) to estimate the network's best hyperparameters. The model extracted 500 features and tested them with several classifiers: deep evolutionary multilayer perceptron (DEvo MLP), LSTM, AdaBoost deep evolutionary multilayer perceptron (AdaBoost DEvo MLP), and AdaBoost LSTM. The accuracy results were 96.11%, 96.86%, 96.23%, and 97.06%, while the training times were 16.66, 65.11, 32.88, and 594.55 seconds, respectively. In study [8], the authors implemented four feature extraction methods: One Rule (OneR), Bayes Network (BN), Info-Gain, and Symmetrical Uncertainty. This study conducted a classification using single and ensemble methods on a dataset to determine the best result. The single models were: OneR, RT, Sequential Minimal Optimization (SMO), Naive Bayes (NB), BN, Logistic Regression (LR), and MLP. In contrast, the ensemble models were RF, Vote, and AdaBoost of random forest (AdaBoost RF). In the single model, MLP with Info-Gain achieved the highest accuracy by 94.89%. Simultaneously, RF with Info-Gain gained the best accuracy in ensemble methods by 97.89%. On the other hand, this experiment [9] converted the EEG signs to 2D and 3D convolutional neural network (CNN) images. In the beginning, the authors used three feature selection methods: Kullback-Leibler Divergence, OneR, and Symmetrical Uncertainty. The best results for each feature selection of 2D CNN were 98.22%, 97.28%, and 97.12%. In contrast, the accuracy of 3D CNN was 97.28%, 96.97%, and 97.12%, respectively.

On the other hand, multiple studies used a popular dataset called SEED [22]; the signals were collected from 15 participants using 15 Chinese film clips. Each participant experimented three times, and they classified the films into three classes: negative, positive, and neutral. Authors in [10] used the SEED dataset to propose a model called (STRNN) which integrated spatial and temporal dependencies with a recurrent neural network (RNN). The proposed approach had

two layers: a multi-direction spatial RNN (SRNN) and a bi-direction temporal RNN (TRNN) layer. These layers captured spatial and temporal information within the sequence signal. The accuracy of this experiment was 89.50%. While Asadur Rahman et al., in [11] used PCA and t-statistics to extract five channels and used them with four classifiers: support vector machine (SVM), artificial neural network (ANN), linear discriminant analysis (LDA), and k-nearest neighbor (KNN). The classification accuracy was 85.85%, 86.57%, 82.50%, and 73.42%, respectively.

DEAP [23] dataset signals were collected from 32 participants; each participant watched 40 music videos for one minute per video. The participants classified the videos based on their levels of valence, arousal, liking, dominance, and familiarity. It is another common dataset used broadly in the state of the art. For instance, Doma and Pirouz in [12] used the DEAP dataset with PCA, and without PCA, with several classifiers: SVM, logistic regression, decision tree, KNN, and naive bayes. The classifiers achieved accuracy ranging from 55% to 75% and an F1 score between 70% and 86%. The better F1-score was 84.73% obtained by SVM with PCA of arousal classification. However, Hemanth in [13] used a Kohonen neural network (KohonenNN) with several modifications to achieve better accuracy. The modified KohonenNN I and II improved the accuracy by 1% to 2%. The best accuracy was achieved using KohonenNN I, which was 87%. In this study [14] suggested a novel model called ECLGCNN. ECLGCNN consists of three layers. The first layer is a graph convolutional neural network (GCNNs) devoted to calculating the relationship between two channels of EEG signals and extracting the graph domain features from differential entropy. The second layer is LSTM, which handles memorizing the changes between two EEG channels. The final layer is the dense layer, which focuses on classifying the emotions. The study conducted two experiments; the first experiment was subject-dependent, and it achieved 90.45% and 90.60% for valence and arousal. The second experiment was subject-independent and achieved lower accuracy, which was 84.81% and 85.27%. In this study [15], authors suggested using an empirical mode decomposition (EMD) to decompose the signals. Then they extracted the features using second-order difference plots (SODP), which are the mean, area, and measure of central tendency. The experiment used an SVM and two-hidden layers of MLP to classify the multi-class emotions. MLP achieved good results in each classification. However, the best accuracy was 100% for high and low of arousal. Another study [16] carried out several feature extraction algorithms: power, entropy, fractal dimension, and statistical. Additionally, they used several classifiers: SVM, KNN, and decision tree. They adopted PCA to select features. The system achieved the overall best accuracies of 78.96%, 77.62%, and 77.60% for arousal, valence, and dominance, respectively, using SVM with statistical features.

Some studies have used multiple emotional datasets. For instance, Cheng et al. [24] used the DEAP and DREAMER [17] datasets. The authors used the deep forest to extract spatial and temporal information from these two datasets. Both experiments achieved good results. The accuracy of the first dataset was 97.69% and 97.53% for valence and arousal, respectively. In contrast, the second dataset's accuracies were 89.03%, 90.41%, and 89.89% for valence, arousal, and dominance, respectively. This study considered the running time; the experiment took 693.4861 seconds in the first experiment and 1307.406 seconds in the second experiment. In contrast, another study [18] used SEED and DEAP datasets to implement their model. The research used a flexible analytic wavelet transform (FAWT) with information potential (IP) to extract the features. The experiment was tested using two classifiers: RF and SVM. SVM was better than RF, where it achieved 59.06% accuracy on DEAP and 83.33% accuracy on the SEED database. Additionally, authors in [19] used the same two datasets; they used three algorithms to extract the EEG signals. The three algorithms are variance, discrete wavelet transform (DWT), and fast Fourier transform (FFT). DEAP and SEED were used to validate the model. The DEAP dataset contains four states: valence, arousal, dominance, and liking. In contrast, the SEED has three states: negative, positive, and neutral. The classifier used spiking neural networks (SNNs), which achieved 78%, 74%, 80%, and 86.27% accuracy, respectively, on the first dataset, while on the second dataset, it achieved 96.67% accuracy.

On the other hand, some studies have used a different dataset. For example, authors in [20] used Indian films to classify emotions into happy, sad, fear, and relax. The study worked in two stages to remove the noise of the dataset. The first stage used empirical mode decomposition (EMD), while the second stage used variational mode decomposition (VMD). A Multi-class least squares SVM (MC-LS-SVM) classifier was used to classify the emotions alongside the morlet wavelet (MW) kernel function. This model's best accuracies were 92.79%, 88.98%, 87.62%, and 93.13% for happy, sad, fear, and relaxed emotions, respectively. Seal et al. in [21] used a dataset that classified the EEG into four categories: sad, fear, happy, and neutral. The experiment used a discrete wavelet transform and an extreme learning machine (ELM). The best accuracy was 94.72% from the FP1-F7 channel in the subband of gamma.

III. METHOD

In this section, we present the proposed model's method, as shown in Fig. 1. We experimented on the emotional state dataset used by some researchers, as mentioned in the literature review section [7]. The database classified a person's emotional state into positive, negative, and neutral. Data were collected from one man and one woman using electrodes via an EEG headband while watching six clips for one minute per clip. Three films stimulated positive emotions, while the other three stimulated negative emotions. Six minutes were recorded for each person in a neutral or normal state. This experiment resulted in a dataset of 2,549 attributes and 2,132 rows. Two algorithms of the feature extraction stage, PCA and FastICA, were applied in this experiment. In addition, three ensemble classifiers were included in the classification stage: RF, XGBoost, and AdaBoost. Moreover, each experiment was tested using 10-cross-fold validation accuracy. Each experiment will be discussed in detail in the following subsections.

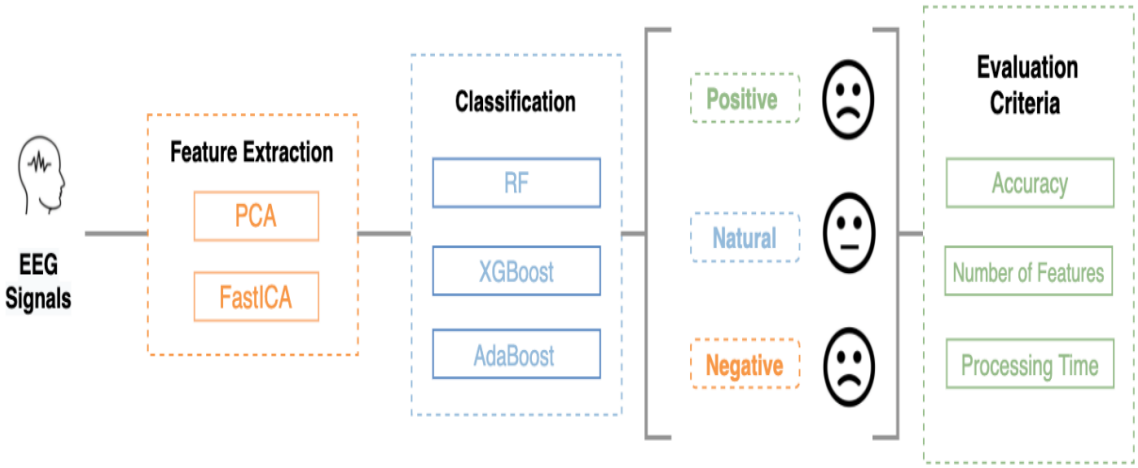


Fig. 1. Research Methodology. The research methodology included three main steps. The first step was the feature extraction step, which contained two methods: PCA and FastICA. The second step was the classification step, which consisted of three classifiers: random forest, XGBoost, and adaptive boost. Finally, the third step was the evaluation step, which was based on three criteria: the accuracy of the model, the number of features, and the processing time.

A. Feature Extraction

This model used two feature extraction algorithms to extract the most important EEG signals of emotions: PCA and FastICA. Several studies used different extraction feature algorithms based on several specifications. In this study, we used PCA for its fast-processing capabilities [25]. Similarly, FastICA is a fast version of ICA and is suitable for large datasets [26]. The two following subsections will discuss the two selected extract feature algorithms.

1) Principal component analysis: The main concept of principal component analysis is dimension reduction by transforming the correlated variables into new uncorrelated components, which maintain a maximum variation of the original components [27]. Based on [28], there are some steps to extracting the features using a PCA, which are presented as follows:

a) Suppose there is a matrix of m × n size.

b) Convert the matrix to an N dimension vector with input data x as $x_0, x_{0,1}, \ldots, x_{m,n}$.

c) Calculate the mean vector as follows:

$$X_{mean} = \left(\frac{1}{N}\right) \sum_{i=1}^{N} X_i \quad (1)$$

d) Calculate the covariance matrix.

e) Compute the eigenvalues and eigenvectors.

f) Select the components using k-eigenvectors of the highest eigenvalues, then construct a w matrix of $d \times k$ dimensions.

g) Construct the principal components using the w matrix to transform the samples into a new subspace.

In this study, PCA was used with each classifier algorithm to obtain a good result. Based on multiple experiments, we have seen that the best selections were 36 components with RF, 33 components with XGBoost, and 28 components with AdaBoost.

2) Fast independent component analysis: Fast independent component analysis is a type of ICA algorithm responsible for separating the unknown mixed signals to obtain useful independent signals using the source signal's independent and non-Gaussian nature [29]. An algorithm of FastICA works faster and is iteratively used at constant points with a simple structure and fast convergence [30].

To implement FastICA, some steps follow, as mentioned in [31]:

a) Remove the mean of x by centralization.

b) Transform x linearly to obtain an uncorrelated vector called z. Thus, the covariance matrix of z will be: $E\{zz^T\} = I$.

c) Construct an operation matrix called w, which satisfies: $\|w\| = 1_{\circ}$.

d) Update the separation matrix w, then iterate it based on the Newton iteration method to obtain w^*.

e) Normalize w^* to be $w = w^*/\|w^*\|_{\circ}$.

f) Judge the coverage of w. If it is good, then the best estimate of the source signal $wx_{\circ}$; if not, go back to step 3.

In FastICA experiments, we selected 33 components with RF and 31 components with AdaBoost and XGBoost.

B. Classifiers

This subsection shows the three classifiers used in this study. We focused on using ensemble classifiers based on their powerful performance. RF and AdaBoost were selected based on their good performance in classifying the EEG signals in some studies [32] [33]. Additionally, we used XGBoost due to its effective label classifications and fast computation [34]. The following subsections will discuss the selected classifiers.

1) Random forest: RF consists of multiple decision trees. Each decision tree makes a classification separately and then provides its result. The final result was identified based on the voting of all the decision trees [35]. The RF algorithm describes the steps according to Evans et al., in [36] as follows:

a) Construct iterative N bootstraps of n size sampled from z population.

b) Grow an RF tree T_b randomly at each node and define m variables from M to permute over each node to know the best split using the Gini entropy index.

c) Use out-of-bag data to validate every tree T_b.

d) Produce several RF trees:

$$\{T_b\} = \frac{B}{1} \quad (2)$$

To predict a new observation x_i, $\hat{C}_b(x)$ be a class prediction of Bth RF tree, then:

$$\hat{C}_{rf}^{B}(x) = majorityvote\{\hat{C}_b(x)\}\frac{1}{B} \quad (3)$$

In this study, we tuned parameters where the best parameters of RF with PCA and Fast ICA are illustrated in Table II.

2) Extreme gradient boosting: Extreme gradient boosting (XGBoost) is a robust algorithm based on a gradient boosting system [37]. It is a tree-boosting system; however, there is a major difference between RF and gradient-boosted machines (GBM): RF trees are built independently. In contrast, the GBM added a new tree to complete the previously built ones [38]. According to Duan et al. [39], the general points of XGBoost algorithms are presented as follows:

- Assume $D = \{(x_i, y_i)\}$ is a dataset of n samples and m features ($D = n, x_i \in R^m, y_i, \in R$).
- The model uses z additive functions to approximate the response of the system, as follows:

$$\hat{y}_i = \phi(x_i) = \sum_{z=1}^{z} f_z(x_i), f_z \in F \quad (4)$$

where, F is the regression trees space, and it is defined as:

$$F = \{f(x) = w_q(x)\}(q: R^m \to T, w \in R^T) \quad (5)$$

where, q stands for the tree structure, T indicates the number of leaf nodes, and the w represents the weights. Besides, f_k is a function that illustrates that w and q are compatible with an independent tree.

Table III illustrates the parameters we used with the XGBoost classifier with PCA and FastICA based on the tuning parameters process.

3) Adaptive boost: Adaptive boost (AdaBoost) is an algorithm that improves learners' accuracy by changing the sample weight distribution [40]. As mentioned in [41] AdaBoost works to reduce the exponential loss greedily. Eq. (6) shows that:

$$FT(X) = \sum_{t=1}^{T} f_t(x) \quad (6)$$

where, ft indicates to a weak learner and x is the object used.

$$FT(X) = \sum_i E\left[F_{t-1}(x_i) + \alpha_t\, h(x_i)\right] \quad (7)$$

where, h(x_i) is the hypothesis made by a weak learner, and αt is the parameter of it to minimize the sum of error in training.

In this experiment, AdaBoost was used two times, first with PCA while the second was with FastICA. The parameters we have used in the two experiments are shown in Table IV.

C. Performance Measures

The proposed model was evaluated in terms of accuracy, number of features, and processing times. The formulation of accuracy is presented as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

where:

- True Positive (TP): means when the actual class was positive, and the model was predicted to be positive.
- True Negative (TN): means when the actual class was not positive, and the model was predicted to be not positive.
- False Positive (FP): means when the actual class was not positive, but the model was predicted to be positive.
- False Negative (FN): means when the actual class was positive, but the model was predicted to be not positive.

TABLE II. PARAMETERS USED BY RF SELECTED BASED ON PARAMETERS TUNING

PARAMETERS	BOOTSTRAP	MAX_DEPTH	MAX_FEATURES	MIN_SAMPLES_LEAF	MIN_SAMPLES_SPLIT	N_ESTIMATORS
PCA	FALSE	100	AUTO	1	2	400
FASTICA	TRUE	60	SQRT	1	5	400

TABLE III. PARAMETERS USED BY XGBOOST SELECTED BASED ON PARAMETERS TUNING

PARAMETERS	BOOTSTRAP	MAX_DEPTH	MAX_FEATURES	MIN_SAMPLES_LEAF	MIN_SAMPLES_SPLIT	N_ESTIMATORS
PCA	TRUE	20	AUTO	1	2	2000
FASTICA	TRUE	20	AUTO	1	2	800

TABLE IV. PARAMETERS USED BY ADABOOST SELECTED BASED ON PARAMETERS TUNING

Parameters	Algorithm	Learning_Rate	n_Estimators
PCA	SAMME	0.01	2000
FastICA	SAMME	0.3	1600

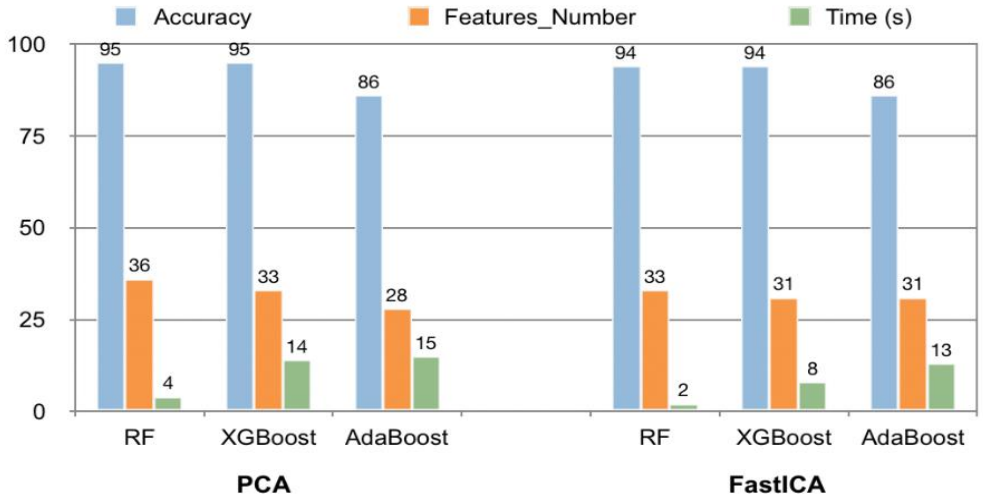


Fig. 2. Proposed model results.

To achieve scientific accuracy, we tested each classifier ten times with PCA and FastICA, avoiding biased results of the classifier accuracy. Then, we calculated the average of these experiments.

Additionally, several measures were calculated to give complete view of the results, namely precision, recall, and F1-score.

Precision was calculated as follows:

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

This means the percentage of the relevant results. Recall was calculated as follows:

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

This indicates the average of precision and recall.

IV. Result and Discussion

One of the goals of this study was to reduce the number of features to fewer than 63, which is the lowest number of features achieved by previous studies that used the same dataset [8]. We succeeded in reducing it by 48%, which equals 33 features using PCA with an accuracy equal to 95% using the XGBoost classifier. On the other hand, FastICA decreased the features by 51%, which is 31 features using the XGBoost classifier with 94% accuracy. According to the time, we noticed that all classifiers' processing times were between 2 seconds and 15 seconds. Notably, all classifiers take a short time to process; thus, we decided to take time as the third factor in the evaluation process. However, RF was the faster algorithm; it took just four and two seconds of processing time with PCA and FastICA.

On the other hand, AdaBoost had the lowest accuracy, achieving 86% in the two experiments. Additionally, AdaBoost was the slowest algorithm in both experiments. The highest accuracy achieved in the experiments was 95% using the XGBoost classifier, with 33 features extracted by PCA. In contrast, the same accuracy was achieved by RF using PCA, but with 36 features. Fig. 2 shows the results of the three classifiers for the three evaluation criteria. The figure shows the performance of the three classifiers with PCA and FastICA. The accuracy of the RF and XGBoost were high and equal in both experiments. However, RF outperformed in the time processing criteria, while XGBoost outperformed in the number of features. AdaBoost was the lowest and slowest classifier in this work. Table V shows the results of each measure of the three emotions in all experiments.

Generally, some existing works achieved better accuracy than this experiment, but with more features, such as [7]. However, we think 95% is not a bad percentage, especially within a short processing time of 14 seconds. Table VI presents a performance comparison of the proposed work and existing works that used the same emotional dataset.

TABLE V. Classification Performance

Feature Extraction	Classifier	Class	Precision	Recall	F1-score
PCA	RF	Negative	0.96	0.98	0.97
		Natural	0.95	0.95	0.95
		Positive	0.93	0.91	0.92
	XGBoost	Negative	0.95	0.98	0.96
		Natural	0.95	0.97	0.96

		Positive	0.95	0.89	0.92
	AdaBoost	Negative	0.81	0.97	0.88
		Natural	0.93	0.91	0.92
		Positive	0.85	0.66	0.74
FastICA	RF	Negative	0.95	0.97	0.96
		Natural	0.94	0.97	0.95
		Positive	0.93	0.88	0.91
	XGBoost	Negative	0.95	0.97	0.96
		Natural	0.94	0.96	0.95
		Positive	0.92	0.88	0.90
	AdaBoost	Negative	0.82	0.94	0.87
		Natural	0.94	0.89	0.91
		Positive	0.76	0.67	0.71

TABLE VI. PERFORMANCE COMPARISON BETWEEN THE MODELS USING THE SAME DATASET IN THIS WORK

Ref	Number of features	Best accuracy	Time (s)
[7]	500	97.06%	595
[8]	63	97.89%	-
[9]	-	98.43%	-
Proposal model: XGBoost with PCA	33	95%	14
Proposal model: XGBoost with FastICA	31	94%	8

V. CONCLUSION

This paper presented an EEG signals emotion prediction model that concerns three factors: accuracy, number of features, and processing time. PCA and FastICA were used to extract the features from the signals. RF, XGBoost, and AdaBoost have been used to classify the signals. The XGBoost results were the best in the two experiments. The best accuracy of this work was 95% using 33 features extracted by PCA; the classification process took 14 seconds. In contrast, RF achieved the same accuracy in the PCA experiment but used 36 features within four seconds. AdaBoost achieved the lowest accuracy and longest time in both experiments. XGBoost was the fastest classifier in both experiments. This model can be adopted in smart devices, such as smartwatches that offer dynamic monitoring of mentally patients to protect them from the probable implications. In future work, the authors will utilize other methods and techniques to obtain better accuracy with fewer features and processing time.

REFERENCES

[1] Y. Bao, Y. Sun, S. Meng, J. Shi, and L. Lu, "2019-nCoV epidemic: address mental health care to empower society," The Lancet, vol. 395, no. 10224. Lancet Publishing Group, pp. e37–e38, 2020. doi: 10.1016/S0140-6736(20)30309-3.

[2] M. Srividya, S. Mohanavalli, and N. Bhalaji, "Behavioral Modeling for Mental Health using Machine Learning Algorithms," J Med Syst, vol. 42, no. 5, pp. 1–12, 2018, doi: 10.1007/s10916-018-0934-5.

[3] S. G. Alonso et al., "Data Mining Algorithms and Techniques in Mental Health: A Systematic Review," Journal of Medical Systems, vol. 42, no. 9. Springer New York LLC, 2018. doi: 10.1007/s10916-018-1018-2.

[4] Q. Gao, C. han Wang, Z. Wang, X. lin Song, E. zeng Dong, and Y. Song, "EEG based emotion recognition using fusion feature extraction method," Multimed Tools Appl, vol. 79, no. 37–38, pp. 27057–27074, 2020, doi: 10.1007/s11042-020-09354-y.

[5] Y. Yao, J. Plested, and T. Gedeon, "Deep Feature Learning and Visualization for EEG Recording Using Autoencoders," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2018, pp. 554–566. doi: 10.1007/978-3-030-04239-4_50.

[6] P. R. Bhise, S. B. Kulkarni, and T. A. Aldhaheri, "Brain Computer Interface based EEG for Emotion Recognition System: A Systematic Review," in 2nd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2020 - Conference Proceedings, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 327–334. doi: 10.1109/ICIMIA48430.2020.9074921.

[7] J. J. Bird, D. R. Faria, L. J. Manso, A. Ekárt, and C. D. Buckingham, "A deep evolutionary approach to bioinspired classifier optimisation for brain-machine interaction," *Complexity*, vol. 2019, 2019, doi: 10.1155/2019/4316548.

[8] J. B. Jordan, E. Aniko, D. B. Christopher, and R. F. Diego, "Mental Emotional Sentiment Classification with an EEG-based Brain-machine Interface," in *The International Conference on Digital Image & Signal Processing (DISP'19)*, 2019, pp. 1–7. [Online]. Available: https://www.researchgate.net/publication/329403546_Mental_Emotional_Sentiment_Classification_with_an_EEG-based_Brain-machine_Interface

[9] J. J. Bird, D. R. Faria, L. J. Manso, P. P. S. Ayrosa, and A. Ekárt, "A study on CNN image classification of EEG signals represented in 2D and 3D," *J Neural Eng*, vol. 18, no. 2, p. 26005, 2021, doi: 10.1088/1741-2552/abda0c.

[10] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial-Temporal Recurrent Neural Network for Emotion Recognition," *IEEE Trans Cybern*, vol. 49, no. 3, pp. 939–947, 2019, doi: 10.1109/TCYB.2017.2788081.

[11] M. Asadur Rahman, M. Foisal Hossain, M. Hossain, and R. Ahmmed, "Employing PCA and t-statistical approach for feature extraction and classification of emotion from multichannel EEG signal," *Egyptian Informatics Journal*, vol. 21, no. 1, pp. 23–35, Mar. 2020, doi: 10.1016/j.eij.2019.10.002.

[12] V. Doma and M. Pirouz, "A comparative analysis of machine learning methods for emotion recognition using EEG and peripheral physiological signals," *J Big Data*, vol. 7, no. 1, p. 18, 2020, doi: 10.1186/s40537-020-00289-7.

[13] D. J. Hemanth, "EEG signal based Modified Kohonen Neural Networks for Classification of Human Mental Emotions," *Journal of Artificial Intelligence and Systems*, vol. 2, no. 1, pp. 1–13, 2020, doi: 10.33969/ais.2020.21001.

[14] Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, "EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM," *Appl Soft Comput*, vol. 100, p. 106954, 2021, doi: 10.1016/j.asoc.2020.106954.

[15] N. Salankar, P. Mishra, and L. Garg, "Emotion recognition from EEG signals using empirical mode decomposition and second-order difference plot," *Biomed Signal Process Control*, vol. 65, p. 102389, 2021, doi: 10.1016/j.bspc.2020.102389.

[16] R. Nawaz, K. H. Cheah, H. Nisar, and V. V. Yap, "Comparison of different feature extraction methods for EEG-based emotion recognition," *Biocybern Biomed Eng*, vol. 40, no. 3, pp. 910–926, 2020, doi: 10.1016/j.bbe.2020.04.005.

[17] S. Katsigiannis and N. Ramzan, "DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals from Wireless Low-cost Off-the-Shelf Devices," *IEEE J Biomed Health Inform*, vol. 22, no. 1, pp. 98–107, 2018, doi: 10.1109/JBHI.2017.2688239.

[18] V. Gupta, M. D. Chopda, and R. B. Pachori, "Cross-Subject Emotion Recognition Using Flexible Analytic Wavelet Transform from EEG Signals," *IEEE Sens J*, vol. 19, no. 6, pp. 2266–2274, 2019, doi: 10.1109/JSEN.2018.2883497.

[19] Y. Luo *et al.*, "EEG-Based Emotion Classification Using Spiking Neural Networks," *IEEE Access*, vol. 8, pp. 46007–46016, 2020, doi: 10.1109/ACCESS.2020.2978163.

[20] S. Taran and V. Bajaj, "Emotion recognition from single-channel EEG signals using a two-stage correlation and instantaneous frequency-based filtering method," *Comput Methods Programs Biomed*, vol. 173, pp. 157–165, 2019, doi: 10.1016/j.cmpb.2019.03.015.

[21] A. Seal et al., "An EEG Database and Its Initial Benchmark Emotion Classification Performance," Comput Math Methods Med, vol. 2020, 2020, doi: 10.1155/2020/8303465.

[22] W. L. Zheng and B. L. Lu, "Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks," IEEE Trans Auton Ment Dev, vol. 7, no. 3, pp. 162–175, 2015, doi: 10.1109/TAMD.2015.2431497.

[23] S. Koelstra et al., "DEAP: A database for emotion analysis; Using physiological signals," IEEE Trans Affect Comput, vol. 3, no. 1, pp. 18–31, 2012, doi: 10.1109/T-AFFC.2011.15.

[24] J. Cheng et al., "Emotion Recognition From Multi-Channel EEG via Deep Forest," IEEE J Biomed Health Inform, vol. 25, no. 2, pp. 453–464, 2020, doi: 10.1109/jbhi.2020.2995767.

[25] S. Shaik and V. Kakulapati, "Curvelet Transform Based EEG Signal Analysis using PCA," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 9, no. 3, pp. 1631–1634, 2020, doi: 10.35940/ijitee.C8479.019320.

[26] Y. Xie and S. Oniga, "A Review of Processing Methods and Classification Algorithm for EEG Signal," Carpathian Journal of Electronic and Computer Engineering, vol. 13, no. 3, pp. 23–29, 2020, doi: 10.2478/cjece-2020-0004.

[27] T. R. Gadekallu et al., "Early Detection of Diabetic Retinopathy Using PCA-Firefly Based Deep Learning Model," Electronics (Basel), vol. 9, no. 2, p. 274, 2020, doi: 10.3390/electronics9020274.

[28] R. M. S. Priya et al., "An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture," Comput Commun, vol. 160, pp. 139–149, 2020, doi: 10.1016/j.comcom.2020.05.048.

[29] A. Moneta and G. Pallante, "Identification of Structural VAR Models via Independent Component Analysis: A Performance Evaluation Study," LEM Papers Series, 2020, [Online]. Available: https://ideas.repec.org/p/ssa/lemwps/2020-24.html

[30] K. Cai, H. Yue, B. Li, W. Chen, and W. Huang, "Combining Chrominance Features and Fast ICA for Noncontact Imaging Photoplethysmography," IEEE Access, vol. 8, pp. 50171–50179, 2020, doi: 10.1109/ACCESS.2020.2979991.

[31] B. Liu, Z. Zhou, Q. Dai, and W. Tong, "FastICA and total variation algorithm for geochemical anomaly extraction," Earth Sci Inform, vol. 13, no. 1, pp. 153–162, 2020, doi: 10.1007/s12145-019-00412-0.

[32] X. Liu, J. Shen, and W. Zhao, "EPILEPTIC EEG IDENTIFICATION BASED on HYBRID FEATURE EXTRACTION," J Mech Med Biol, vol. 20, no. 6, 2020, doi: 10.1142/S0219519420500256.

[33] A. Subasi, S. Jukic, and J. Kevric, "Comparison of EMD, DWT and WPD for the localization of epileptogenic foci using Random Forest classifier," Measurement (Lond), vol. 146, pp. 846–855, 2019, doi: 10.1016/j.measurement.2019.07.026.

[34] S. Bhattacharya et al., "A Novel PCA-Firefly Based XGBoost Classification Model for Intrusion Detection in Networks Using GPU," Electronics (Basel), vol. 9, no. 2, p. 219, 2020, doi: 10.3390/electronics9020219.

[35] T. Zhang, W. Chen, and M. Li, "AR based quadratic feature extraction in the VMD domain for the automated seizure detection of EEG using random forest classifier," Biomed Signal Process Control, vol. 31, pp. 550–559, 2017, doi: 10.1016/j.bspc.2016.10.001.

[36] J. S. Evans, M. A. Murphy, Z. A. Holden, and S. A. Cushman, Modeling species distribution and change using random forest. Springer New York, 2011. doi: 10.1007/978-1-4419-7390-0_8.

[37] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[38] M. Luckner, B. Topolski, and M. Mazurek, "Application of XGboost algorithm in fingerprinting localisation task," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2017, pp. 661–671. doi: 10.1007/978-3-319-59105-6_57.

[39] J. Duan, P. G. Asteris, H. Nguyen, X. N. Bui, and H. Moayedi, "A novel artificial intelligence technique to predict compressive strength of recycled aggregate concrete using ICA-XGBoost model," Eng Comput, pp. 1–18, 2020, doi: 10.1007/s00366-020-01003-0.

[40] Q. Liu, X. Wang, X. Huang, and X. Yin, "Prediction model of rock mass class using classification and regression tree integrated AdaBoost algorithm based on TBM driving data," Tunnelling and Underground Space Technology, vol. 106, p. 103595, 2020, doi: 10.1016/j.tust.2020.103595.

[41] A. R. Javed, Z. Jalil, S. A. Moqurrab, S. Abbas, and X. Liu, "Ensemble Adaboost classifier for accurate and fast detection of botnet attacks in connected vehicles," Transactions on Emerging Telecommunications Technologies, p. e4088, 2020, doi: 10.1002/ett.4088.

Selection of a Trustworthy Technique for Fraud Prevention in the Digital Banking Sector

Bandar Ali M. Al-Rami Al-Ghamdi
Faculty of Computer Studies, Arab Open University, Riyadh 11681, Saudi Arabia

***Abstract*—Digital banking fraud poses a threat to the global economy and fintech applications. Sustainable models are essential to address this issue and minimize its economic impact. Hybrid methods have been developed to assess strategies for preventing digital banking fraud, aiding global stakeholders in making well-informed judgments. However, many of these models concentrate on the numerical features of digital banking ratios while overlooking crucial financial fraud protection qualities. This paper introduces a computational method for discovering and measuring the influence of digital banking fraud prevention strategies on sustainable fraud prevention. This innovative approach combines intuitionistic fuzzy set theory and the analytical network process for decision-making. Initially, an intuitionistic fuzzy expert system prioritizes crucial indices based on the preferences of financial decision-makers. This technique is then compared to alternative decision-making models across multiple variables. Empirical data demonstrate the superiority of the intuitionistic fuzzy-based decision-making system, outperforming other models and facilitating the recognition of financial statement fraud in global banking networks. Consequently, it offers a sustainable fintech solution. The findings of this study are pertinent to fintech scholars and practitioners engaged in the global battle against digital banking fraud.**

***Keywords*—Digital banking fraud; analytical network process; intuitionistic fuzzy sets; fraud prevention and detection**

I. Introduction

Fraudsters in Saudi Arabia have rapidly adapted to modern technological systems, leaving old brick-and-mortar institutions behind [1–2]. Criminal techniques are constantly evolving, posing a threat to the entire payment ecosystem and prompting regulatory attention across all banks, regardless of their level of digital transformation [3–4].

The adoption of online banking in Saudi Arabia is expected to rise by 16.7% from 2024 to 2028 [5], indicating significant growth in KSA's digital banking. After fifteen years of continuous growth, the online banking rate is predicted to peak at 68.16 percent in 2028. It is noteworthy that Saudi Arabia has consistently increased internet banking penetration in recent years, demonstrating a persistent trend of people using digital banking services for financial interaction and management.

The growing digitization of global banking has necessitated an evaluation of hybrid decision-making (DM) methods to combat digital banking fraud [6–8]. Digital banking fraud encompasses phishing, spoofing, identity theft, account fraud, and transaction fraud. Phishing and spoofing, involving unwary recipients receiving misleading emails, phone calls, and texts from seemingly reputable sources, are common occurrences [9–10]. These attempts aim to steal sensitive data or gain access to computer networks, resulting in financial losses and identity theft.

Given the severity of these frauds, Saudi business boards must implement stronger and more effective fraud prevention measures. Fraud detection and mitigation are not only cost-effective but also essential for reputation protection. Saudi Arabian specialists use multiple characteristic decision-making (MCDM) methods to evaluate digital banking fraud prevention technologies [5–7]. The Analytical Network Process (ANP) serves as a suitable model for comparing characteristics to solutions. However, due to decision-makers' subjectivity and uncertainties in input data, a more realistic strategy that effectively addresses these uncertainties is needed.

The intuitionistic fuzzy ANP emerges as a better and more realistic method for evaluating alternative solutions to Saudi Arabia's digital banking fraud problem. It assists decision-makers in choosing actions that align with goals by considering user expertise and historical perspectives on the issue. By incorporating the views of Saudi Arabian subject-matter experts, the intuitionistic fuzzy-ANP model addresses local fintech differences effectively.

This research enhances the evaluation of fraud prevention models in the financial accounts of Saudi Arabian fintech applications using the intuitive fuzzy-ANP methodology. The study has two objectives: first, to understand the drivers of Saudi fintech apps and digital banking fraud, and second, to assess how well the intuitionistic fuzzy-ANP utilizes these characteristics to prevent wrongdoing. The study recommends a systematic and integrated approach to investigation and prevention to identify abnormalities and conduct comprehensive fraud assessments [8–12]. These early investigations lay the groundwork for further analyses and in-depth inquiries to resolve the issue.

This article analyzes past trends and statistics to provide readers with a brief introduction to fraud analysis and prevention in fintech apps. The next section is a literature overview of pertinent studies by earlier researchers. The final section discusses features used to evaluate digital banking fraud models and suggests a hierarchical architecture for Saudi Arabian fraud detection and prevention. The study employs intuitionistic fuzzy-ANP to numerically deconstruct the hierarchical problem. The discussion will include a comparative study after presenting the findings. The paper concludes with a summary of the complex discussion and outlines study limitations.

II. PREVENTING DIGITAL BANKING FRAUDS

Saudi Arabia needs a comprehensive digital banking fraud prevention strategy to guide managers and key departments in recognizing, assessing, and responding to possible fraud [13-15]. This well-developed strategy details each organization's function and its duties. Its success depends on senior management's active participation in its formulation. Effective communication between decision-makers, managers, law enforcement agencies, customers, and external organizations helps develop robust risk analysis, prediction, and DM procedures and fosters seamless collaboration.

Comprehensive training for the entire team is necessary to accurately apply the policy and handle situations in accordance with the fraud prevention strategy [16-17]. Before preventing fraud, this training is required. To ensure efficacy, the fraud prevention plan should be thoroughly examined, rehearsed, and checked for flaws. This rigorous preparation is necessary since the fraud prevention plan executes activities and ensures fraud detection and prevention. The organization's nature, size, and operating environment have all played a significant role in shaping the Saudi strategy.

Before establishing the best decision support tools, Saudi Arabia needs a well-defined network to manage its many fraud prevention techniques. Fraud in digital banking applications is diverse; thus, decision support systems should be tailored to the individual difficulties [5-6]. The following paragraph examines the fundamental principles of a proactive preventative strategy and its components. The following paragraph elaborates on this strong discussion and analysis.

A. Fraud Prevention Planning

Upper management's active involvement across all business units and their unwavering support underpin the plan's concepts in Saudi Arabia [15]. It has four main steps, as explained below. After refining with Saudi-specific examples, the process was reduced to four parts. The organization's management, fraud risk assessment, views of fraud detection, and a complete fraud prevention policy for stakeholders are covered in these phases. Each phase is interdependent, and sub-processes are carefully structured to match. The hierarchical design in Fig. 1 shows how to build a fraud prevention plan model. Organizational management subprocesses include management attitude, integrity, policy commitment, and strict enforcement. Below is a full description of each subprocess:

1) Organization's management: Any fraud prevention plan begins with a thorough management structure review. Senior management must be involved in organisational units while developing a fraud prevention strategy. This involvement should be regular to be effective. The top of the firm oversees all plan execution processes as a protective cover. This level includes management's thinking, unshakable dedication to integrity policies, and rigorous implementation. Fig. 1 shows the complex interaction of these mechanisms [4–5].

2) Fraud risk assessment: Any comprehensive fraud protection plan's second crucial step is a thorough analysis of digital banking fraud risk. This phase assesses the like-lihood of illegal activity. It involves creating specialised teams, mitigating risks, and monitoring and controlling the strategy. This step includes the creation of the risk as-sessment team, the creation of a rigorous risk assessment methodology, and the im-plementation of compliance monitoring measures [9–12]. Fig. 1 shows the fraud risk assessment procedure in steps.

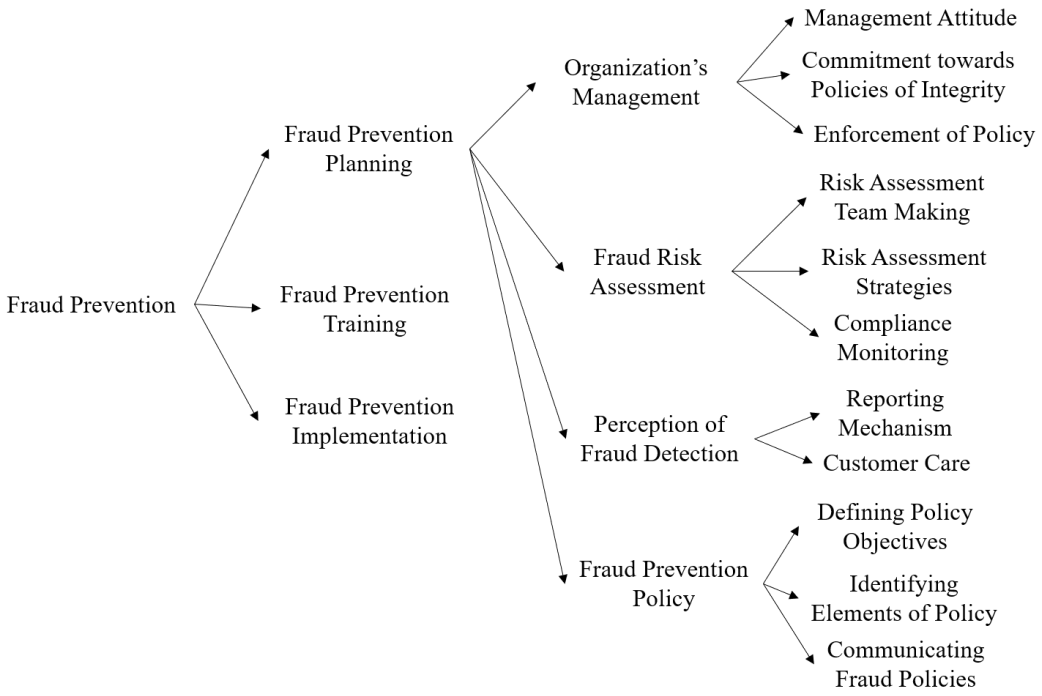


Fig. 1. Basic characteristics.

3) Perception of fraud detection: This phase includes all fraud detection efforts [16–17]. Customer service and fraud reporting are key to this level. Fig. 1 shows that fraud detection and its subsidiary procedures are the fundamental processes in this situation.

4) Fraud prevention policy: The methodical process an organisation follows to create and implement a digital banking fraud prevention policy is called the "fraud prevention policy." This process involves identifying essential fraud policy elements, developing effective communication mechanisms, and rigorously documenting the policy [17]. Fig. 1 shows the three essential subprocesses of creating policy objectives, identifying critical fraud prevention policy elements, and disseminating the fraud policy. It also shows the first step in the fraud prevention policy process.

B. Fraud Prevention Training

Saudi Arabia simulates the fraud prevention approach in this planning phase. At this level, fraud prevention plans and methods are tested and practiced. This involves developing testing processes, creating test scenarios, and carefully assessing the fraud protection plan's performance. Management must also train their Saudi Arabian staff and agents to effectively implement the fraud prevention plan [17].

C. Fraud Prevention Implementation

Saudi Arabian enterprises need comprehensive fraud detection and prevention solutions due to the rising frequency of financial crimes, cyberattacks, and digital fraud. The main goal is to reduce the damage or costs from these illegal operations [17]. Financial crimes, hacking, and digital fraud are on the rise in Saudi Arabia. Organizations can avoid the high costs and consequences of unchecked fraud by implementing efficient fraud detection and prevention methods. This strategic approach protects Saudi Arabian organizations from unbridled fraud [17].

Fraud prevention must be integrated into a comprehensive plan to identify and prevent fraudulent transactions or banking activities in Saudi Arabia, preventing financial and reputational harm to clients and financial institutions. This approach helps identify fraudulent transactions and banking activity and prevent them from harming the Saudi financial sector.

III. Methodology

The proposed methodology for selecting dependable visual analytics tools tailored for medical data analysis in the context of Saudi Arabia is a multifaceted approach encompassing three key phases [5-6]. The methodology commences with the meticulous establishment of a comprehensive network encompassing all variables potentially impacting the problem at hand. This entails a rigorous examination of the entire chain under consideration, with a keen focus on identifying any potential vulnerabilities. Subsequently, the outcomes are categorized based on shared attributes, following the delineation of intersections. This procedural step warrants periodic reevaluation, especially when significant alterations within the chain transpire.

In the subsequent phase, the methodology involves the assignment of weights to the criteria, a crucial step in the process. Intuitionistic fuzzy ANP is the chosen method for this purpose, incorporating the valuable input of domain experts. Within this framework, the criteria are diligently assessed under specific key categories.

The third phase of the methodology encompasses the scrutiny of results derived from multiple criteria evaluations. This comprehensive analysis aids in the identification of key patterns, trends, and insights essential for informed DM.

Finally, the fourth step of the methodology is pivotal in determining the organization’s readiness to employ raw materials. To facilitate this DM process, the intuitionistic fuzzy TOPSIS technique is employed. This step serves as the culmination of the methodology, shaping the organization’s course of action.

In adapting this methodology to the unique landscape of Saudi Arabia, it is paramount to consider the distinct contextual factors, healthcare requirements, and societal aspects that underpin the nation's healthcare sector. The utilization of a reliable knowledge-based fuzzy expert system within this framework holds substantial promise in advancing medical data visualizations, enhancing patient care, and contributing to the evolving healthcare landscape of Saudi Arabia.

A. Intuitionistic Fuzzy ANP

Intuitionistic fuzzy ANP is an extension of ANP that seamlessly incorporates intuitionistic fuzzy set theory into its framework [5-7]. In intuitionistic fuzzy ANP, intuitionistic fuzzy scales are a key part of showing how different parts of different criteria compare in terms of how intense they are. Consequently, this approach enables the creation of an intuitionistic fuzzy decision matrix, with intuitionistic fuzzy quantities representing the ultimate scores assigned to the alternatives [8-10]. By applying specific algebraic operators to these intuitionistic fuzzy integers, the optimal solution is derived, effectively encapsulating the entirety of weight vectors and the judgment matrix.

To gauge the relative importance of one criterion compared to another, intuitionistic fuzzy ANP employs intuitionistic fuzzy numbers in the formulation of an intuitionistic fuzzy judgment matrix for each measure [12-13]. These evaluation vectors, in tandem with the intuitionistic fuzzy pairwise comparison matrix, facilitate the allocation of relative significance to each criterion. The representation of linguistic expressions as intuitionistic fuzzy integers is elucidated in Table I, while Table II delineates the computation of the consistency ratio (CR) through the utilization of the random index (RI). Furthermore, the intuitionistic fuzzy membership function for linguistic terms in both criteria and sub-criteria is visually depicted in Fig. 2.

It is crucial to acknowledge that the final weightings result from a meticulous process of organizing, modifying, and reviewing linguistic term assignments, frequently with the help of expert consensus. By using intuitionistic fuzzy ANP in the context of Saudi Arabia, this method is likely to improve the evaluation of a reliable knowledge-based intuitionistic fuzzy

expert system for medical data visualizations that fits the unique needs of the Saudi healthcare system.

TABLE I. SCALE

Numeric Value	Verbal Value	Intuitionistic based Triangular Fuzzy Number (TFN)
1	Equally significant	1, 1, 3
3	Less significant	1, 3, 5
5	Very significant	3, 5, 7
7	Incredibly Essential	5, 7, 9
9	Extremely Essential	7, 9, 11

TABLE II. SCALE OF RANDOM INDEX

Size (n)	1	2	3	4	5	6	7	8
RI	0	0	0.52	0.89	1.11	1.25	1.35	1.40

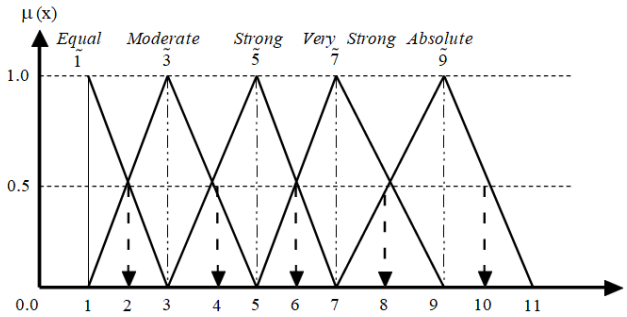


Fig. 2. Function of fuzzy membership.

B. Intuitionistic Fuzzy TOPSIS

Applying intuitionistic fuzzy ANP and intuitionistic fuzzy TOPSIS entails the following three procedures: One must first discover what elements are having an impact on the DM procedure, then perform the calculations for the intuitionistic fuzzy ANP, and finally utilize the intuitionistic fuzzy TOPSIS and Table III to rank the options.

In Saudi Arabia, using intuitionistic fuzzy ANP and intuitionistic fuzzy TOPSIS requires a structured approach that includes three integral procedures. These procedures are delineated as follows: initially, the identification of influential factors affecting the DM process; subsequently, the execution of calculations employing intuitionistic fuzzy ANP; and ultimately, the utilization of intuitionistic fuzzy TOPSIS, along with the insights presented in Table III, to ascertain the ranking of available options.

TABLE III. RATING SCALE

Verbal Values	Equivalent TFNs
Very Poor (VP)	1, 1, 3
Poor (P)	1, 3, 5
Medium (M)	3, 5, 7
Good (G)	5, 7, 9
Very Good (VG)	7, 9, 11

In the initial phase, the identification of potential attributes for the foundational system is conducted. Criteria that hold significance in impacting the DM process are identified, and a hierarchical structure is established to facilitate informed DM. This pivotal phase culminates with the approval of the DM chain of command by the team of decision-makers. An inherent strength of the proposed methodology lies in its adeptness at addressing ambiguity, particularly in terms of criteria and resources. Several key strategies are employed within this method to mitigate or eliminate uncertainty:

Alignment of variables across different facets of the model to gauge the problem-solving capabilities of field experts effectively.

Replacement of numerical data with descriptive terminology through the application of distinct intuitionistic fuzzy membership functions, which are aptly suited for resolving the discussed issues.

Consideration of multiple weighted sources relevant to the subject matter, thereby enhancing the reliability of medical data.

Employing a two-module structure that accommodates varying levels of granularity and uncertainty inherent in medical data sources, encompassing precise qualitative and intangible quantitative medical data, as well as tangible quantitative medical data and data derived from field surveys.

The intuitionistic fuzzy TOPSIS technique demonstrates remarkable efficacy in addressing real-world application challenges within an intuitionistic fuzzy environment. Rooted in traditional multi-attribute-based DM systems, TOPSIS [17] operates on the principle that the alternative selected should be the one farthest from both the positive ideal solution (PIS) and negative ideal solution (NIS). Additionally, TOPSIS features a flexible and user-friendly calculation methodology, enabling the concurrent consideration of multiple criteria with differing units [17]. The procedural steps of the intuitionistic fuzzy ANP-TOPSIS assessment are elaborated below.

Step 1: Determine the relative importance of each criterion for evaluation. In this study, the intuitionistic fuzzy ANP is used to determine preference weights.

Step 2: Build the intuitionistic fuzzy decision/performance matrix and use the criteria to select the right linguistic variables for the different options as shown in Eq. (1) and Eq. (2).

$$\widetilde{DA} = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & \dots & \dots & f_{1n} \\ f_{21} & f_{22} & \dots & \dots & f_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ f_{m1} & f_{m2} & & & f_{mn} \end{bmatrix} \quad (1)$$

i=1,2,3,….m and j=1,2,…n

$$x_{ij} = \frac{1}{K}(x_{ij}^{-1} \oplus \dots \oplus x_{ij}^{-k} \oplus \dots x_{ij}^{-K}) \quad (2)$$

Where x_{ij}^{-k} is performance rating of attribute $A_i$with respect to C_j evaluated by kth expert and $\tilde{x}_{ij}^{k} = (l_{ij}^{k}, m_{ij}^{k}, u_{ij}^{k})$.

Step 3: Next step is constructing the intuitionistic fuzzy decision matrix. The unprocessed medical data are then

normalized by employing a linear scale conversion in order to bring the scales of the different criteria into a comparable format. The intuitionistic fuzzy decision matrix for the attributes $(\widetilde{DA})$ is built as follows Eq. (3) to Eq. (5).

$$(\widetilde{DA}) = [\tilde{d}_{ij}]_{m\times n} \quad (3)$$

Where i=1,2,3,….m and j=1,2,…n

$$\tilde{d}_{ij} = \left(\frac{x_{ij}}{z_j^*}, \frac{y_{ij}}{z_j^*}, \frac{z_{ij}}{z_j^*}\right) \text{ and } z_j^* = \max_i z_{ij} \quad (4)$$

$$\tilde{d}_{ij} = \left(\frac{x_j^-}{z_{ij}}, \frac{x_j^-}{y_{ij}}, \frac{x_j^-}{x_{ij}}\right) \text{ and } z_j^* = \max_i z_{ij} \quad (5)$$

Step 4: Build the weighted and normalized matrix.

Multiplying the weights (wj) of calculating criteria by the normalized intuitionistic fuzzy decision matrix is how one arrives at the weighted normalized matrix (wj) for criteria.$\tilde{d}_{ij}$ see Eq. (6).

$$\tilde{V} = [v_{ij}]_{m\times n} \quad \text{i=1,2,….n ; j=1,2,3……m where } v_{ij} = \tilde{d}_{ij}(\cdot)W_j \quad (6)$$

Note that $\tilde{v}_{ij}$ is a TFN represented by $(\tilde{x}_{ijk}, \tilde{y}_{ijk}, \tilde{z}_{ijk})$

Step 5: Following is the calculation that is used to determine the intuitionistic fuzzy PIS and intuitionistic fuzzy NIS of the attributes as shown in Eq. (7) and Eq. (8):

$$F^* = (\tilde{v}_1^*, \tilde{v}_2^* \dots \dots . \tilde{v}_n^*) \text{ where } \tilde{v}_j^* = (\tilde{z}_j^*, \tilde{z}_j^*, \tilde{z}_j^*) \ \tilde{z}_j^* = \max_i\{\tilde{z}_{ij}\} \quad (7)$$

$$F^- = (\tilde{v}_1^-, \tilde{v}_2^- \dots \dots . \tilde{v}_n^-) \text{ where } \tilde{v}_j^- = (\tilde{x}_j^-, \tilde{x}_j^-, \tilde{x}_j^-) \ \tilde{x}_j^- = \min_i\{\tilde{x}_{ij}\} \quad (8)$$

i=1,2,3…..m; j=1,2,….n

Step 6: The subsequent step is to determine the distance between individual attributes.

The distance (d_i^+, d_i^-) of individual weighted alternative i=1,2,3……m from the intuitionistic fuzzy PIS and intuitionistic fuzzy NIS is computed as per the following in Eq. (9) to Eq. (11):

$$d_i^+ = \sum_{j=1}^n dv(\tilde{v}_{ij}, \tilde{v}_j^+) \quad i = 1, \dots \dots, m \quad (9)$$

$$d_i^- = \sum_{j=1}^n dv(\tilde{v}_{ij}, \tilde{v}_j^-) \quad i = 1, \dots \dots, m \quad (10)$$

$$d(\tilde{A}, \tilde{B}) = \sqrt{\frac{1}{3}((x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2)} \quad (11)$$

Step 7: This step is computation of the closeness coefficient $ClCo_i$ of individual alternative

The closeness coefficient $ClCo_i$ symbolizes the distances to the intuitionistic fuzzy PIS (d^+) and the intuitionistic fuzzy NIS (d^-) concurrently. The closeness coefficient of separate attribute is measured as in Eq. (12):

$$ClCo_i = \frac{d_i^-}{(d_i^+ + d_i^-)} \quad (12)$$

Step 8: Next step is to rank the attributes

In step 8, the values of the maximum closeness coefficient are used to rank or choose the different attributes in decreasing order.

IV. Results

The author evaluates hybrid DM methods for combating digital banking fraud in Saudi Arabia in this part. The author implements intuitionistic fuzzy ANP and shows how it solves selection through hybrid multi-criteria decision-making. The author created and constructed a network of digital fraud prevention model variables and conducted a case study using Saudi-based digital banking applications to evaluate the intuitionistic fuzzy-ANP technique. In addition, the ANP method establishes characteristic weights in phases. 147 Saudi decision-makers and experts evaluated the qualities and options. Three important features, four level 1 characteristics, and eleven level 2 characteristics are compiled from the literature review. After reviewing Saudi Arabian literature and industry standards, the author chose decision network of Fig. 1.

Planning, training, and implementation of fraud prevention are the primary evaluation characteristics. Saudi Arabia's organizational management fraud prevention plan includes fraud risk assessment, perceptions of fraud detection, and fraud prevention policies. Management sub-characteristics include attitude, integrity, and policy implementation. Sub-characteristics of fraud risk assessment include risk assessment teams, methodologies, and compliance monitoring. Customer service and reporting models are fraud detection sub-characteristics. A complete fraud prevention strategy should include planning, training, and implementation. Saudi Arabia uses a matrix to weigh each criterion and alternative.

Consequently, the primary objective of this study is to conduct a case study involving five alternatives (digital fraud prevention models), exploring the attributes that determine their suitability for selecting trustworthy fraud prevention models for the digital banking sector. The chosen approach encompasses a comprehensive set of potential identifiers, each accompanied by rating evaluations. This approach was determined before the commencement of the investigation. The selection of these five digital fraud prevention models as alternatives for comparative trustworthiness evaluation was based on consensus among domain owners and subject matter experts, ensuring a robust evaluation process. To enhance the productivity and reliability of this study, we have conducted an ANP-TOPSIS analysis within an intuitionistic fuzzy framework. This evaluation is centred on assessing the ideality of a hybrid medical expert system for selecting a trustworthy fraud prevention model for the digital banking sector, following the equations detailed in the methodology section, ranging from Eq. (1) to Eq. (12).

After converting linguistic expressions into numerical values Steps 1 to 4 and Eq. (1) to Eq. (6) these values were refined within the intuitionistic fuzzy framework (see Step 5) into clear numerical values. Subsequently, numerical computations were performed to construct a pairwise comparison matrix, with the results summarized in Table IV, as detailed below. The algorithm progressed by introducing intuitionistic fuzzy integer values and then transitioning into crisp numerical values within an intuitionistic fuzzy framework

to present the final results in Table IV. Following that, numerical calculations were executed to generate a pairwise comparison matrix, and the summarized results, which are displayed in Table IV and presented below, are elaborated in the subsequent paragraph.

The method was changed to include intuitionistic fuzzy wrappers as in Eq. (1) to Eq. (5), triangular number estimation, and the degree of possibility in order to get the final results shown in Table V. Ultimately, the experts established the pairwise comparison matrix using Eq. (7) to Eq. (8). Table V showcases the defuzzified values of the group's characteristics, computed using Eq. (9), and the Table V and Fig. 3 were constructed accordingly. Table VI displays the normalized weights of the group's characteristics after calculating local priority vectors, weighted super matrix, and super matrix formation. The comprehensive findings of this investigation are concisely summarized below for reference. After that, numerical calculations were done to figure out the absolute weight vector for row values and figure out which traits were the most important, as shown in Eq. (10) to Eq. (12). The intuitionistic fuzzy data from the judgment matrices were put together to make a pairwise contribution matrix. Table VII and Fig. 4 shows the results of the network as a whole.

TABLE IV. WEIGHTS OF THE CHARACTERISTICS AND THE SUB CHARACTERISTICS

	Characteristics Weight	Sub Characteristics at Level 2	Characteristics Weight	Sub Characteristics at Level 3	Characteristics Weight	Final Weight (CW*SCW3)
Fraud Prevention Planning	0.7200	Organization's Management	0.1601	Management Attitude	0.3312	0.0382
				Commitment towards Policies of Integrity	0.3453	0.0398
				Enforcement of Policy	0.3235	0.0373
		Fraud Risk Assessment	0.3023	Risk Assessment Team Making	0.3111	0.0677
				Risk Assessment Strategies	0.3224	0.0702
				Compliance Monitoring	0.3665	0.0798
		Perception of Fraud Detection	0.2111	Reporting Mechanism	0.3424	0.0520
				Customer Care	0.6576	0.0999
		Fraud Prevention Policy	0.3265	Defining Policy Objectives	0.3222	0.0757
				Identifying Elements of Policy	0.3546	0.0834
				Communicating Fraud Policies	0.3232	0.0760
Fraud Prevention Training	0.1200	-	-	-	-	0.1200
Fraud Prevention Implementation	0.1600	-	-	-	-	0.1600

TABLE V. THE PRIORITY OF THE CHARACTERISTICS

Final Characteristics	Final Weight	Percentage	Overall Priority
Management Attitude	0.0382	3.82 %	12
Commitment towards Policies of Integrity	0.0398	3.98 %	11
Enforcement of Policy	0.0373	3.73 %	13
Risk Assessment Team Making	0.0677	6.77 %	9
Risk Assessment Strategies	0.0702	7.02 %	8
Compliance Monitoring	0.0798	7.98 %	5
Reporting Mechanism	0.0520	5.20 %	10
Customer Care	0.0999	9.99 %	3
Defining Policy Objectives	0.0757	7.57 %	7
Identifying Elements of Policy	0.0834	8.34 %	4
Communicating Fraud Policies	0.0760	7.60 %	6
Fraud Prevention Training	0.1200	12.00 %	2
Fraud Prevention Implementation	0.1600	16.00 %	1

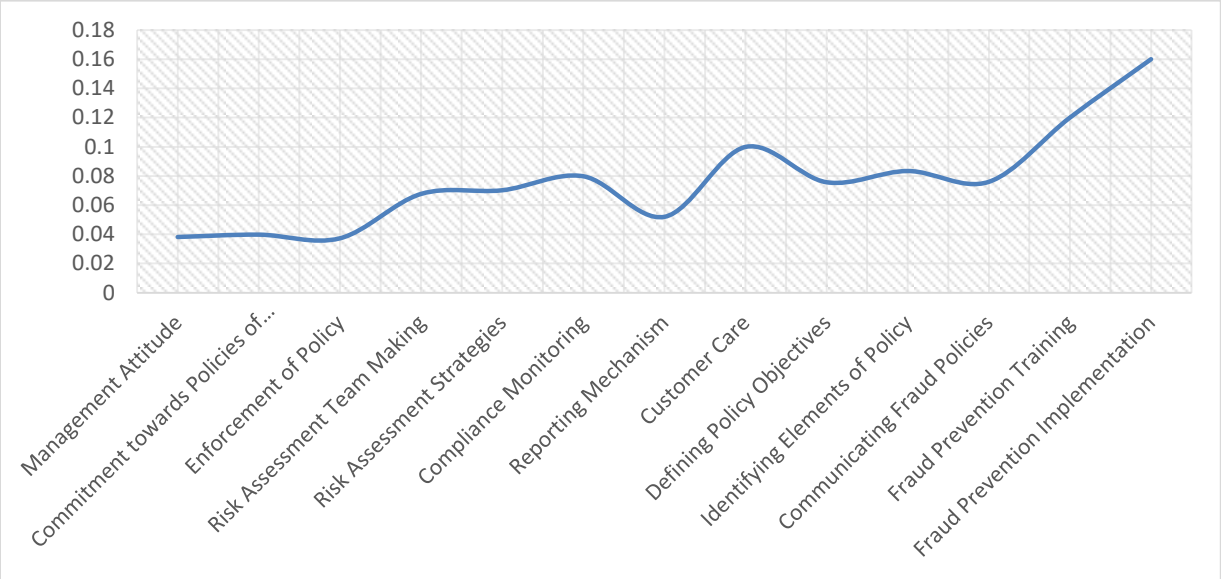


Fig. 3. Graphical representation of characteristics.

Subsequently, the following section of this study will provide a practical assessment of the findings, focusing on the evaluation of a trustworthy fraud prevention model for the digital banking sector. To accomplish this, an ANP approach was applied under conditions of fuzziness to derive the composite weights of features. Subsequently, the intuitionistic fuzzy TOPSIS method was employed to determine the overall ranking of competing alternatives, utilizing the feature weights obtained earlier. It is important to note that these evaluations are particularly relevant in the context of Saudi Arabia.

After completing several intermediary steps, we obtained the normalized intuitionistic fuzzy decision matrix for five fraud prevention models: Riskified [1], Nudata Security [2], GBG Services [18], Feedzai [19], and Featurespace [20]. The results of our analysis are encapsulated within this matrix. To calculate the normalized performance values of the intuitionistic fuzzy decision matrix, we utilized Eq. (8) to Eq. (9). Table VI presents the definitive findings, computed by applying Eq. (10) to Eq. (11) to establish the positive and negative idealness of each alternative concerning each characteristic. These equations were combined to ascertain the ideality of each alternative. The presentation of these results follows the chronological order in which they were obtained.

Furthermore, Eq. (12) were employed to calculate the relative closeness score for each choice, which was then used to determine the degree of satisfaction [21]. The results of this computation can also be found in Table VII and Fig. 4. This additional calculation serves as a valuable post-analysis assessment to provide a comprehensive view of the findings. These evaluations are essential for informed DM in the Saudi Arabian context.

TABLE VI. COMBINED RATINGS WITH RESPECT TO FINAL CHARACTERISTICS

Final Characteristics	Riskified	Nudata Security	GBG Services	Feedzai	Featurespace
Management Attitude	(0.60,0.30,0.40)	(0.65,0.35,0.35)	(0.50,0.30,0.50)	(0.60,0.40,0.40)	(0.55,0.30,0.45)
Commitment towards Policies of Integrity	(0.50,0.30,0.50)	(0.60,0.40,0.40)	(0.70,0.20,0.30)	(0.75,0.20,0.25)	(0.60,0.30,0.40)
Enforcement of Policy	(0.70,0.20,0.30)	(0.75,0.20,0.25)	(0.55,0.30,0.45)	(0.85,0.30,0.20)	(0.75,0.20.0.30)
Risk Assessment Team Making	(0.55,0.30,0.45)	(0.85,0.30,0.20)	(0.60,0.40,0.40)	(0.65,0.30,0.35)	(0.60,0.25,0.40)
Risk Assessment Strategies	(0.60,0.40,0.40)	(0.65,0.30,0.35)	(0.65,0.35,0.35)	(0.60,0.20,0.40)	(0.50,0.40,0.50)
Compliance Monitoring	(0.65,0.35,0.35)	(0.50,0.30,0.50)	(0.60,0.40,0.40)	(0.55,0.30,0.45)	(0.60,0.45,0.40)
Reporting Mechanism	(0.50,0.30,0.50)	(0.60,0.40,0.40)	(0.55,0.30,0.45)	(0.60,0.30,0.40)	(0.60,0.40,0.40)
Customer Care	(0.70,0.20,0.30)	(0.75,0.20,0.25)	(0.60,0.30,0.40)	(0.75,0.20.0.30)	(0.65,0.30,0.40)
Defining Policy Objectives	(0.55,0.30,0.45)	(0.85,0.30,0.20)	(0.50,0.30,0.50)	(0.60,0.40,0.40)	(0.55,0.30,0.45)
Identifying Elements of Policy	(0.50,0.30,0.50)	(0.60,0.40,0.40)	(0.55,0.30,0.45)	(0.75,0.20,0.25)	(0.60,0.30,0.40)
Communicating Fraud Policies	(0.70,0.20,0.30)	(0.75,0.20,0.25)	(0.60,0.30,0.40)	(0.85,0.30,0.20)	(0.75,0.20.0.30)
Fraud Prevention Training	(0.55,0.30,0.45)	(0.85,0.30,0.20)	(0.75,0.20.0.30)	(0.65,0.30,0.35)	(0.60,0.25,0.40)
Fraud Prevention Implementation	(0.60,0.40,0.40)	(0.65,0.30,0.35)	(0.60,0.25,0.40)	(0.60,0.20,0.40)	(0.50,0.40,0.50)

TABLE VII. THE OVERALL SCORE OF DIFFERENT ALTERNATIVES

Alternatives	Ideal Best	Ideal Worst	Ideal Best+ Ideal Worst	Degree of Closeness	Ranking
Riskified	0.0600	0.1200	0.1900	0.67895	1
Nudata Security	0.0700	0.1300	0.2100	0.64587	3
GBG Services	0.0800	0.1600	0.2500	0.66985	2
Feedzai	0.1000	0.1100	0.2200	0.61235	5
Featurespace	0.0800	0.1400	0.2000	0.63528	4

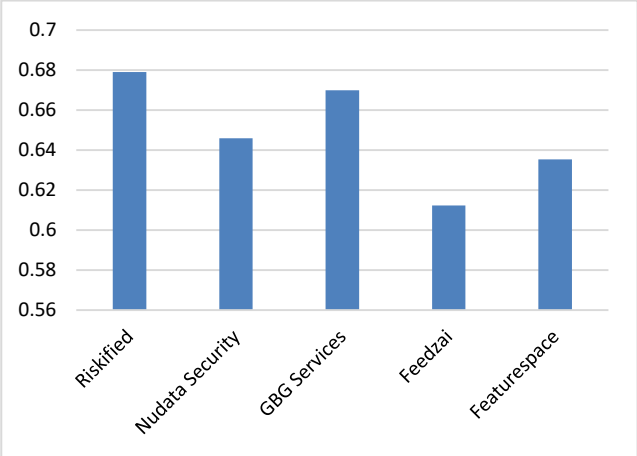


Fig. 4. Impact of alternatives.

The investigation was conducted on five distinct alternatives within the digital banking applications, revealing that categorization is a preferable and successful approach to addressing issues related to fraud prevention in the context of Saudi Arabia. The evaluation was based on the selected characteristics chosen to serve as the foundation for assessment in the Saudi Arabian digital banking landscape.

The relative weights assigned to each of the numerous factors used for selection have a significant impact on the presentation order of various choices. Careful adjustments to the proportional weights of the selection characteristics are essential in the Saudi context to avoid potential ranking changes as a direct consequence of these adjustments. The authors followed a sensitivity analysis approach outlined in [5] to assess the level of confidence in the findings, allowing them to validate their results within the Saudi digital banking environment. By progressively adding a 5% penalty to the weights of each selection criterion one by one, the authors evaluated the sensitivity of the final outcomes to performance variations in Saudi Arabia. These steps enable an exploration of the sensitivity of the results, specifically in the Saudi context. The conclusions derived from the sensitivity analysis are graphically represented in Table VIII and Fig. 5, included for the sake of clarity and convenience for Saudi stakeholders. The results unequivocally demonstrate the ongoing consistency of their practices within the Saudi digital banking sector.

TABLE VIII. SENSITIVITY ANALYSIS

Final Characteristics	Riskified	Nudata Security	GBG Services	Feedzai	Featurespace
Original Outcomes	0.67895	0.64587	0.66985	0.61235	0.63528
Exp-1	0.67789	0.64545	0.66956	0.61985	0.63545
Exp-2	0.67524	0.64756	0.66445	0.61478	0.63458
Exp-3	0.67785	0.64123	0.66658	0.61869	0.64568
Exp-4	0.67265	0.64444	0.66236	0.61856	0.60562
Exp-5	0.67456	0.64445	0.66789	0.61478	0.63789
Exp-6	0.67256	0.64666	0.66985	0.61689	0.63852
Exp-7	0.67235	0.64555	0.66563	0.61658	0.63548
Exp-8	0.67474	0.64223	0.66236	0.60256	0.63658
Exp-9	0.67111	0.64289	0.66745	0.61567	0.63653
Exp-10	0.67874	0.64789	0.66569	0.61856	0.63789
Exp-11	0.67744	0.64987	0.66265	0.61789	0.63356
Exp-12	0.67000	0.64256	0.66485	0.61456	0.63897
Exp-13	0.67446	0.64998	0.67856	0.62568	0.63564

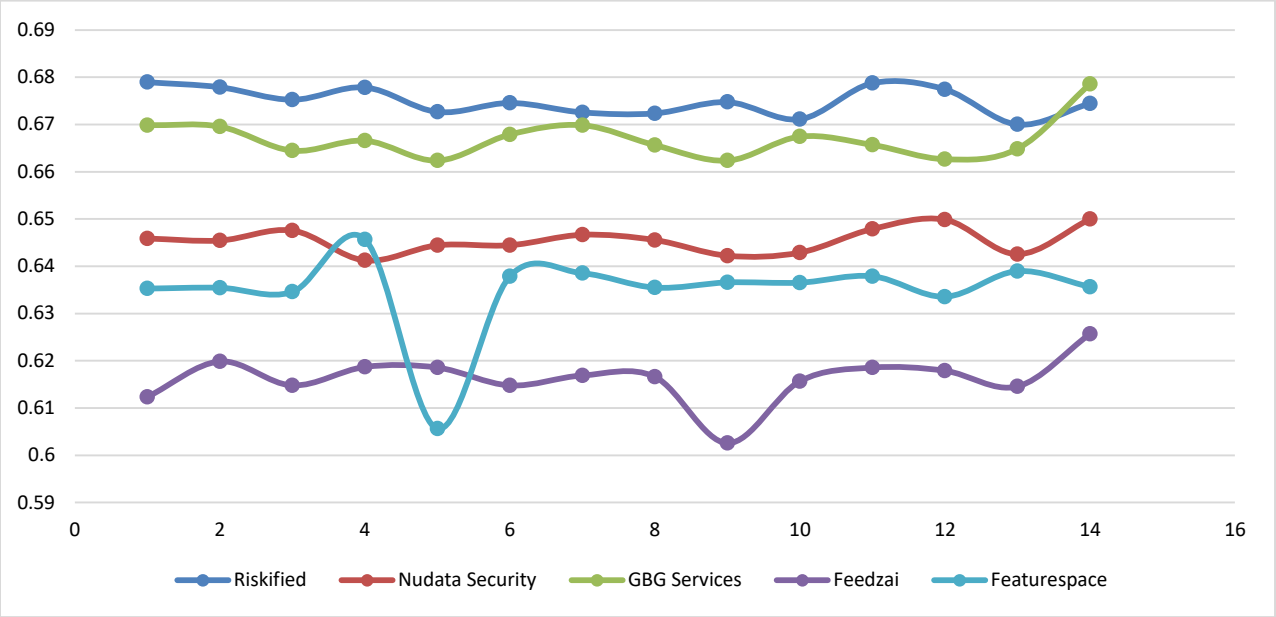


Fig. 5. Graphical representation of sensitivity analysis.

V. Conclusion

This study sheds light on the profound impact of financial theft on Saudi Arabia's economy and culture, underscoring the considerable financial strain it places on individuals. The complex business landscape in the Kingdom of Saudi Arabia (KSA) poses challenges for traditional methods of digital banking fraud detection. Our research set out with several objectives, with a primary emphasis on developing robust DM models within the Saudi financial sector, specifically geared toward effective fraud detection and prevention. Through the application of a novel intuitionistic fuzzy-based DM approach, we aimed to create models capable of assessing fraud detection and prevention in Saudi digital banking systems. The outcomes presented in this paper highlight the efficacy of our proposed approach in detecting and preventing fraud in Saudi Arabian digital banking applications, surpassing the performance of both fuzzy ANP and conventional ANP models. As a final recommendation, we advocate for the implementation of an artificial intelligence (AI) DM program to mitigate fraud in Saudi Arabia. Given the hybrid nature of DM processes in Saudi Arabia, such an algorithm holds promise for significantly enhancing effectiveness and application, ultimately playing a vital role in fortifying fraud protection within the dynamic landscape of KSA's digital banking ecosystem.

Acknowledgment

This research was supported by Arab Open University (AOU)/ KSA. Author is thankful for providing the fund to carry out the work.

References

[1] Unleash your ecommerce growth, Riskified, [Online]. Available at: https://www.riskified.com/.

[2] Trust the person behind the device, Nudatasecurity, [Online]. Available at: https://nudatasecurity.com/.

[3] AffairsCloud YouTube Channel, Affairscloud, [Online]. Available at: https://www.youtube.com/channel/UCkpXde9qr9rmEB1mWGfzfIQ/videos.

[4] D. Krause, "Mitigating Risks for Financial Firms Using Generative AI Tools," SSRN, [Online]. Available at: https://ssrn.com/abstract=4452600.

[5] Penetration rate of online banking in Saudi Arabia from 2013 to 2028, Statista. [Online]. Available at: https://www.statista.com/forecasts/1150349/online-banking-penetration-forecast-in-saudi-arabia.

[6] Payments Fraud and Control Report, J.P. Morgan, [Online]. Available at: https://www.jpmorgan.com/content/dam/jpm/commercial-banking/insights/cybersecurity/highlights-afp-2022-payments-fraud-and-control-report.pdf

[7] A. K. S. Yadav and M. Sora, "Fraud detection in digital banking using text mining models: A review," IOP Conference Series: Materials Science and Engineering, vol. 1020, no. 1, p. 012012, 2021.

[8] S. Chen, "Detection of fraudulent digital banking using the hybrid data mining approach," SpringerPlus, vol. 5, no. 1, pp. 1-16, 2016.

[9] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," Statistical Science, vol. 17, no. 3, pp. 235-255, 2002.

[10] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 50-59, 2004.

[11] P. Ravisankar, V. Ravi, G. R. Rao, and I. Bose, "Detection of digital banking fraud and feature selection using data mining techniques," Decision Support Systems, vol. 50, no. 2, pp. 491-500, 2011.

[12] J. N. Dharwa and A. R. Patel, "A data mining with a hybrid approach-based transaction risk score generation model (TRSGM) for fraud detection of online financial transactions," International Journal of Computer Applications, vol. 16, no. 1, pp. 18-25, 2011.

[13] Z. Zojaji, R. E. Atani, and A. H. Monadjemi, "A survey of credit card fraud detection techniques: data and technique-oriented perspective," arXiv preprint arXiv:1611.06439, 2016.

[14] W. Y. Moon and S. D. Kim, "Adaptive fraud detection framework for fintech based on machine learning," Advanced Science Letters, vol. 23, no. 10, pp. 10167-10171, 2017.

[15] R. Wedge et al., "Solving the false positives problem in fraud prediction using automated feature engineering," in Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part III, pp. 372-388, Springer International Publishing, 2019.

[16] B. Stojanović et al., "Follow the trail: Machine learning for fraud detection in Fintech applications," Sensors, vol. 21, no. 5, p. 1594, 2021.

[17] T. Pi, H. Hu, J. Lu, and X. Chen, "The analysis of Fintech risks in China: Based on fuzzy models," Mathematics, vol. 10, no. 9, p. 1395, 2022.

[18] Every day we build, collaborate and partner to create a world where everyone can transact online with confidence, Gbgplc, [Online]. Available at: https://www.gbgplc.com/en/about-us/.

[19] Transactin Trust, Feedzai, [Online]. Available at: https://feedzai.com/.

[20] Game-changing innovation. Generative AI for good., Featurespace, [Online]. Available at: https://www.featurespace.com/.

[21] M. Taqi et al., "Village fund financial fraud prevention model using the analytical network process model," Jurnal Organisasi dan Manajemen, vol. 17, no. 2, pp. 203-216, 2021.

Arabic Regional Dialect Identification (ARDI) using Pair of Continuous Bag-of-Words and Data Augmentation

Ahmed H. AbuElAtta, Mahmoud Sobhy*, Ahmed A. El-Sawy, Hamada Nayel
Department of Computer Science-Faculty of Computers and Artificial Intelligence,
Benha University, Egypt

***Abstract*—Author profiling is the process of finding characteristics that make up an author's profile. This paper presents a machine learning-based author profiling model for Arabic users, considering the author's regional dialect as a crucial characteristic. Various classification algorithms have been implemented: decision tree, KNN, multilayer perceptron, random forest, and support vector machines. A pair of Continuous Bag-of-Word (CBOW) models has been used for word representation. A well-known data set has been used to evaluate the proposed model and a data augmentation process has been implemented to improve the quality of training data. Support vector machines achieved a 50.52% f1-score, outperforming other models.**

***Keywords*—Dialect identification; continuous Bag-of-Words; data augmentation; text classification.**

I. Introduction

The Arabic language presents a captivating and challenging duality. Its source stems from its historical significance, the strategic importance of its native speakers and their region, along with its abundant cultural and literary legacy. Simultaneously, its complex linguistic framework poses difficulties [1]. More than 330 million people speak Arabic as their native tongue, and as a Semitic language, it has several distinctive linguistic features such as right-to-left writing, and the presence of a dual number of nouns. One of the most prominent features observed in Semitic languages, including Arabic, is the utilization of both female and male genders alongside the root. This aspect stands out significantly and distinguishes these languages [1].

The various Arabic dialects that exist today are together referred to as the Arabic language. There is one "written" form that is used to write Modern Standard Arabic (MSA), while numerous "spoken" variants based on the regional dialect have been used. Due to its use in official settings and written communication, the sole variant undergoes standardization, regulation, and formal instruction in educational institutions. When compared to MSA, regional dialects, which are mostly employed for spoken communication and daily interactions, are still slightly absent from written communication. The same letters used in MSA and the same (mainly phonetic) spelling conventions of MSA can, however, be utilized to create Dialectal Arabic (DA) text [2].

The possible data source is of importance to two key sectors, the commercial sector where marketing intelligence places a larger value on client information including age, gender, nationality, and native language, and the security industry is responsible for guarding against crimes like plagiarism and identity theft, among others, on the internet. As a result, the research community encourages scientists to find and create efficient procedures and methodologies in related disciplines like plagiarism detection and author profiling [3].

The use of DA is prevalent on social media platforms. Computational linguists could generate vast datasets that could be employed in statistical learning environments by gathering information from such sources. It is a challenge to differentiate and separate the dialects from one another; as all Arabic dialects share the same character set and a large portion of their vocabulary. There are six main Arabic regional dialects in addition to MSA, which is typically not commonly spoken as a primary language. Fig. 1 shows the regional dialects of Arabic world.

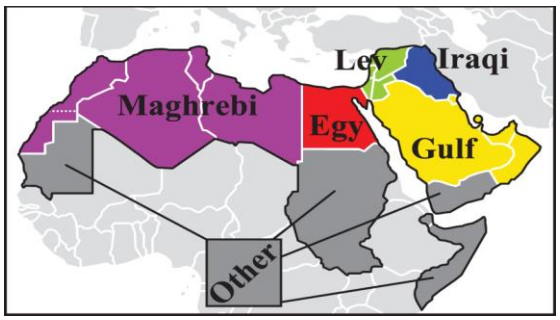


Fig. 1. Regional dialects of Arabic world [2].

- Egyptian: The dialect that is most generally known and understood, due to Egypt's strong film and television industries. As well as it's significant influence through-out a significant portion of the 20th century [4].
- Levantine: A group of dialects linked to Aramaic that sound somewhat different and have different intonations but are substantially comparable when written [5].
- Gulf: The regional dialect that is most like MSA as the current version of MSA is developed from an Arabic dialect that originated in the Gulf region. Compared to other variants, the Gulf dialect has retained a greater portion of MSA's verb conjugation, despite the variances [6].

- Iraqi: Although it has distinctive qualities of its own in terms of prepositions, verb conjugation, and sound, it is occasionally regarded as one of the Gulf dialects [6].
- Maghrebi: French and Berber had a big influence on this dialect. In spoken form, the western-most dialects may be incomprehensible to speakers from other Middle Eastern countries [7].

Due to the complexity of the Arabic language's morphology, the dearth of datasets, and most of the available datasets are imbalanced. Arabic research obtained little attention in its primary phases, especially regarding dialect identification. There are many challenges caused by the high similarity of dialects, particularly in short phrases, such as:

- The same word might have similar meanings in different dialects, for example, the word "**كتاب**" (pronounced "Ketab") means book.
- For the same dialect, there are different short phrases with the same meanings. For example, in Egyptian dialect, the words "**طيب**" (pronounced "Tayb"), "**حاضر**" (pronounced "Hader"), "**عنيا**" (pronounced "Enya"), "**انت تؤمر**" (pronounced "Enta To'mor"), "**إشطه**" (pronounced "Eshta") means "ok".

Effective dialect identification improves the performance of different applications and services, such as machine translation, Automatic Speech Recognition (ASR), remote access, e-commerce, e-learning, and exposing forensic evidence. Arabic dialect identification has been performed at the regional-level (e.g., Levant, Gulf) [8], country-level (e.g., Egypt, Saudia Arabia) [9], and province-level (e.g., Cairo, Al-Madinah) [10]. This work concentrates on the challenge of Arabic Regional Dialect Identification (ARDI) for social media users. We propose machine learning-based classification models to perform ARDI for Arabic tweets. A word embedding model has been used for word representations and a data augmentation process has been applied to improve the quality of data. In the classification step, Decision Tree (DT), K-Nearest Neighbor (KNN), Multi-Layer Perceptron (MLP), Random Forest (RF), and Support Vector Machines (SVM) algorithms were used.

The remaining portions of the paper are structured as follows: Related work is introduced in Section II. Section III describes the dataset that has been used for model development. Section IV introduces the general architecture of the proposed model. The results obtained by the proposed model presented in Section V. The detailed explanation of the results is represented in Section VI, while Section VII concludes the paper.

II. Related Work

There are major efforts that have been carried out for ARDI; some of these works will be described in this section. The First Nuanced Arabic Dialect Identification Shared Task (NADI 2020) has been presented in [11]. This shared task includes the identification of Arabic countries as subtask-1, and the identification of Arabic provinces as subtask-2. The dataset for NADI 2020 covers 21 Arab countries including 100 provinces obtained from Twitter. The baseline, Google's mBERT, model was fine-tuned with 50 tokens as a sequence's maximum length and 8 batches. Various approaches have been applied for NADI such as Machine Learning (ML) approaches and Deep Learning (DL) approaches, and the best performed model achieved a 26.78% f1-score for the first subtask and 6.39% for the second subtask. NADI 2021 was the second shared task that aimed at identifying the linguistic diversity of brief texts based on small geographical regions of origin in Arabic dialects [10]. In NADI 2021, an unlabeled corpus for 10M tweets has been added for optional use. The same baseline model for NADI 2020 was fine-tuned in addition to the ML and DL approaches. The best winner model reported 22.38%, 32.26%, 6.43%, and 8.60% f1-score for country-level-MSA, province-level-MSA, country-level-Dialectal Arabic, and province-level-Dialectal Arabic, respectively. Another NADI shared task in 2022 [9] aimed at the identification of Arabic country-level dialects. Three baselines were finetuned in NADI 2022, Baseline-mBERT, Baseline-XLMR, and Baseline-MARBERT in addition to some pre-trained models based on the BERT model. The top systems reported 36.48% and 18.95% f1-score for Test-A and Test-B sets respectively.

Antoun et al. [12], introduced the AraBERT model which trained on a large Arabic corpus and achieved state-of-the-art on various Arabic NLP tasks, including sentiment analysis, named entity recognition, and question answering. The BERT-base configuration was 12 encoder blocks, 768 hidden dimensions, 12 attention heads, 512 maximum sequence length, and a total of 110M parameters. The model outperforms the multilingual version of BERT and other previous approaches. Another transformer-based model designed for Arabic language understanding was introduced by Abdul-Mageed et al. [8]. They introduced ARBERT and MARBERT, bidirectional transformers for Arabic language processing, focusing on MSA and Arabic Dialects respectively. A random 1B Arabic tweets were selected to train MARBERT, and a dataset of about 6B tweets was formed. Tweets greater than two Arabic words were only included. MARBERT was trained for 36 epochs with 256 batch size and 128 sequence length and achieved high scores in various Arabic dialects datasets.

Talafha et al. [13] used BERT architecture for NADI and achieved a 26.78% f1-score. Also, Gaanoun and Benelallam [14] presented an Arabic-BERT model combined with ensemble methods and data augmentation for NADI. The Arabic-BERT model was trained on the provided training data; then data augmentation was performed by splitting the training data into three parts and mixing them for each country. The augmented data was used to train multiple models, including the "Mix" model, which showed good performance and obtained an f1-score of 23.26% and 5.75% for country-level and province-level respectively. A combination of BERT and N-GRAM characteristics was presented in [3]. The authors investigated the task for the identification of dialects at the national and provincial levels. They introduced an ensemble model that achieved promising results. M-NGRAM uses TF-IDF with character and word n-grams and a Stochastic Gradient Descent (SGD) classifier. The ensemble method achieved a f1-score of 25.99% and 6.39% for country-level identification and province-level identification, respectively.

The contributions of this paper include: (1) develop five classification models based on a pair of Continuous Bag-of-Words model (CBOW) for ARDI, (2) build a new corpus from various datasets to use for building our CBOW model, (3) apply the data augmentation process to enhance the quality of training data.

III. DATA

The dataset that has been used in this research is ArSarcasm [8]. Which is a collection of Arabic sentiment analysis datasets called SemEval 2017 [15] and ASTD [16]. The dataset contains 10,547 tweets modified by adding dialect labels. The distribution of the train set, and test set over all classes is shown in Table I.

TABLE I. DISTRIBUTION OF TRAINING SET AND TEST SET IN EACH CLASS OF ARSARCASM DATASET

Arabic Region	Number of documents in train set	Number of documents in test set
msa	5652	1410
egypt	1904	479
levant	439	112
gulf	414	105
maghreb	28	4
Total	8,437	2,110

The dataset contains the following fields:

- tweet: the text of the original tweets.
- sarcasm: a Boolean value indicating whether a tweet is sarcastic.
- sentiment: the new annotation's sentiment (good, neutral, or negative).
- source: the original tweet's SemEval or ASTD source.
- dialect: the Arabic regional dialect used in tweets, msa, egypt, levant, gulf, and maghreb, which were shown in Section I.

IV. METHODOLOGY

The proposed model uses classic ML-based classifiers integrated with a data augmentation approach for word representation. As shown in Fig. 2, the proposed system is divided into five primary phases, including data augmentation, text preprocessing, feature extraction, classification, and evaluation.

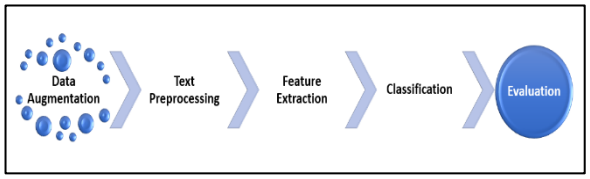


Fig. 2. General architecture of the proposed system.

In the data augmentation phase, the training data was augmented with extra Arabic text from itself and other datasets. Some text cleaning steps and adjustments were applied in the text preprocessing phase. In the feature extraction phase, all documents have been represented based on the embeddings of words in each document. In the classification phase, we classify each vector in the feature vectors to its correct class. Finally, the proposed model has been evaluated by measuring its performance in the evaluation phase. All steps are explained in the following sections.

A. Data Augmentation

- Data augmentation basically allows one to artificially increase the training set by making updated copies of training samples using existing data [17]. It involves modifying the dataset slightly or creating new data points using deep learning approaches. Data augmentation is used for several purposes; (1) to prevent overfitting; (2) when the initial training set is insufficient; (3) to increase the accuracy of models; and (4) to handle the unbalanced dataset. As shown in Table I, the training data is unbalanced as “msa” and “egypt” classes were much bigger than the other classes, as well as the portion of class “maghreb” is too small. In this work, several text augmentation approaches have been used, like:
- Rearranging words or sentences in a random manner.
- Substituting words with their synonyms.
- Rephrasing sentences using the same meanings.
- Insertion or deletion of words at random.

Furthermore, to increase the quality of the dataset, the original training data has been augmented with other datasets in the same domain. In this step, we selected some datasets which have the same Arabic dialect texts as our dataset. Especially, we selected records of data with the same regions in our dataset. The first external dataset is NADI 2022 [9] which focused on nuanced Arabic dialect identification at country-level for Arabic tweets and covers 18 dialects (a total of approximately 20K tweets). The second external dataset is NADI 2021 [10], which covers MSA and DA. The dataset contains a training set of 21,000 tweets, a development set of 5,000 tweets, and a test set of 5,000 tweets. Habibi is the third external dataset [18] which is the earliest Arabic song lyrics corpus. More than 30,000 Arabic song lyrics by vocalists from 18 different Arabic countries are included in the corpus, which includes songs in six Arabic dialects.

TABLE II. DISTRIBUTION OF TRAINING SET FOR EACH CLASS BEFORE AND AFTER THE AUGMENTATION PROCESS.

Class (Arabic region)	Number of documents before augmentation	Number of documents after augmentation
msa	5652	5652
egypt	1904	6187
levant	439	4227
gulf	414	4484
maghreb	28	3554
Total	8,437	24.104

More than 500,000 sentences (song verses) and more than 3.5 million words make up the lyrics [18]. The benefit of this corpus is that all words are written in DA not in MSA. After the data augmentation step was done, the dataset became balanced somewhat and the next step was data preprocessing. Table II shows the distribution of training data in each class before and after the augmentation process.

B. Text Preprocessing

To get the data ready for training, light preprocessing has been used which preserves a true representation of the text that naturally appears. Emojis, Latin letters, URLs, mentions, numerals, and non-Arabic characters were all excluded from the data because Arabic texts, particularly those found on social media, are unstructured and exceedingly loud. In addition, the following steps have been implemented.

- Convert the various forms of Arabic characters into their unique forms, such as "ة " (pronounced as Haa) and " ه " to be " ه".
- Deleting extraneous Arabic forms, such as," ال " (pronounced "al") and it operates as a determiner.
- Deleting punctuation marks such as {'?'; '.'; '!'; '$'} which make more unnecessary features that can expand the dimension of the feature space.
- Reducing the letter repetition since Arabic tweets tend to be less structured. Clearing the letters from the extraneous tokens helps in reducing feature space. In this work, we considered the letter, which is repeated more than twice as redundant. For example, the word "كاااامل" "(pronounced as "Kamel") which means "complete" will be decreased to "كامل", also the word "رهيييب"(pronounced as "Rahib") which means "awesome") will be reduced to "رهيب ".

C. Feature Extraction

In this phase, all tweets are represented as feature vectors, each of which contains an embedding for each word in the tweet. A variety of approaches are employed to get the word embedding vector from the context in which the words are found. In this study, a pair of Word2vec models has been used. Google has suggested the Word2vec neural network [19] to analyze text input. The Word2vec model is a neural network with three layers: an input layer, an output layer, and a hidden layer without activation function. Additionally, the number of neurons in the hidden layer is the same size as the word embeddings' vector's dimensions. The Word2Vec model makes use of huge datasets during training to precisely capture the semantic and syntactic structure of the words, allowing for the efficient measurement of word similarity [19].

Continuous Bag of Words (CBOW) and Skip-gram are the two learning models included in Word2Vec: CBOW predicts the word given its context, while Skip-gram predicts the context given a word as shown in Fig. 3. The window size and vocabulary size are two hyperparameters that are shared by the two methods. The window size indicates the number of words in the context. Given the near future and historical words, the CBOW technique classifies the projected middle word using a log-linear classifier.

The number of words in the context is equal to the size of the sliding window; for example, if the sliding window is seven words, then there are six words in the context. Additionally, while predicting a word, the context of the preceding three words and the succeeding three words of the middle word must be considered.

The first Word2Vec model that has been used was built using UNLABELED-10M, a set of 10 million unlabeled Arabic tweets provided by NADI [9] in the form of tweet IDs. A ready implementation of Word2Vec model, gensim [20], using CBOW has been used to generate word vectors of size 300.

Nevertheless, the embedding vectors for words in the first model were not enough to cover all the words in the dataset. So, we created another Word2Vec model. We employed CBOW to generate the word vectors as it has higher computing speed, and it is more efficient with frequent words than Skip-gram [21]. The vocabulary has been built from the entire training data and some external aforementioned datasets, NADI 2021 shared task dataset [10], and Habibi corpus [18]. Previously, if the embedding vector did not exist in the corpus, the vector was set randomly. Now, it is obtained from the second corpus, and this increases the correctness of the vector. Following the training phase, a vector is used to represent each word.

Next, we construct the high dimension matrix. Rows in the matrix represent the training tweets and columns represent words. The classification phase, which is described in the following section, follows the creation of the feature vector matrix for all training instances.

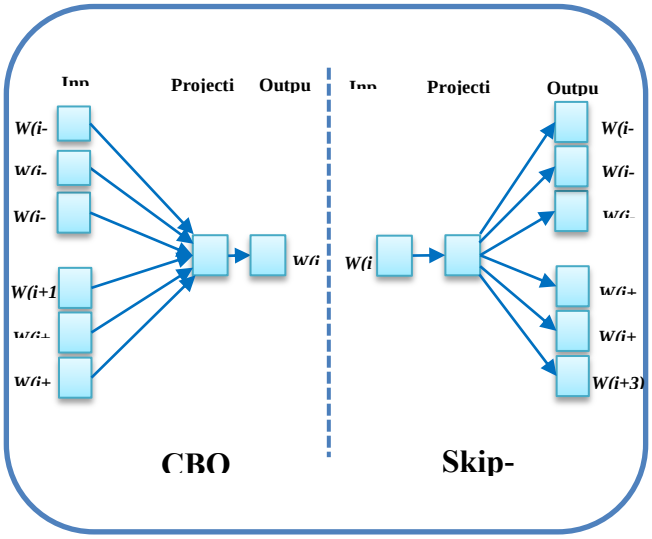


Fig. 3. The architecture of CBOW model.

D. Classification

Five classification algorithms, DT, KNN, MLP, RF, and SVM, were used in this study. To improve the performance of these classifiers, their hyperparameters were utilized.

The SVM is a linear classifier that uses training samples close to the borders of classes [22]. The SVM model uses kernel functions for classifying non-linear data such as linear,

sigmoid, and Radial Basis Function (RBF) kernels, which were used in this work. The KNN algorithm assumes that the new sample and the available samples are similar, and it places the new sample in the category that resembles the available categories [23]. The DT classifier [24] utilizes the decision tree as a model for making pre-dictions based on observations about the items that are represented in the tree's branches to inferences about the target value of items that are represented in the decision tree's leaves.

The RF is an average-based meta-estimator that is used to increase predictive accuracy and reduce overfitting by applying several decision tree classifiers to various dataset sub-samples [25]. The MLP is a completely connected class of feedforward Artificial Neural Network (ANN) [26]. A typical MLP has an input layer, a hidden layer, and an output layer, which make up together less than or equal to three layers of nodes. Every node uses a nonlinear activation function, apart from the input nodes.

E. Evaluation

All algorithms mentioned in the paper that were evaluated on the ArSarcasm dataset. The evaluation metrics that have been used are Accuracy (Acc), Precision (P), Recall (R), and f1-score [27]. Accuracy measures the number of truly classified tweets divided by all tweets. Precision is another metric that calculates the number of correctly classified tweets divided by all classified tweets. Another metric is Recall, which calculates the number of correct correctly classified tweets divided by all correct tweets. The macro-averaged f1-score is the official metric for most NLP tasks as it is the most basic aggregation for the f1-score. The macro-averaged f1-score is the unweighted mean of the f1-scores determined for each class. The formula for macro f1-score [27] is:

$$f1\text{-}score = 2 * \frac{R*P}{R+P} \quad (1)$$

V. Experimental Results and Discussion

To evaluate the proposed model, several experiments have been carried out using different parameters as shown in Table III.

TABLE III. Parameters of Some Proposed Models.

Algorithm	Parameters
SVM	Kernel functions = {linear, sigmoid, RBF}
KNN	Number of neighbors (*n*) = {30, 40, 50}
MLP	Number of hidden layers (*h*) = {5, 10, 20}

The proposed models have been tested using two variations of word embedding models. The results of the proposed models with the first embedding (UNLABELED-10M) without augmentation and with augmentation are shown in Table IV and Table V respectively. The results of the proposed models with the second embedding model (CBOW) without augmentation and with augmentation are shown in Table VI and Table VII respectively.

As shown in Table I above, the test data were unbalanced; the number of documents with the label (msa) is 1410, whereas it was only 4 with the label (maghreb). This was a big challenge to classify at least two documents with the label (maghreb) correctly. Fig. 4 shows the confusion matrix plot for the SVM (RBF) classifier, which has the best results.

It is clear that, the SVM (RBF-kernel) with the augmentation of data and the pair of Word2Vec models outperformed all other classifiers. Table IV shows the results of using the CBOW model, UNLABELED-10M, without data augmentation. The SVM classifier with RBF kernel function has achieved the highest accuracy, precision, and f1-score while MLP with five hidden layers has achieved the highest recall. Table V shows that the SVM classifier with RBF kernel function has achieved the highest accuracy, and F1-score, while MLP with five hidden layers has achieved the highest precision and the highest recall has achieved by MLP with 10 hidden layers.

TABLE IV. Performance of using UNLABELED-10M Model (without augmentation).

Algorithm	Parameter	P	R	f-score	Acc
SVM	linear kernel	46.683	39.737	41.902	77.014
	sigmoid kernel	39.296	35.618	36.684	70.900
	RBF kernel	**52.495**	40.650	43.478	**78.027**
KNN	n = 30	47.853	38.152	40.250	77.014
	n = 40	48.833	37.947	40.057	77.156
	n = 50	48.261	37.777	39.198	76.730
DT		32.024	32.815	32.365	64.739
RF		43.333	33.292	35.057	76.398
MLP	h = 5	44.125	45.456	42.512	74.041
	h = 10	47.568	**47.453**	44.542	73.652
	h = 20	48.225	47.342	**44.654**	72.850

TABLE V. Performance of using UNLABELED-10M Model (with Augmentation).

Algorithm	Parameter	P	R	f-score	Acc
SVM	linear kernel	48.195	45.783	43.214	75.877
	sigmoid kernel	39.896	42.723	40.115	69.384
	RBF kernel	**56.008**	46.530	**46.050**	**78.341**
KNN	n = 30	48.460	38.436	40.570	77.204
	n = 40	47.262	37.287	40.900	69.005
	n = 50	46.334	36.564	38.152	68.512
DT		35.701	36.096	35.849	67.915
RF		43.420	30.964	32.567	75.450
MLP	h = 5	49.105	46.466	43.758	75.071
	h = 10	48.940	**48.403**	45.655	74.739
	h = 20	47.334	47.042	44.259	73.791

In Table VI, the results of using two CBOW models, UNLABELED-10M, and our own CBOW model, without data augmentation. This table shows that the SVM classifier with RBF kernel function has achieved the highest accuracy and

recall while MLP with five hidden layers has achieved the highest f1-score and precision. Table VII shows the results of using the same pair of CBOW models in Table V after the data augmentation process was done. This table shows that the SVM classifier with RBF kernel function has achieved the highest performance for all metrics.

TABLE VI. PERFORMANCE OF USING PAIR CBOW MODELS (WITHOUT AUGMENTATION).

Algorithm	Parameter	P	R	f-score	Acc
SVM	linear kernel	49.159	44.001	45.777	77.488
	sigmoid kernel	38.310	35.141	36.059	70.900
	RBF kernel	**55.018**	41.684	44.556	**78.863**
KNN	n = 30	48.460	38.436	40.570	77.204
	n = 40	49.090	38.015	40.338	76.967
	n = 50	47.450	37.263	39.274	76.872
DT		33.360	34.956	33.999	65.355
RF		46.045	33.991	36.091	76.730
MLP	h = 5	49.250	**44.225**	**46.016**	77.583
	h = 10	48.068	43.591	45.230	77.204
	h = 20	46.603	42.883	44.259	77.062

TABLE VII. PERFORMANCE OF USING PAIR CBOW MODELS (WITH AUGMENTATION).

Algorithm	Parameter	P	R	f-score	Acc
SVM	linear kernel	48.820	50.844	47.087	76.303
	sigmoid kernel	40.619	41.186	38.478	69.100
	RBF kernel	**50.294**	**55.893**	**50.534**	**77.251**
KNN	n = 30	47.557	47.748	41.723	68.578
	n = 40	48.314	47.672	41.506	69.194
	n = 50	48.049	48.080	41.614	69.431
DT		33.013	48.632	32.989	57.630
RF		46.795	46.722	42.798	74.787
MLP	h = 5	47.185	50.548	46.610	75.640
	h = 10	49.018	52.481	47.272	75.924
	h = 20	46.714	54.902	46.545	74.313

TABLE VIII. F1-SCORE OF OUR BEST MODEL AND PREVIOUS MODELS.

Algorithm	F1-score
mBERT [8]	43.81
XLM-R_L [8]	41.83
AraBERT [8]	47.54
SVM-RBF-With-Augmentation	**50.53**

Table VIII compares the performance of the proposed model and state-of-the-art in terms of f1-score. The results show that the proposed model outperformed the state-of-the-art models using low resources than those previous models which use huge resources to train language models using billions of words; as the proposed model depends on two embeddings and augmented data.

Fig. 4 shows the change in the values in the confusion matrix with the three attempts we made in the augmentation step. Let's focus on the true and predicted values of maghreb. In Fig. 4(a), without augmenting training data, the four test samples were misclassified. In Fig. 4(b), after the augmentation of the training samples of the maghreb class, 25% of the test samples were correctly classified. In Fig. 4(c), after the augmentation of all training samples, 50% of the test samples were correctly classified.

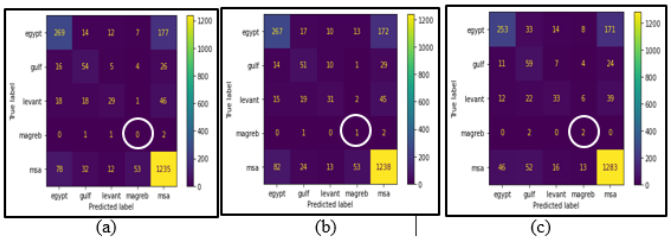


Fig. 4. Confusion matrix of classes after applying SVM with RBF kernel: (a) without augmentation; (b) with augmentation of magreb documents only; (c) with augmentation of documents of all classes.

I. CONCLUSION

Due to its complexity, ARDI becomes a challenge. This work proposed a machine learning-based model that implements DT, KNN, MLP, RF, and SVM for ARDI. A pair of CBOWs has been used for data representation and a data augmentation approach has been implemented to overcome the problem of imbalanced data. SVM reported 50.53% f1-score which was higher than previous work. The results proved that using a pair of CBOW models is better than using only one and proved that data augmentation was very useful for improving the quality of data. In future work, different representation models can be used to improve the performance of ARDI such as BERT models.

Authors' Contribution: Conceptualization, M.A. and H.N.; methodology, M.A. and H.N.; software, A.A. and M.A.; validation, H.N., A.E. and A.A.; formal analysis, M.A. and H.N.; resources, M.A. and H.N.; data curation, A.A. and A.E.; writing—original draft preparation, M.A. and H.N.; writing—review and editing, A.A. and A.E.; visualization, M.A. and H.N.; supervision, A.E.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: the dataset that has been used in this article is open source and can be downloaded from https://github.com/iabufarha/ArSarcasm.

Conflicts of Interest: We confirm that neither the manuscript nor any parts of its content are currently under consideration or published in another journal. All authors have approved the manuscript and agree with its submission to the journal "Information".

REFERENCES

[1] I. Al-Huri, "Arabic Language: Historic and Sociolinguistic Characteristics English Literature and Language Review Arabic Language: Historic and Sociolinguistic Characteristics," vol. 1, no. 4, pp. 28–36, 2015, doi: 10.13140/RG.2.2.16163.66089/1.

[2] O. F. Zaidan and C. Callison-Burch, "Arabic Dialect Identification," 2014, doi: 10.1162/COLI.

[3] A. Abbassi, S. Mechti, L. Hadrich Belguith, and R. Faiz, "Author Profiling for Arabic Tweets based on n-grams." [Online]. Available: http://www.internetlivestats.com/

[4] N. Haeri, "Sacred language, ordinary people: Dilemmas of culture and politics in Egypt," Sacred Language, Ordinary People: Dilemmas of Culture and Politics in Egypt, pp. 1–184, Jan. 2003, doi: 10.1057/9780230107373/COVER.

[5] A. H. Aliwy, H. A. Taher, and Z. A. Abutiheen, "Arabic Dialects Identification for All Arabic countries," pp. 302–307, 2020.

[6] R. Bassiouney, A. Sociolinguistics, and M. Amara, "Reem Bassiouney: Arabic Sociolinguistics," Language Policy 2010 9:4, vol. 9, no. 4, pp. 379–381, May 2010, doi: 10.1007/S10993-010-9169-0.

[7] M. Tilmatine, "Substrat et convergences: le berbère et l¿arabe nord-africain," 1999.

[8] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic," Dec. 2020, [Online]. Available: http://arxiv.org/abs/2101.01785

[9] M. Abdul-Mageed, C. Zhang, A. Elmadany, H. Bouamor, and N. Habash, "NADI 2022: The Third Nuanced Arabic Dialect Identification Shared Task," Oct. 2022. Available: http://arxiv.org/abs/2210.09582

[10] M. Abdul-Mageed, C. Zhang, A. Elmadany, H. Bouamor, and N. Habash, "NADI 2021: The Second Nuanced Arabic Dialect Identification Shared Task," Mar. 2021. Available: http://arxiv.org/abs/2103.08466

[11] M. Abdul-Mageed, C. Zhang, H. Bouamor, and N. Habash, "NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task." pp. 97–110, 2020. Accessed: Aug. 08, 2023. Available: https://aclanthology.org/2020.wanlp-1.9

[12] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," Feb. 2020. Available: http://arxiv.org/abs/2003.00104

[13] B. Talafha et al., "Multi-Dialect Arabic BERT for Country-Level Dialect Identification," 2020. Available: https://github.com/mawdoo3/Multi-dialect-Arabic-BERT

[14] K. Gaanoun and I. Benelallam, "Arabic dialect identification: An Arabic-BERT model with data augmentation and ensembling strategy," 2020.

[15] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 Task 4: Sentiment Analysis in Twitter," pp. 502–518, Accessed: Aug. 06, 2023. Available: https://trends24.in/

[16] M. Nabil, M. Aly, and A. F. Atiya, "ASTD: Arabic Sentiment Tweets Dataset," pp. 17–21, 2015, Accessed: Aug. 06, 2023. Available: https://github.com/boto/boto.

[17] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text Data Augmentation for Deep Learning," Journal of Big Data 2021 8:1, vol. 8, no. 1, pp. 1–34, Jul. 2021, doi: 10.1186/S40537-021-00492-0.

[18] M. El-Haj, "Habibi-a multi Dialect multi National Arabic Song Lyrics Corpus," pp. 11–16, 2020, Accessed: Aug. 06, 2023. [Online]. Available: www.figure-eight.com/

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space", Accessed: Aug. 06, 2023. [Online]. Available: http://ronan.collobert.com/senna/

[20] R. Rehurek and P. Sojka, "Gensim -- Statistical Semantics in Python," 2011.

[21] D. Suleiman and A. Awajan, "Comparative Study of Word Embeddings Models and Their Usage in Arabic Language Applications," ACIT 2018 - 19th International Arab Conference on Information Technology, Mar. 2019, doi: 10.1109/ACIT.2018.8672674.

[22] K. P. Ukey and A. S. Alvi, "Text Classification using Support Vector Machine", Accessed: Aug. 09, 2023. [Online]. Available: www.ijert.org

[23] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved K-nearest-neighbor algorithm for text categorization," Expert Syst Appl, vol. 39, no. 1, pp. 1503–1509, Jan. 2012, doi: 10.1016/J.ESWA.2011.08.040.

[24] F. Harrag, E. El-Qawasmeh, and P. Pichappan, "Improving Arabic text categorization using decision trees," 2009 1st International Conference on Networked Digital Technologies, NDT 2009, pp. 110–115, 2009, doi: 10.1109/NDT.2009.5272214.

[25] N. Jalal, A. Mehmood, G. S. Choi, and I. Ashraf, "A novel improved random forest for text classification using feature ranking and optimal number of trees," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 6, pp. 2733–2742, Jun. 2022, doi: 10.1016/J.JKSUCI.2022.03.012.

[26] H. Alla, L. Moumoun, Y. Balouki, and J. Gou, "A Multilayer Perceptron Neural Network with Selective-Data Training for Flight Arrival Delay Prediction," Sci. Program., vol. 2021, Jan. 2021, doi: 10.1155/2021/5558918.

[27] H. Dalianis, "Evaluation Metrics and Evaluation," in Clinical Text Mining: Secondary Use of Electronic Patient Records, Cham: Springer International Publishing, 2018, pp. 45–53. doi: 10.1007/978-3-319-78503-5_6.

Advanced Metering Infrastructure Data Aggregation Scheme Based on Blockchain

Hongliang TIAN, Naiqian ZHENG, Yuzhi JIAN
School of Electrical Engineering, Northeast Electric Power University, Jilin 132012, China

***Abstract*—Smart grid stands as both the cornerstone of the modern energy system and the pivotal technology for addressing energy-related challenges. Advanced Metering Infrastructure constitute a critical component within the smart grid ecosystem, providing real-time energy consumption data to power utility companies. Advanced Metering Infrastructure enables these companies to make timely and accurate decisions. Hence, the issue of data security pertaining to Advanced Metering Infrastructure assumes profound significance. Presently, Advanced Metering Infrastructure data confronts challenges associated with centralized data storage, rendering it susceptible to potential cyberattacks. Moreover, with the burgeoning number of electricity consumers, the resultant data volumes have swelled considerably. Consequently, the transmission of this data becomes intricate and its efficiency is compromised. To address these issues, this paper presents a lightweight blockchain data aggregation scheme. By integrating fog computing and cloud computing, a three-tier blockchain-based architecture is devised. Initially, digital signatures are employed to ensure the validity and integrity of user data. The innate attributes of blockchain technology are harnessed to safeguard the security of electricity energy data. Through secondary data aggregation, the privacy-sensitive user data is efficiently compressed and subsequently integrated into the blockchain, thereby mitigating the storage pressure on the blockchain and enhancing data transmission efficiency. Ultimately, through rigorous theoretical analysis and simulated experimentation, the paper demonstrates that, in comparison to existing methodologies, lightweight blockchain data aggregation scheme exhibits heightened security. Additionally, lightweight blockchain data aggregation scheme holds a competitive advantage in terms of computational and communication costs.**

***Keywords*—Smart grid; blockchain; advanced metering infrastructure; data aggregation**

I. Introduction

Smart grid epitomizes the evolution and transformation of the power and energy industry, constituting a pivotal platform for the execution of new energy strategies and the optimization of energy resource allocation. Functioning as an intelligent power network, the smart grid is constructed upon the foundation of an integrated and high-speed bidirectional communication network. Leveraging cutting-edge sensing and measurement technologies, advanced equipment methodologies, sophisticated control techniques, and state-of-the-art decision support systems, the power grid is imbued with traits of reliability, safety, cost-effectiveness, efficiency, environmental consciousness, and user security [1]. The deployment of the smart grid has notably catalyzed interaction between users and power utility companies, fostering a two-way exchange of power and data that has substantiated substantial economic and societal gains. The amalgamation of distributed generation, user-oriented energy consumption management, and remote monitoring has been actualized [2]. The realization of these outcome rests upon the bedrock of essential components like smart meters, the proliferation of which has engendered a prodigious volume of corresponding power data. The effective transmission and processing of this data stands as a vital prerequisite for the smart grid's success, as underscored by computational and communication cost considerations [3]. Moreover, the security and confidentiality of this data are of paramount importance, given its role in shaping customer electricity billing and guiding the decisions of power utility companies. Data security concerns encompass tampering, data falsification, and database attacks. In this context, blockchain technology, a decentralized data processing paradigm, emerges as a robust safeguard, with all network nodes being collectively responsible for data storage, thus ensuring comprehensive data security. In tandem with the continuous advancement of blockchain technology, numerous countries have synergistically integrated smart grid infrastructures with blockchain [4]. Within the smart grid domain, Advanced Metering Infrastructure (AMI) store substantial volumes of private user data, thereby adopting blockchain to fortify AMI data protection and storage within the blockchain, effectively mitigating the potential repercussions of data breaches [5]. Simultaneously, leveraging its distributed architecture, fog computing facilitates computations at the network's edge. In comparison to cloud computing, fog computing holds distinct advantages in processing smart meter data [6]. Notably, within smart meters, fog nodes can expedite the processing of user privacy data [7]. Given the shared architectural underpinnings of fog computing and blockchain, their integration holds great potential. Building upon prior research, this paper introduces a lightweight data aggregation schema (LDAS-BC), amalgamating fog computing, blockchain technology, and the Paillier homomorphic encryption algorithm. This schema refines the Paillier encryption system and employs a two-tier aggregation model to achieve granular data aggregation. Moreover, the scheme leverages the lightweight, one-way irreversible properties of hash algorithms to authenticate components, thereby minimizing computational and communication overheads. Overall, the primary contributions of this endeavor are as follows:

1) The introduction of cloud computing and fog computing in a three-tier architecture—comprising the user layer, fog layer, and cloud layer—attenuates performance

limitations arising from the constrained storage and computing resources of network edge devices. The intermediary fog node routinely collects user data from smart meters and significantly enhances data transmission efficiency through secondary aggregation operations.

2) The LDAS-BC scheme harnesses blockchain technology and harnesses fog nodes to formulate a fog chain. This dual-pronged approach not only permits fault tolerance for select fog nodes via consensus and master node selection algorithms but also furnishes clouds with stable and dependable data services.

3) By synergizing improved additive homomorphic encryption techniques, the schema presents a lightweight data aggregation mechanism predicated on blockchain to ensure the privacy of aggregated and transmitted data. Simultaneously, digital signatures founded on hash algorithms underwrite data integrity and validity during transmission.

The ensuing sections delineate the research structure: Section II reviews pertinent literature; Section III expounds upon the network model of the schema; Section IV introduces the master node algorithm tasked with electing a master node from the entire pool; Section V elaborates on the blockchain-rooted data aggregation mechanism. Section VI is dedicated to simulation experiments and performance analyses, culminating in a final comparison of data. Lastly, Section VII encapsulates this paper's findings and outlines potential future directions.

II. Related Work

In recent years, several privacy-conscious data aggregation methodologies have emerged, aiming to safeguard the confidentiality of transmitted data within the smart grid context. Chen et al. [8] introduced a data aggregation scheme founded on the Paillier homomorphic encryption algorithm. Nevertheless, this scheme neglected the constrained computational capabilities of smart meters and was incapable of executing pairing operations. Liang et al. [9] proposed a protocol centered on total homomorphic encryption, yet the intricate implementation of total homomorphic encryption posed a challenge. Gope et al. [10] devised a gradual data aggregation scheme that employed an aggregation tree to consolidate users' energy consumption data. The mechanism required all smart meters to partake in the aggregation process to ensure scheme accuracy. Despite this safeguard against aggregation interruption due to smart meter failure through Ping tests and third-party aggregator (TPA) involvement, the scheme incurred substantial communication overhead. Furthermore, excessive reliance on TPA engendered issues of trust and single points of failure. Singh et al. [11] proposed a privacy-ensuring data aggregation model integrating deep learning and homomorphic encryption to mitigate the adverse effects of flash memory workload on predictive model accuracy. The model facilitated secure data aggregation at low computational costs. However, it failed to account for data volume and suffered from sluggish transmission efficiency.

And in the context of blockchain. Guan et al. [12] introduced a blockchain-based methodology for privacy protection and secure, efficient data aggregation. The scheme employed pseudonyms to mask user identities and maintained data on a private blockchain. Identity authentication primarily relied on Bloom Filters within this framework. Chen et al. [13] advanced a dual blockchain-supported secure and anonymous data aggregation protocol named DA-SADA. The scheme formed a three-tier data aggregation structure via fog computing, incorporating secure and anonymous data aggregation mechanisms involving Paillier additive homomorphic encryption, aggregate signatures, and anonymous authentication, with minimal computational overhead. Faiza et al. [14] conceived PrivDA, a blockchain and homomorphic encryption-based privacy-preserving IoT data aggregation approach, enabling consumer users to create smart contracts to stipulate terms of service and requested IoT data. Cristina et al. [15] innovated a privacy-enhancing distributed security protocol leveraging blockchain and homomorphic encryption for data aggregation, employing homomorphic encryption for data encryption and blockchain smart contracts for aggregation. Bao et al. [16] advanced the BBNP paradigm, a blockchain-grounded model employing data aggregation protocols to safeguard data privacy and communication confidentiality, deploying identity authentication mechanisms for data integrity, and employing the subjective logical reputation model for consensus to address single point of failure. Zhang et al. [17] presented a potent blockchain-oriented multidimensional data aggregation framework using the Byzantine consensus mechanism to designate master nodes for data aggregation and secret sharing-based user management. Zhao et al. [18] introduced a blockchain-rooted privacy protection billing framework underpinned by the BGN encryption scheme, safeguarding users' private billing data. Notwithstanding, none of these solutions offered comprehensive data protection within the smart grid milieu. Yu et al. [19] proposed a privacy-preserving data aggregation and quality assessment protocol driven by smart contracts, storing data on the Inter Planetary File System (IPFS), deriving summary outcomes to evaluate data quality, and allocating rewards based on data quality.

In smart grid and edge computing scenarios. Lu et al. [20] introduced an edge blockchain-aided lightweight privacy-protecting data scheme dubbed EBDA within the smart grid context. Zhang et al. [21] devised LPDA-EC, a lightweight privacy-preserving data aggregation framework tailored for edge computing, ensuring data confidentiality and privacy. Fan et al. [22] introduced DPPDA, a distributed privacy-protecting data aggregation methodology for the smart grid, uniting master node algorithms, the Paillier encryption system, Boneh-Lynn-Shacham short signatures, and SHA-256 capabilities to meet security and privacy requisites for data aggregation in a fledgling grid, ensuring equitable and secure smart grid communications.

III. Smart Grid Network Model Based on Blockchain

A. Main Objectives of the System Model

Currently, substantial challenges persist in ensuring data privacy within the smart grid framework. Firstly, at the user level, the possibility of malevolent entities fabricating spurious data for transmission to fog nodes or tampering with data conveyed by smart meters poses a threat to data validity and

integrity. Consequently, data source verification becomes imperative to establish the authenticity and integrity of transmitted data. Secondly, the issue of raw data exposure remains prominent, as malicious actors could potentially gain access to such data. To mitigate this, encryption of original data is necessary, safeguarding data security from the inception of transmission through storage. Thirdly, the quandary pertains to data storage methodology. The conventional practice of storing data on fog node servers renders it susceptible to malevolent infiltration, jeopardizing power data confidentiality. To counter this, blockchain serves as a robust solution, averting single points of failure and safeguarding against server attacks by storing data in a decentralized manner. Lastly, as user numbers proliferate, so does the voluminous power consumption data. This exponential data growth necessitates data compression to curtail transmission duration, enhance efficiency, and facilitate swift decision-making at the cloud computing center. In light of these challenges, this paper introduces the LDAS-BC scheme, chiefly targeting data integrity, validity, and privacy, while also addressing the efficacy of data storage and transmission.

B. System Model

This section introduces the blockchain-centered smart grid network model, as proposed within the scope of this paper. The schematic representation of this network model is depicted in Fig. 1. The blockchain-based smart grid network model comprises three distinct strata: the user layer, the fog layer, and the cloud layer.

1) User layer: This layer predominantly encompasses an extensive array of smart meters, which, in alignment with their geographic distribution, establish connections with adjacent fog nodes. Concurrently, each smart meter dispatches encrypted data in the form of user reports to the respective area's fog node.

2) Fog layer: Comprising fog nodes endowed with computational capabilities, this layer undertakes the verification of received user reports. Once confirmed as accurate, these nodes perform primary aggregation operations on the ciphertext received from the user layer. Subsequently, an elected master node oversees secondary aggregation, culminating in the storage of the resultant aggregated data onto the blockchain, followed by data transmission to the cloud.

3) Cloud layer: The cloud layer is chiefly responsible for system initialization, data retrieval, decryption, storage, and analysis, among other functionalities. The cloud server is capable of disseminating electricity consumption reports, deduced via analysis, to the associated fog nodes. This facilitates low-latency real-time electricity consumption data queries for users.

The crux of the network model comprises three core entities: the smart meter (SM), the fog node, and the measurement data management system (MDMS). Within this intricate network framework, the Advanced Metering Infrastructure (AMI) network is methodically subdivided into 'm' distinct subregions. Each such subregion encompasses 'n' smart meters, primarily tasked with the collection of users' energy consumption data. $SM_{ij}\left(0 \le i \le n, 0 \le j \le n\right)$ is the i-th smart meter in region j, and smart meters form a user layer. In each region of the user layer, a fog node is strategically positioned for deployment. These fog nodes undertake the pivotal role of data concentration and possess the capability to aggregate the collected data. Spatially situated at the network's periphery, these fog nodes bridge the gap between the user layer and the cloud infrastructure. The measurement data management system resides within the cloud environment, equipped with the capacity to access aggregated data residing on the fog chain. Upon retrieval, this system is proficient in decryption procedures, thereby enabling analysis of the decrypted data. The outcomes of this analysis are subsequently harnessed to formulate pertinent supply strategies. Notably, the MDMS encompasses functionalities akin to cloud servers and reputable third-party entities, thereby encapsulating multifaceted roles within its purview.

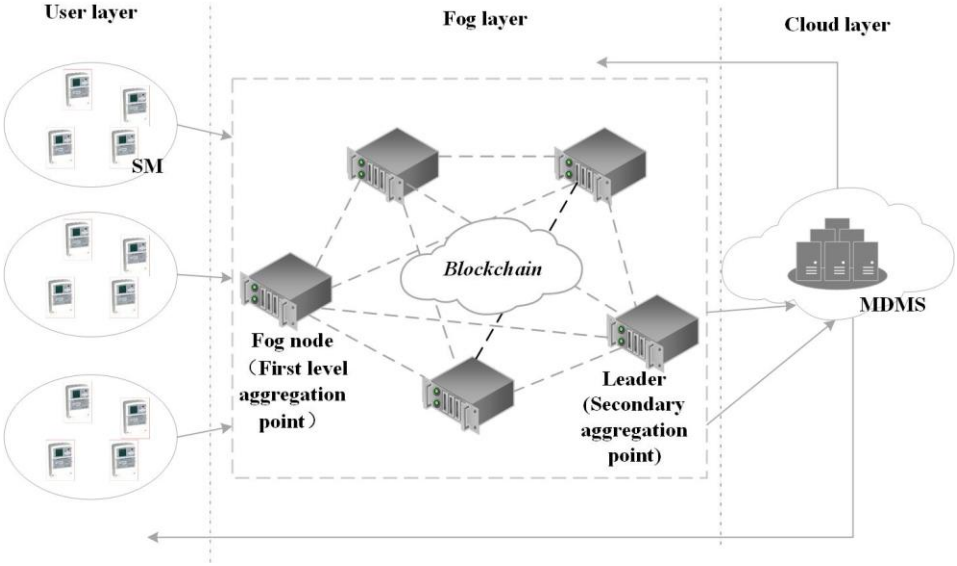


Fig. 1. Network model of LDAS-BC.

IV. Election of the Leader Node

In the realm of blockchain technology, the master node assumes a pivotal role as a fundamental component within the blockchain network. Its principal functions encompass the verification of operations executed by other nodes, the preservation of blockchain integrity, and the packaging of transactions into fresh blocks for incorporation into the blockchain. Within the context of fog nodes, the role of a primary node becomes essential in performing analogous tasks. This section sets forth a novel approach for the election of a primary node, designated to undertake the aforementioned operations.

Algorithm1 Election of the Master Node

1 When $n > 3f + 1$, $DC_i \rightarrow Follower(i \in 1,2,...,n)$, where f is the number of faulty nodes;
2 Set the term number to 0, that is $TN_{DC_i} = 0(i \in 1,2,...,n)$;
3 Set the initial vote number to 0, that is $N_v = 0$;
4 Start timing and express as $Times$;
5 Set a time threshold, that is T_{out};
6 While $Times > T_{out}$ do;
7 $Follower \rightarrow Candidate$;
8 $TN + 1$;
9 $Times$ reset to zero and restart the timer;
10 $N_v + 1$;
11 Send voting requests to other nodes and wait for responses;
12 if the replies from other nodes are received, then the cumulative number of votes N_v is calculated;
13 if $N_v > n/2 + 1$, Where n is the number of nodes, then $Candidate \rightarrow Leader$;
14 end If
15 else (Primary node has been identified);
16 $Candidate \rightarrow Follower$;
17 else
18 Repeat steps 7-11 to start a new election;
19 end If
20 end While

The master node, pivotal in linking aggregated data, subsequently transmitting this data to the cloud, is underscored by three distinct states within each fog node: candidate (sole initiator of the election), follower (participant in the voting process), and master node (sole entity authorized to modify operations). The algorithm governing the master node unfolds across three phases, as delineated below, and is encapsulated in the steps presented within Algorithm 1.

1) Preparation stage: In cases where a primary node is non-existent, the primary node election process is activated. All fog nodes are initialized with a term of 0, and the initial vote tally stands at 0.

2) Voting stage: Upon exceeding the designated time threshold, all nodes transition from follower nodes to candidate nodes, thereby heralding the commencement of the voting process.

3) End stage: Upon the vote count of a given node exceeding half of the total votes, that node ascends to the status of primary node. Subsequently, this node broadcasts its identity as the primary node. Other candidate nodes, in turn, assume the status of followers.

This method ensures the seamless selection of a primary node, vital for executing essential operations within the fog node domain.

V. Data Aggregation Mechanism Based on Blockchain

This section elucidates the LDAS-BC scheme delineated in this paper, which ingeniously amalgamates blockchain technology with Paillier homomorphic encryption technology. The scheme encompasses four integral components: system initialization, user report generation, fog chain generation, and data reading and analysis. A visual representation of these components is visually depicted in Fig. 2.

A. System Initialization

Select *k* as the system security parameter, and compute the Paillier algorithm, public key $(n = p \cdot q, g)$, private key $(\lambda = lcm(p-1, q-1), \mu)$, where p, q are large prime numbers satisfying the $|p| = |q| = |k|$ condition, let $g = N + 1$, ensure that $\mu = \left(L\left(g^{\lambda} \bmod N^2\right)\right)^{-1} \bmod N$ exists. The system randomly selects $r \in Z_N^*$ and calculates $s = r^N \bmod N^2$. Define the function $L(u) = u - 1/N$ and select the safe hash function $h: \{0,1\}^* \rightarrow \{0,1\}^l$.

B. Generation of User Reports

The energy consumption data acquired from smart meters inherently harbors users' confidential information, necessitating the encryption of such data during the collection process. For contextual clarity, this paper postulates a scenario where the private data gathered by a smart meter is transmitted to a fog node in 15-minute intervals. It's worth noting that the smart meter employs encryption mechanisms to secure the energy consumption data every 15 minutes, ensuring data integrity and privacy through digital signatures. During a designated timeframe, each fog node undertakes the aggregation of data submitted by individual smart meters.

1) SM_{ij} generates the corresponding energy consumption data, which is represented by Formula 1, including user ID_{ij}, energy consumption data d_{ij} and time stamp T_p.

$$m_{ij} = ID_{ij} \| d_{ij} \| T_p \tag{1}$$

2) Use the corresponding key (g, n, s) to calculate ciphertext C_{ij}, which can be expressed as Formula 2.

$$C_{ij} = g^{d_{ij}} \cdot s = (N+1)^{d_{ij}} \cdot s = (1 + d_{ij} n) \cdot s \tag{2}$$

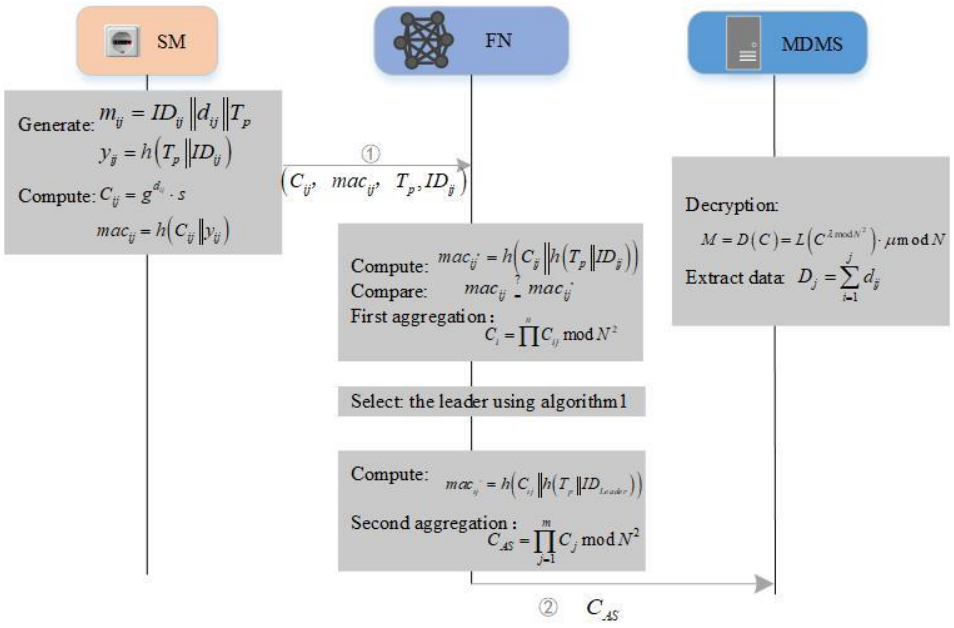


Fig. 2. LDAS-BC working flowchart.

3) After data encryption is completed, SM_{ij} uses hash function h to calculate the digital signature, which can be expressed as Formula 3, where $y_{ij} = h(T_p \| ID_{ij})$.

$$mac_{ij} = h(C_{ij} \| y_{ij}) \tag{3}$$

4) SM_{ij} sends the user report $(C_{ij}, mac_{ij}, T_p, ID_{ij})$ to the fog node.

C. Generation of Fog Chain

This section unfolds across five primary stages, each uniquely handling data in distinct ways. The algorithmic representation of state transitions for each fog node is illustrated in Algorithm 1. During the initial three stages, all fog nodes actively engage, while subsequent to the selection of the primary node, only the fog node designated as the primary node partakes in the final two stages.

1) Data verification: Upon the receipt of a user report, the fog node conducts digital signature computation, incorporating its own identity information. This process can be formally represented as Eq. (4). Subsequently, a comparison is made with the baseline condition, expressed as $mac_{ij} \stackrel{?}{=} mac_{ij}'$. If the equation holds true, the verification process is deemed successful, warranting further progression. Conversely, if the comparison fails to satisfy the condition, the received user report is deemed invalid and subsequently rejected.

$$mac_{ij}' = h(C_{ij} \| h(T_p \| ID_{ij})) \tag{4}$$

2) Initial data aggregation: Within the purview of this stage, every fog node orchestrates the aggregation of ciphertext originating from all smart meters under its jurisdiction. The resultant aggregated ciphertext is delineated by Formula 5. The generation of the corresponding signature for this aggregated ciphertext is elucidated through Formula 6. Where $y_j = h(T_p \| ID_j)$.

$$C_i = \prod^{n} C_{ij} \bmod N^2 \tag{5}$$

$$mac_j = h(C_j \| y_j) \tag{6}$$

3) Primary node election: In an endeavor to curtail the risk of regional data loss and data compromise stemming from single points of failure, and to mitigate the concentration of processing numerous data streams originating from fog nodes via the measurement data management system, the algorithmic methodology outlined in Algorithm 1 is harnessed for primary node selection. Each of the participating fog nodes is equipped with the prospect of ascending to the role of a master node.

4) Subsequent data aggregation: Following the receipt of the aggregation report denoted as (C_j, mac_j, T_p) dispatched by the respective fog node, the master node undertakes a sequence of actions. Initially, the master node initiates the verification process of the aggregation report. Subsequently, it computes the digital signature, as depicted in Formula 7. In contrast to $mac_{ij} \stackrel{?}{=} mac_{ij}'$, if the equation is valid and T_p is within the validity period, then the verification passes. Then, the master node will perform a secondary aggregation of the aggregation report, and the second-level aggregation ciphertext is C_{AS}, as shown in Formula 8.

$$mac_{ij}' = h(C_{ij} \| h(T_p \| ID_{leader})) \tag{7}$$

$$C_{AS} = \prod_{j=1}^{m} C_j \bmod N^2 \tag{8}$$

5) Creation of new block: The master node is tasked with the creation of transaction $T_x = \left(C_{AS}, Leader_{fog_j}, T_p\right)$. In the network model outlined within this section, all fog nodes bear the responsibility of upholding and overseeing the fog blockchain. However, the role of generating blocks is solely entrusted to one and only one master node. The architectural blueprint of this process is succinctly represented in Fig. 3. Upon the master node's encapsulation of the transaction within a block, it disseminates the block across the network, transmitting it to all nodes. Upon receipt of acknowledgments from over a third of the nodes, the block is seamlessly integrated into the longest blockchain, ensuring its enduring persistence and security.

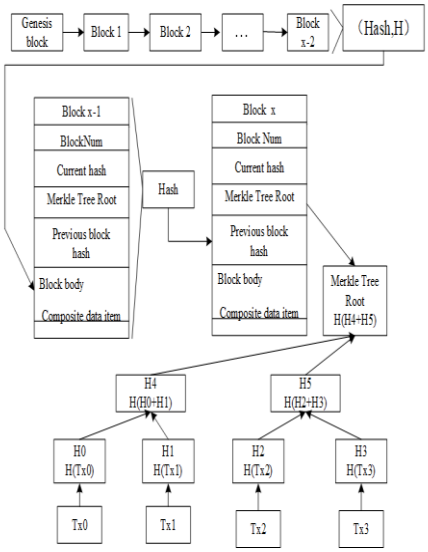


Fig. 3. Block structure and connections.

D. Data Reading and Analysis

Algorithm2 Extraction of Regional Data

Input： M and R

Output： $\left(D_1, D_2, ..., D_n\right)$

1 Set $x_0 = M/R$; $a_1 = R^1, a_2 = R^2, ..., a_m = R^m$; $x_0 = D_1 + R^1 D_2 + ... + R^{m-1} D_m$;

2 For $j = m$; $j > 1$; $j--$ do;

3 $D_j = x_{j-1} \bmod R$;

4 $x_j = x_{j-1}/R$;

5 End for;

6 return $\left(D_1, D_2, ..., D_n\right)$.

The measurement data management system possesses the capability to retrieve information from the fog chain at η minute intervals. Commencing this process, the system initiates the decryption of the second-level aggregate ciphertext through the utilization of the Paillier homomorphic decryption algorithm. To streamline this decryption process for the aggregated ciphertext, certain definitions are introduced:

$$M = a_1 \sum_{i=1}^{n} d_{i1} + a_2 \sum_{i=1}^{n} d_{i2} + ... + a_n \sum_{i=1}^{n} d_{in} \tag{9}$$

$$R = \prod_{j=1}^{n} r_j \tag{10}$$

The ciphertext can then be converted to the form of Formula 11.

$$C = g^M \cdot R^N \bmod N^2 \tag{11}$$

The final aggregated ciphertext still follows the Paillier encryption algorithm, so the measurement data management system can perform Paillier decryption using the private keys λ and $L(\mu)$ to obtain the aggregated plaintext M:

$$M = D(C) = L\left(C^{\lambda \bmod N^2}\right) \cdot \mu \bmod N \tag{12}$$

The ultimate goal of the measurement data management system is to obtain the fine-grained power consumption of each region. In order to achieve this goal, the region data can be obtained through algorithm 2 and $D_1, D_2, ..., D_n$ can be extracted from M:

$$D_j = \sum_{i=1}^{j} d_{ij} \tag{13}$$

VI. Simulation Experiment and Performance Analysis

A. Performance Test

This paper uses Hyperledger Caliper to test the performance of the deployed blockchain network, mainly measuring transaction throughput, transaction latency, and docker container volume. The host hardware parameters used were as follows: The experiment was carried out on an Apple M1 CPU@4*3.2GHz+4*2.064GHz computer.

This paper builds a blockchain network based on Hyperledger fabric v2.4, and the test network consists of two organizations, each with two nodes. A workload named AMI DATA is used to simulate the reading and writing of AMI data, and the performance of the blockchain is evaluated by controlling the number of transactions sent. The test results are shown in Table I.

TABLE I. Performance of Parameters of AMI Data Transaction

AMI data	Maximum delay (s)	Handling capacity (TPS)
1000	1.12	75.4
2000	1.20	79.4
3000	2.05	80.5
4000	1.08	81.6
5000	1.08	98.2

According to the experimental results, with the increase of the number of transactions, the maximum delay basically does not change and remains at about 1S, and the whole system is in

a fairly stable state. The blockchain system can process all the AMI data in a very short time. It can meet the time requirements of smart grid data processing. The throughput changes from 75.4TPS to 98.2TPS, which is a small change and uses less system resources. It can meet the needs of daily power grid equipment. As shown in Fig. 4, the throughput changes with the number of AMI devices. It can be seen that with the increase of AMI devices, the throughput gradually becomes stable. The growth rate is smaller. The changes in throughput are shown in Fig. 4.

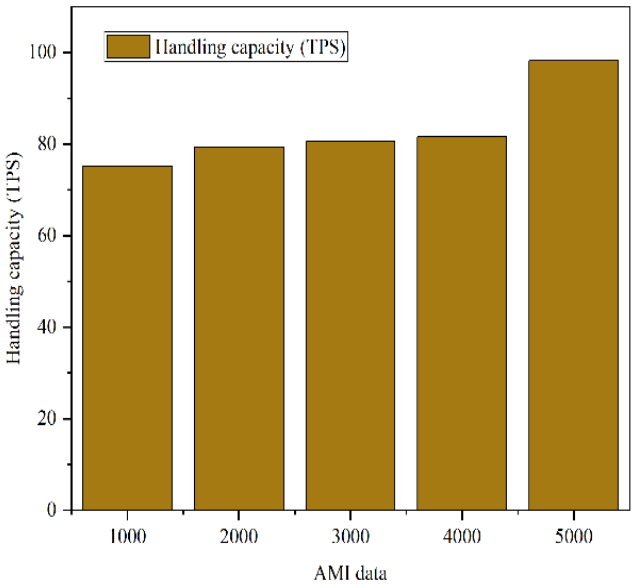


Fig. 4. Changes in throughput.

B. Calculation of Cost

Within this section, a comprehensive analysis of the overall system's computational cost is undertaken. In this analysis, it is posited that the number of fog nodes traverses a spectrum ranging from 5 to 50, which is $m \in [5,50]$. Each of these fog nodes governs jurisdiction over 20 smart meters, which is $n = 20$. For the sake of facilitating a comprehensive performance evaluation of the proposed scheme, a comparative analysis was conducted involving other proposed schemes such as EBDA [20] and LPDA-EC [21]. This comparative assessment aims to provide a robust understanding of the proposed scheme's relative merits and performance attributes in contrast to these alternative approaches. Define T_{E_1} as the exponential operation time in $Z_{n^2}^*$, T_{E_2} as the exponential operation time in G , T_M as the multiplication operation time, and T_P as the pairing operation time. Within this section, the execution of each aforementioned operation is carried out leveraging a pair-based encryption library (PBE). To ensure comparability across experiments, the operation times stipulated in study [13] are adopted. Table II presents the diverse operations alongside their corresponding execution times. Notably, due to the significantly minute time allocation for hash operations relative to other operations, these hash-related procedures are deemed negligible and thus omitted from the computational cost assessment.

TABLE II. TIME TO RUN THE OPERATION

Notations	Descriptions	Time cost （ms）
T_{E_1}	Exponentiation operation in $Z_{n^2}^*$	1.60
T_{E_2}	Exponentiation operation in G	1.62
T_M	Multiplication operation	0.06
T_P	Pairing operation	17.70

At the user level, computing ciphertext C_{ij} requires $2mn$ multiplication T_M . In the fog layer, each fog node performs the first data aggregation, then the entire fog layer requires a total of mn times multiplication operation, that is, mn times T_M . The master node performs a second data aggregation, requiring the m multiplication operation T_M . When the measurement data management system in the cloud needs to read the contents of the fog chain, the operation of Formula 12 is performed, which requires a power operation T_{E_1} and a multiplication operation T_M . In summary, the total calculation cost of the proposed scheme is $(3mn+m+1)T_M + T_{E_1}$. Table III shows a comparison of calculated costs.

TABLE III. COMPARES THE CALCULATED COSTS

Option	Calculate the cost
LPDAEC	$m\left[2nT_{E_1} + 4nT_M + (3n+1)T_{E_2}\right] + 2mT_P + 2mT_{E_1}$
EBDA	$m\left[nT_{E_1} + 2(n+1)T_M\right] + (m+1)T_P + (m+1)T_M + T_{E_1}$
LDASBC	$(3mn+m+1)T_M + T_{E_1}$

Fig. 5 provides a comparative illustration of the cumulative computational costs associated with the three schemes under examination. Evidently, the proposed scheme presented in this paper markedly exhibits significantly lower computational costs in comparison to the other two schemes. This advantageous performance differential becomes even more pronounced with the escalation in the number of smart meters within the system. By comparing the graphs, it can be concluded that LDAS-BC can significantly reduce the computational cost of AMI data and has strong scalability.

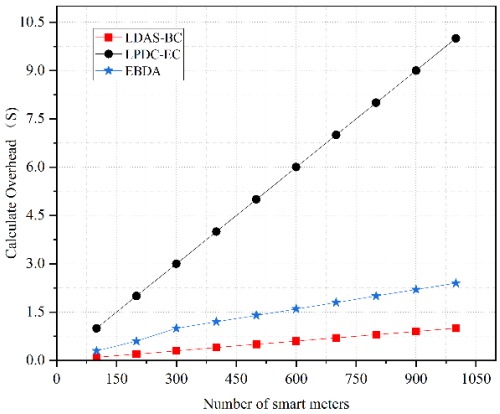


Fig. 5. Calculated cost comparison.

C. Communication Cost

The communication cost within the proposed scheme encompasses two primary components: the communication between the smart meter and the fog node, and the subsequent communication between the fog node and the master node. First, the smart meter generates a user report $\left(C_{ij}, mac_{ij}, T_p, ID_{ij}\right)$ and sends it to the fog node, assuming that $\left|N^2\right|$ is 2048*bit*, $|p|$ is 160*bit*, $|ID|$ and $\left|T_p\right|$ are 160*bit*, then $S_{user\to fog} = \left|C_{ij}\right| + \left|mac_{ij}\right| + \left|T_p\right| = 2368bit$. The fog node then sends $\left(C_{ij}, mac_{ij}, T_p, ID_{ij}\right)$ to the master node, which has $S_{fog\to Leader} = \left|C_{ij}\right| + \left|mac_{ij}\right| + \left|T_p\right| = 2368bit$. As a consequence of the aggregation operation, the aggregate communication overhead remains unaffected by the number of smart grids. This intrinsic property signifies that even as the quantity of smart grids escalates, the communication overhead between the fog node and the master node remains consistent. Table IV tabulates the communication costs associated with the three alternatives.

TABLE IV. COMPARISON OF COMMUNICATION COSTS

Option	Communication Cost
LPDA-EC	$mn\left(\left\|C_{ij}\right\|+160+\left\|T_p\right\|+160\right)+(m-1)\left(\left\|C_j\right\|+\left\|ID_j\right\|+\left\|\sigma_j\right\|+T_p\right)$
EBDA	$mn\left(\left\|C_{ij}\right\|+\left\|h_{ij-s}\right\|+\left\|mac_{ijs}\right\|\right)+(m-1)\left(\left\|C_j\right\|+\left\|ID_j\right\|+\left\|\sigma_j\right\|\right)$
LDAS-BC	$mn * S_{user\to fog}+(m-1) * S_{user\to Leader}$

Fig. 6 presents a comparative analysis of the communication costs inherent in the three schemes. The visual depiction highlights that the proposed scheme boasts a marginal advantage over the other two alternatives in terms of communication costs. This advantage is poised to intensify as the Advanced Metering Infrastructure (AMI) scale expands during subsequent stages. Through the comparison of the figures, it can be concluded that LDAS-BC can significantly reduce the cost of AMI data transmission, and after secondary aggregation, the effect is more obvious.

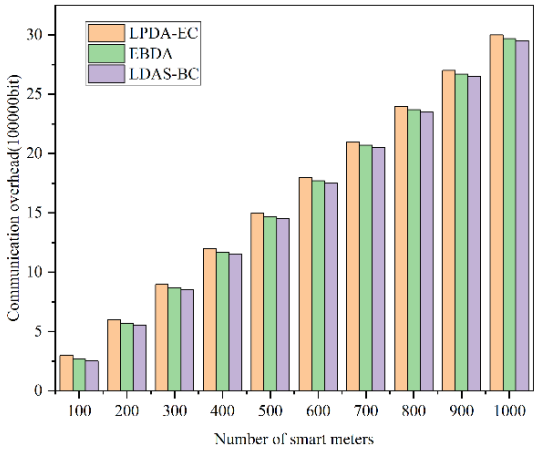


Fig. 6. Calculated cost comparison.

To sum up, compared with existing schemes, LDAS-BC has great advantages in terms of data security, data privacy, data aggregation, as well as computational costs and communication costs. LDAS-BC is a lightweight, low-risk secondary aggregation scheme for blockchain data.

VII. CONCLUSIONS

With the objective of enhancing the data transmission efficiency within the context of advanced measurement systems, as well as curtailing computational and communication overhead, this study introduces the LDAS-BC data aggregation scheme predicated on blockchain technology. This scheme ingeniously amalgamates the Paillier homomorphic encryption algorithm to achieve the aggregation of ciphertext, significantly mitigating the risk of breaching user privacy data. Furthermore, the scheme incorporates the tenets of fog computing and cloud computing, effectively addressing issues tied to computational limitations and restricted storage resources associated with network edge devices. The integration of blockchain technology further reinforces the security and reliability of on-chain data. The proposed LDAS-BC scheme's efficacy is substantiated through theoretical analysis and experimental simulations, which collectively validate its safety and dependability. Comparative experiments additionally underscore the scheme's heightened advantages in comparison to alternative approaches. In the future work, we should consider the selection of consensus algorithm, the mutual authentication among different levels, and the calculation cost and communication cost of these problems need to be calculated. Finally, it is envisaged to extend this scheme to the scenario of Internet of Things data. This will become the focus of the later research work.

REFERENCES

[1] I. Colak, R. Bayindir and S. Sagiroglu, “The Effects of the Smart Grid System on the National Grids,” in 2020 8th International Conference on Smart Grid (icSmartGrid), Paris, France, 2020, doi: 10.1109/icSmartGrid49881.2020.9144891.

[2] K. Sha, N. Alatrash and Z. Wang, "A Secure and Efficient Framework to Read Isolated Smart Grid Devices," IEEE Transactions on Smart Grid, vol. 8, no.6, pp. 2519-2531, 2017, doi: 10.1109/TSG.2016.2526045.

[3] Y.Y. Sun, J.J. Yuan and M.Y. Zhai, “Cloud-Based Data Analysis of User Side in Smart Grid,” in 2016 2nd International Conference on Open and Big Data (OBD), Vienna, Austria, 2016, doi: 10.1109/OBD.2016.13.

[4] P. Zhuang, T. Zamir and H. Liang, "Blockchain for Cybersecurity in Smart Grid: A Comprehensive Survey," IEEE Transactions on Industrial Informatics, vol. 17, no.1, pp. 3-19, 2021, doi: 10.1109/TII.2020.2998479.

[5] H. Tian, Y. Jian and X. Ge, "Blockchain-based AMI framework for data security and privacy protection, " Sustainable Energy, Grids and Networks, vol. 32, pp. 100807, 2022, doi: 10.1016/j.segan.2022.100807.

[6] C. F. Jiang, T. T. Fan, H. H. Gao, W. S. Shi, L. K. Liu et al., "Energy aware edge computing: A survey, " Computer Communications, vol. 151, pp. 556-580, 2020, doi: 10.1016/j.comcom.2020.01.004.

[7] S. L. Chen, H. Wen, J. S. Wu, W. X. Lei, W. J. Hou et al., "Internet of Things Based Smart Grids Supported by Intelligent Edge Computing, " IEEE ACCESS, vol. 7, pp. 74089-74102, 2019, doi: 10.1109/ACCESS.2019.2920488.

[8] L. Chen, R. X. Lu, Z. F. Cao, K. AlHarbi and X. D. Lin, "MuDA: Multifunctional data aggregation in privacy-preserving smart grid communications, " PEER-TO-PEER NETWORKING AND APPLICATIONS, vol. 8, no.5, pp. 777-792, 2015, doi: 10.1007/s12083-014-0292-0.

[9] X. H. Liang, X. Li, R. X. Lu, X. D. Lin and X. M. Shen, "UDP: Usage-Based Dynamic Pricing With Privacy Preservation for Smart Grid, " IEEE TRANSACTIONS ON SMART GRID, vol. 4, no.1, pp. 141-150, 2013, doi: 10.1109/TSG.2012.2228240.

[10] P. Gope and B. Sikdar, "An Efficient Privacy-Friendly Hop-by-Hop Data Aggregation Scheme for Smart Grids" IEEE SYSTEMS JOURNAL, vol. 14, no.1, pp. 343-352, 2020, doi: 10.1109/JSYST.2019.2899986.

[11] P. Singh, M. Masud, M. S. Hossain and A. Kaur, "Blockchain and homomorphic encryption-based privacy-preserving data aggregation model in smart grid, " Computers & Electrical Engineering, vol. 93, pp. 107209, 2021, doi: 10.1016/j.compeleceng.2021.107209.

[12] Z. T. Guan, G. L. Si, X. S. Zhang, L. F. Wu, N. Guizani et al., "Privacy-Preserving and Efficient Aggregation Based on Blockchain for Power Grid Communications in Smart Communities, " IEEE COMMUNICATIONS MAGAZINE, vol. 56, no.7, pp. 82-88, 2018, doi: 10.1109/MCOM.2018.1700401.

[13] S. G. Chen, L. Yang, C. X. Zhao, V. Varadarajan and K. Wang, "Double-Blockchain Assisted Secure and Anonymous Data Aggregation for Fog-Enabled Smart Grid, " Engineering, vol. 8, pp. 159-169, 2022, doi: 10.1016/j.eng.2020.06.018.

[14] F. Loukil, G. G. Chirine, K. Boukadi and A. N. Benharkat, "Privacy-Preserving IoT Data Aggregation Based on Blockchain and Homomorphic Encryption, " Sensors, vol. 21, no. 7, pp. 2452, 2021, doi: 10.3390/s21072452.

[15] C. Regueiro, I. Seco, S. de Diego, O. Lage and L. Etxebarria, "Privacy-enhancing distributed protocol for data aggregation based on blockchain and homomorphic encryption, " Information Processing & Management, vol. 58, no. 6, pp. 102745, 2021, doi: 10.1016/j.ipm.2021.102745.

[16] H. Y. Bao, B. B. Ren, B. B. Li and Q. L. Kong, "BBNP: A Blockchain-Based Novel Paradigm for Fair and Secure Smart Grid Communications, " IEEE INTERNET OF THINGS JOURNAL, vol. 9, no.15, pp. 12984-12996, 2022, doi: 10.1109/JIOT.2021.3107301.

[17] X. H. Zhang, L. You, and G. R. Hu, "An Efficient and Robust Multidimensional Data Aggregation Scheme for Smart Grid Based on Blockchain, " IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, vol. 19 no.4, pp. 3949-3959, 2022, doi: 10.1109/TNSM.2022.3217312.

[18] M. Zhao, Y. Ding, S. J. Tang, H. Liang and H. Y. Wang, "A blockchain-based framework for privacy-preserving and verifiable billing in smart grid, " PEER-TO-PEER NETWORKING AND APPLICATIONS, vol. 16, no.1, pp. 142-155, 2022, doi: 10.1007/s12083-022-01379-4.

[19] R. Y. Yu, A. M. Oguti, D. R. Ochora and S. C. Li, "Towards a privacy-preserving smart contract-based data aggregation and quality-driven incentive mechanism for mobile crowdsensing," Journal of Network and Computer Applications, vol. 207, pp. 103483, 2022, doi: 10.1016/j.jnca.2022.103483.

[20] W. Lu, Z. Ren, J. Xu and S. Chen, "Edge Blockchain Assisted Lightweight Privacy-Preserving Data Aggregation for Smart Grid, " IEEE Transactions on Network and Service Management, vol. 18, no.2, pp. 1246-1259, 2021, doi: 10.1109/TNSM.2020.3048822.

[21] J. Zhang, Y. Zhao, J. Wu and B. Chen, “LPDA-EC: A Lightweight Privacy-Preserving Data Aggregation Scheme for Edge Computing,” in 2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), Chengdu, China, 2018, doi: 10.1109/MASS.2018.00024.

[22] H. B. Fan, Y. N. Liu and Z. X. Zeng, "Decentralized Privacy-Preserving Data Aggregation Scheme for Smart Grid Based on Blockchain," SENSORS, vol. 20, no.18, pp. 5282, 2020, doi: 10.3390/s20185282.

Predicting and Improving Behavioural Factors that Boosts Learning Abilities in Post-Pandemic Times using AI Techniques

Dr. Jaya Gera[1], Dr. Ekta Bhambri Marwaha[2], Dr. Reema Thareja[3], Dr. Aruna Jain[4]
Associate Professor, Department of Computer Science, Shyama Prasad Mukherji College, University of Delhi, India[1]
Associate Professor, Department of Applied Psychology, Shyama Prasad Mukherji College, University of Delhi, India[2]
Assistant Professor, Department of Computer Science, Shyama Prasad Mukherji College, University of Delhi, India[3]
Associate Professor, Department of Computer Science, Bharati College, University of Delhi, India[4]

***Abstract*—Quantifying student academic performance has always been challenging as it hinges on several factors including academic progress, personal characteristics and behaviours relating to learning activities. Several research studies are therefore being conducted to identify the factors so that appropriate measures can be conducted by academic institutions, family and the student to boost his/ her academic performance. The present study investigates personal characteristics, psychological factors, behavioural factors, social factors and learning capabilities, that directly or indirectly affect student's academic performance, which was tapped by administering a self-designed questionnaire. The data was collected from 214 undergraduate students studying in various streams of the University of Delhi and post that semi-structured interview was conducted to get in- depth information. The result proved the correlation between the aforementioned factors and the learning capabilities of the students. Using the results of analysis a machine learning model based on k-nn algorithm was formed to predict student performance. A chatbot is also proposed to provide guidance to students in strenuous situations, motivate them and interact with them without having personal bias.**

Keywords*—Academic performance; machine learning; chatbot; educational data mining; learning analytics

I. Introduction

In today's era of the information revolution, learning analytics, predictive analytics, educational data mining, and machine learning techniques has become a hot area of research [1, 2] as it is beneficial for Teachers, administrators, family and the student him/herself to provide a timely remedy. We found that students have frequent mood swings, internal conflicts and issues that keep them distracted. They are often not comfortable to share their problems with their family members and friends. Accumulation of these issues causes stress and anxiety. In such a situation students become pessimistic, spend more time on social media or get involved in substance use. It is therefore important to identify academic and non-academic parameters that affects a student's performance. Using these parameters, we need a ML model to predict a new student's performance so that teachers and mentors can address their concerns at the earliest and provide them every help to grow with a stable mind.

Supervised learning algorithms are being used to analyse student's learning behaviours in order to predict and classify the students' performance [13,17]. However, the challenging task is to find the optimal algorithm that gives best results. Machine learning algorithms like k-nn, Naïve Bayes, ANN, logistic regression, SVM, decision tree, random forest, etc are used for prediction but the accuracy of results obtained from each model depends on the size and quality of data [11,12].

In this paper, we have used data analytics to identify personal, behavioural and academic factors that affect student's learning capability and empirically proved that there exists a relationship between the identified factors and the learning capability. Then, an AI model is created which when fed with identified factors as input, predicts the how quick learner a student is using the k-nn algorithm. Moreover, to handle the mental health issues which in turn affects student's academic performance as well as their learning capability, an NLP based chatbot model has been proposed to interact and guide them especially in strenuous situations.

The paper is divided into six sections. The first section introduces the title of the paper. The second section presents a summarization of research already done in this area. In section, data science and data analytics have been used to mine information from the dataset and visually represent crucial numbers using tables and charts. The results of analysis are interpreted and explained in Section IV. The section also proposes an AI based solution to find a solution to the issues identified in Sections III and IV. Section V proposes an AI based machine learning model that can accurately predict a new student's future performance by analysing certain parameters and comparing them with data already stored in the database. The paper is finally concluded in Section VI.

II. Literature Review

The drop-out students' ratio from higher education institutions results in immense loss/resource wastage. It also affects the evaluation and assessment processes of these institutions. New technologies like data mining, machine learning should therefore be used to perform simple and effective analysis of student-performance data that could help to improve learning procedures and atmosphere [3,4].

In [5], researchers have explained the use of Bigdata methods for learning analysis that can be used for system performance prediction, visualization of data, student skills estimation, risk detection, course recommendation, grouping of students and collaboration with other students. Predictive analysis is done on student achievements, behaviour, and skill prediction.

During the forecast of student performance [6] data-mining techniques were used to build a predictor framework for the final-marks based on students' achievements and features including features, non-courses variables, out of class student conduct, video watching and post-school tutoring.

Researchers in [7] have evaluated students at the beginning of an academic session to forecast their achievements using collaborative filtering technique that is based on their academic history. Going further, in [8], the author has used historical academic data of students to evaluate a students' performance. The study relied on a factorization of low-range-matrixes and dispersed linear model.

In [9], a student classification system was developed using naive Bayes and decision tree algorithms that used features like the occupation of parents to enhance the correctness of the grade-prediction framework. The Decision Tree classifier performed better than the naive Bayes Classification in terms of accuracy [10].

III. Data Analysis

In the first stage, the raw datasets are collected from 214 students studying different courses in the University of Delhi. The data was then pre-processed to transform categorical data into numerical. Data analytics techniques were then applied using programming languages R and Python. The observations are reported in Table I.

TABLE I. Queries and Corresponding Responses

Query	Response	Query	Response
When you listen to a spiritual leader / motivational speaker, what are your topics of interest?	36.4% for Handling Relationships 66.3% for Enhancing Concentration 57.3% for overcoming anxiety 72.4% for staying motivated 12% do not listen to any spiritual leader	How many hours do you spend on social media per day?	13.1% spend less than 1 hour daily 31.8% spend 1 to 2 hours daily 35.5% spend 2 to 3 hours daily 19.6% spend more than 3 hours daily
How many close friends do you have?	11.7% have no close friend 65.4% have 1 to 3 close friends 15.9% have 4 to 5 close friends 7% have more than 5 close friends	How frequently do you meet your friends in a week?	8.4% meet once in a week 5.1% meet twice in a week 6.1% meet thrice in a week 26.2% meet more than three times 54.2% meet them occasionally
Do you discuss your personal issues with your close friends?	50.5% says certain issues not all 35% feel free to discuss 14.5% do not share	How frequently do you have mood swings?	36.4% experience mood swings multiple times in a day 15.4% once in a day 13.6% once in a week 28.5% once every fortnight Rest occasionally
What is the state of your mind in general?	43.3% varies frequently 24.3% calm 16.8% happy 14.9% highly frustrated and disappointed	How many hours do you devote every day for self-study?	16.8% spend less than 1 hour daily 36.9% spend 1 to 2 hours daily 29% spend 2 to 3 hours daily 12.6% spend 3 to 4 hours daily 4.7% study for more than 4 hours daily
Which method of study do you prefer the most?	59.3% Reading Books or any other sort of Reading Material 29.9% Watching Videos 7.5% Group Studies with Friends 2.3% Listening Audio Podcasts Rest Educational Programs on TV	Who is your role model?	43% Family Member 42.5% have no role model 4.7% sees their teacher as their role model 6.5% sees Sportsperson / army / any eminent person from history as their role model 3.3% sees Actor / Actress as their role model
How do you find your classes?	15.9% find them Boring and unproductive 50% Not boring but putting extra burden 34.1 %Exciting and triggering my thought process	Has your learning capabilities been affected due to the pandemic?	41.6% Badly affected 51.4% Slightly affected 3.3% Not at all affected 3.7% Have become better than before
Were you comfortable with online teaching?	29.9% Yes 41.6% No Rest was unsure	Which mode of teaching do you prefer now?	10.7% Online 47.7% Offline 41.6% Hybrid
Did your learning capabilities improve with digital technology?	53.7% Yes 20.6% No 25.7% are unsure	Did any of the cases given below affect your learning capabilities in the last 2 years?	71.4% were affected due to Stress and Anxiety 21.9% were affected due to Death in the Family 44.8% were affected due to Strained Financial Conditions 63.08% were affected due to Problems in Relationships 72.4% were affected due to Increased exposure to Electronic Gadgets

			18.69% were affected due to All of the above
During pandemic, your studies were affected the most due to which of the following reasons?	46.77% Lack of personal space 20.56% Inadequate ventilation 25.7% Power Cut 54.7% Noise from surroundings 62.14% Technical Issues 77.17% Disturbed routine 42.05% Unhealthy eating habit	Which activity do you practice to keep your mind stable?	27.57% practice Yoga 39.71% practice Meditation 58.4% practice Singing 54.2% practice Dancing 31.7% get involved in Sports Activities
Do you want any type of counselling to handle your stress?	57.5% Yes Rest No	What do you do to burst your stress?	64.4% prefer Talking to someone who is close to me 65.8% prefer Keeping Quiet and Sitting Alone 35.9% prefer Reading Shopping 33.17% prefer Eating too much 68.2% prefer Watching TV or Web Series 54.2% prefer Walking or any Physical Activity
As a student, when was your level of motivation and attention higher?	49.5% Pre-pandemic times 7.5% During pandemic times 14.5% post-pandemic times Rest opted for Never 24.3% Always the same	Do you feel that your memorizing capabilities have been adversely affected due to COVID-19?	52.8% Yes 24.3% No Rest believes Maybe
During online classes, was your learning affected due to any of the following reasons?	28.03% Classes were not conducted on online platform regularly 61.68% Recording of lecturers were not available 50% Sessions were not interactive 70.09% Could not interact with other students in the class 51.8% Could not attend due to poor net connection 42.98% Did not have enough devices to attend classes 35.98% Study material was not available 80.8% Difficult to concentrate online beyond a certain limit	Are you facing self-regulation challenges?	40.65% Often miss the deadlines given by teachers 52.33% Did not get appropriate help during online classes 48.5% Lack the ability to control my own thoughts, emotions, and actions during classes 63.08% Have limited preparation before a class. 60.28% have poor time management skills during online classes. 55.14% fail to properly use online peer learning strategies (i.e., learning from one another to better
Do you face Technological literacy and competency challenges?	27.10% Lack competence and proficiency in using various interfaces or systems 21.49% resist learning technology. 30.37% are distracted by an overly complex technology. 19.6%I have difficulties in learning a new technology. 22.42% lack the ability to effectively use technology to facilitate learning. 23.83% lack knowledge and training in the use of technology. 20.09% are intimidated by the technologies used for learning. 28.5% resist and/or am confused when getting appropriate help during online classes. 30.84% have poor understanding of directions and expectations during online learning. 29.43% perceive technology as a barrier to getting help from others during online classes		

IV. Data Interpretation

When we analysed the data collected, we observed an alarming situation. 90% of the students admitted that their learning capabilities have been affected due to the pandemic. Out of these, 71.4% of the students were affected due to Stress and Anxiety and 41.6% of these students were never comfortable with online teaching.

49.5% students believe that their level of motivation and attention was higher in pre-pandemic times and 52.8% feel that their memorizing capabilities have been adversely affected due to COVID-19.

During online classes, learning abilities of the students were affected as their classes were not conducted on online platform regularly (28.03%), recorded lecturers were not available (61.68%), teaching sessions were not interactive (50%), no interactions with peer group (70.09%), poor net connection (51.8%) and lack of concentration (80.08%). In fact, every student experienced Student isolation challenges (SIC). Due to this, students feel emotionally disconnected or isolated and uncomfortable during online classes. They preferred face-to-face interaction with teachers and traditional classroom methods for socialization [11]. This is evident from the graph shown in Fig. 1.

During pandemic, studies were affected majorly due to lack of personal space (46.77%), noise from surroundings (54.7%), technical Issues (62.14%), disturbed routine (77.17%) and unhealthy eating habits (42.05%). In fact, the students admitted that during exams, they experienced anxiety issues when appearing for interviews or viva. So, they preferred to have offline teaching for better outcomes (refer Fig. 2).

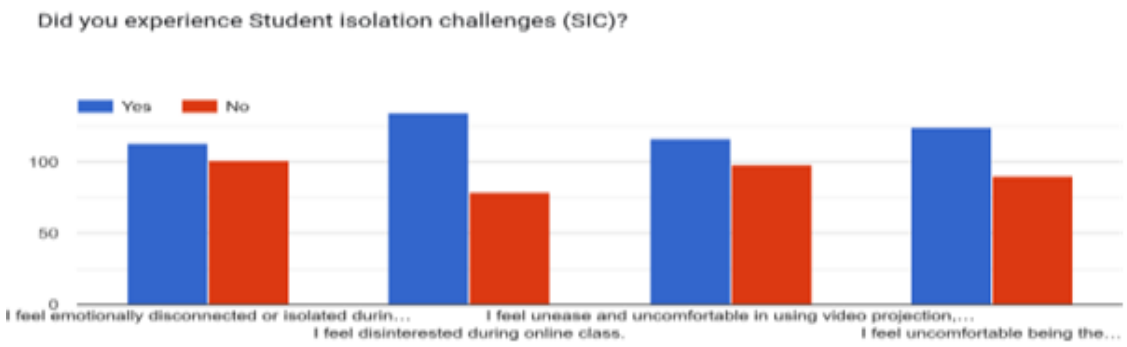


Fig. 1. SIC issues.

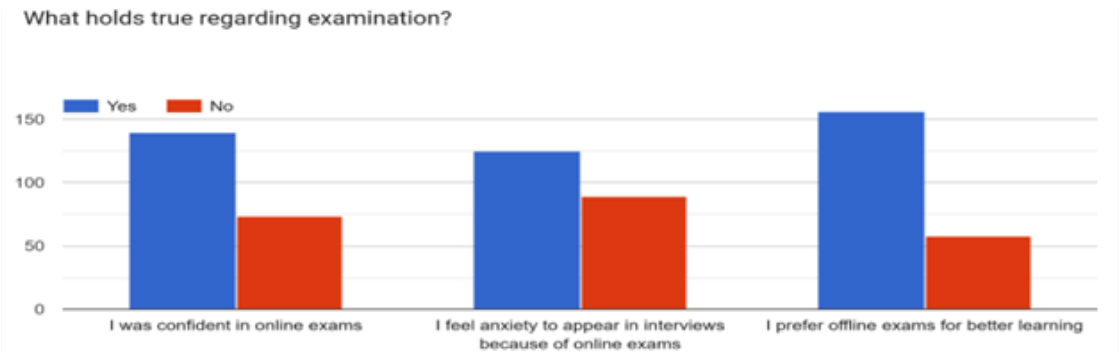


Fig. 2. Mindset regarding examination.

40.65% of the students feel disappointed as they often miss the deadlines given by teachers and 48.5% find it difficult to control their thoughts, emotions, and actions during classes. However, after the colleges reopened post COVID, the situation has become little better. More than 76% of the students agreed that a personal eye contact with the teacher is required for effective learning [12] (refer Fig. 3).

In fact, after colleges started in offline, only 8.9% of the students stated that small intra-class as well as inter-class competitions will not help to augment effective learning.

A. *Stress and Anxiety Hampers Success*

The students accepted that due to stress and anxiety they are unable to perform to their maximum potential. 51.8% of students experience mood swings at least once in a day [13]. 43.3% believes that their state of mind is not calm and happy. Rather, it varies frequently. The situation is worsened as 14.9% of students accept that they are highly frustrated and disappointed.

42.5% of students do not have a role model. This may not be a good sign for their sound mental health. 50% of students feel that their classes add an extra burden to their daily lives. 57.8% students admitted that they need counselling to handle their issues. When asked additional questions related to their mental health, students' responses were recorded as shown in Fig. 4.

To handle such a strenuous situation, students are taking measures to overcome factors affecting their studies, which are in line with other research studies. They are listening to spiritual and motivational speakers for enhancing concentration and overcoming stress as well as anxiety. As evident from the pie chart shown in Fig. 5, more than 75% of the students have expressed that they listen to a spiritual/ motivational speaker at least once in a week. 58.4% practice Singing and 54.2% of the students practice Dancing to keep stress at bay. In extreme strenuous conditions, 64.4% prefer talking to a close friend/relation, 65.8% prefer keeping quiet and sitting alone, 68.2% watch TV or Web Series and 54.2% prefer walking or any other physical activity, which helped improve the mood and reduce stress among students [14].

Fig. 3. Importance of teachers and competitions in effective learning.

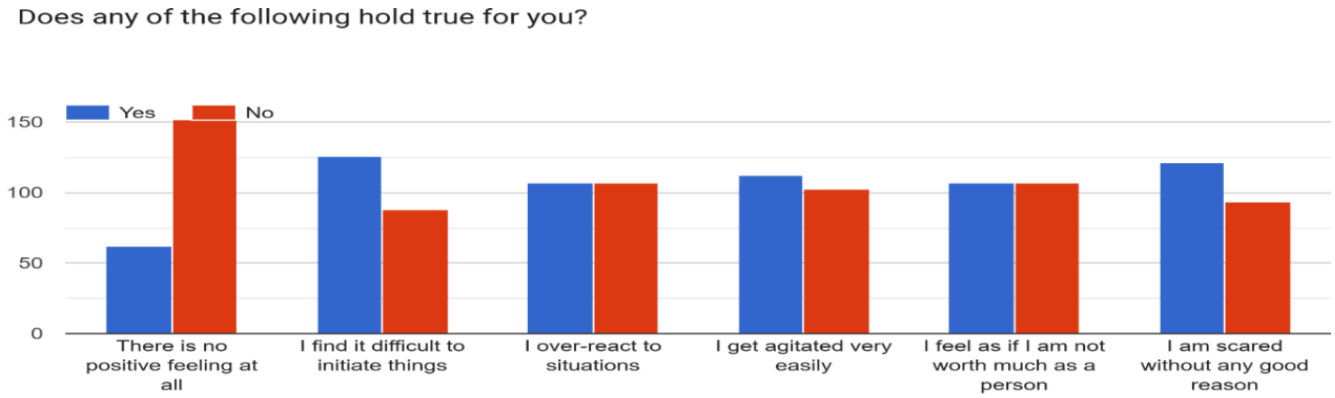


Fig. 4. Analysing mental health of students in post-pandemic times.

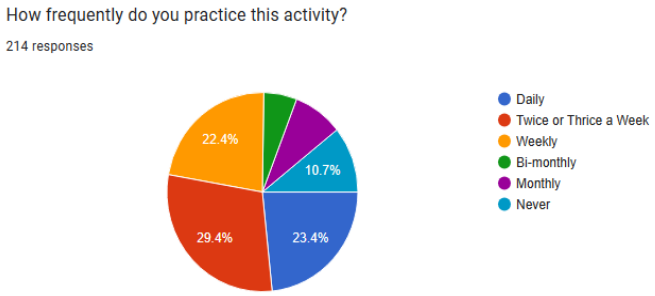


Fig. 5. Frequency of listening to spiritual / motivational speakers.

B. Solution Proposed to Boost Learning Capabilities

In the study, it was observed that 65.4% of the students have 1 to 3 close friends but 54.2% of them meet them occasionally. Moreover, 65% of the students admitted that they feel uncomfortable to discuss issues that disturb them with their close friends.

Moreover, in the post-pandemic times, more than 56% of the students are facing issues due to fatigue. And only 37% of the students are finding their studies interesting. Rest feels overwhelmed due to stress and thus feel fatigued after studying for few minutes.

To help students handle their stress and anxiety issues, we have proposed a chatbot based on the model depicted in Fig. 6. The role of this chatbot is to provide guidance to students in strenuous situations, motivate them and interact with them without having personal bias. Our study reveals that students are not comfortable discussing their personal issues even with their close friends. In such a scenario, they can at least share their feelings with an AI assisted chatbot that can communicate like a human and provide counselling or guidance.

C. Data Exploration to Build the Model

Before designing the machine learning model to make predictions, we first need to explore data to understand the relationships that exists between different parameters [15]. Therefore, we calculated co-relation between some key variables. The co-relation values between two variables are given in Fig. 7. The same values when plotted using a heatmap can be visualized.

From the heat map plotted, it is very clearly evident that there is a strong relationship between Hours_of_Self_Study and Optimistic, Consistent_Learner and Quick_Learner, Stable_Mind and Optimistic, Stable_Mind and Hours_of_Self_Study. To study their underlying relationship in depth, we have plotted bar graphs. Identifying relationships between the variables and their impact on being quick and consistent learner is important, as a successful student has to be quick to learn and learn consistently. The conclusion that a consistent learner is quick to learn and vice versa has been drawn from the heatmap.

We observed that students who believe in group studies with friends are not quick and consistent learners. Quick and consistent learners prefer to study from books or any sort of reading material. Few of them also prefer to learn by watching educational videos on the World Wide Web. A quick and consistent learner spends at least 1 to 3 hours daily (refer Fig. 8) and spends majority of the time reading books/ study material followed by watching educational videos (refer Fig. 9). Even in the digital age, students prefer to read hard bound books or printed text material. They watch videos only for certain topics that they find too hard to understand. Moreover, in a country like India students find difficult to understand terminologies of hard-core English language, so they watch videos in local languages to understand as well as validate their understanding. One more interesting thing that we concluded from our studies is that negligible number of students prefers watching educational programs on Television and listen to podcast for effective learning.

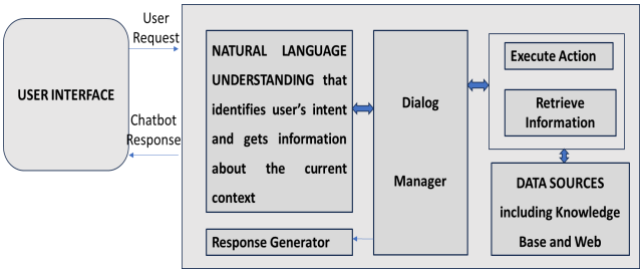


Fig. 6. Proposed chatbot model.

	Hours_SelfStudy	Method_Study	Optimistic	ConsistentLearner	Stable_mind	Quick_Learner
Hours_SelfStudy	1.000000	-0.039559	0.174594	-0.023668	0.098585	-0.015722
Method_Study	-0.039559	1.000000	-0.023559	-0.005205	-0.068424	0.014068
Optimistic	0.174594	-0.023559	1.000000	-0.043759	0.089492	-0.101285
ConsistentLearner	-0.023668	-0.005205	-0.043759	1.000000	0.019614	0.124775
Stable_mind	0.098585	-0.068424	0.089492	0.019614	1.000000	0.022564
Quick_Learner	-0.015722	0.014068	-0.101285	0.124775	0.022564	1.000000

Fig. 7. Correlation matrix and heatmap.

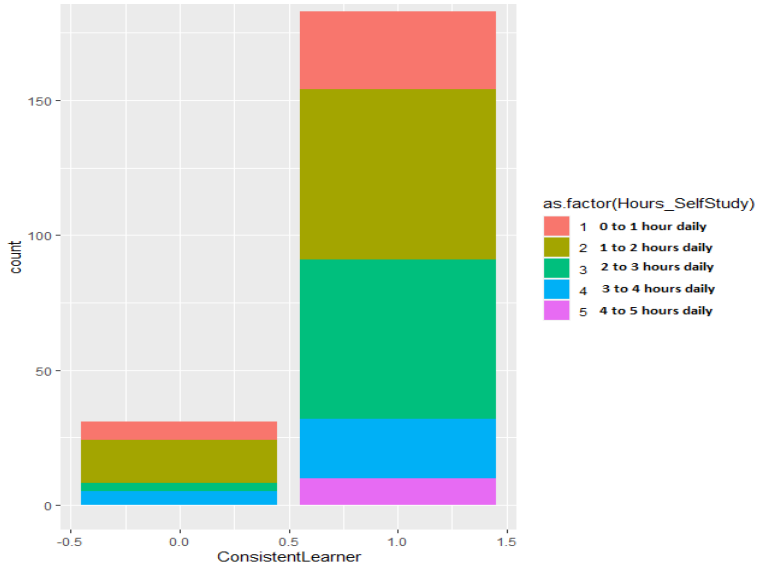


Fig. 8. Relationship between consistent learner and hours of self-study.

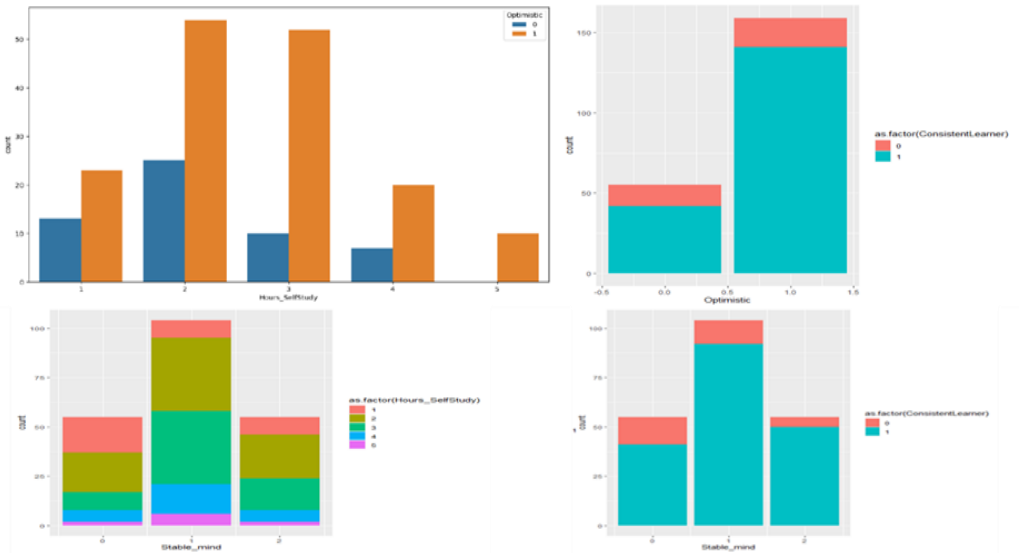


Fig. 9. Relationship between method of study and hours of self-study.

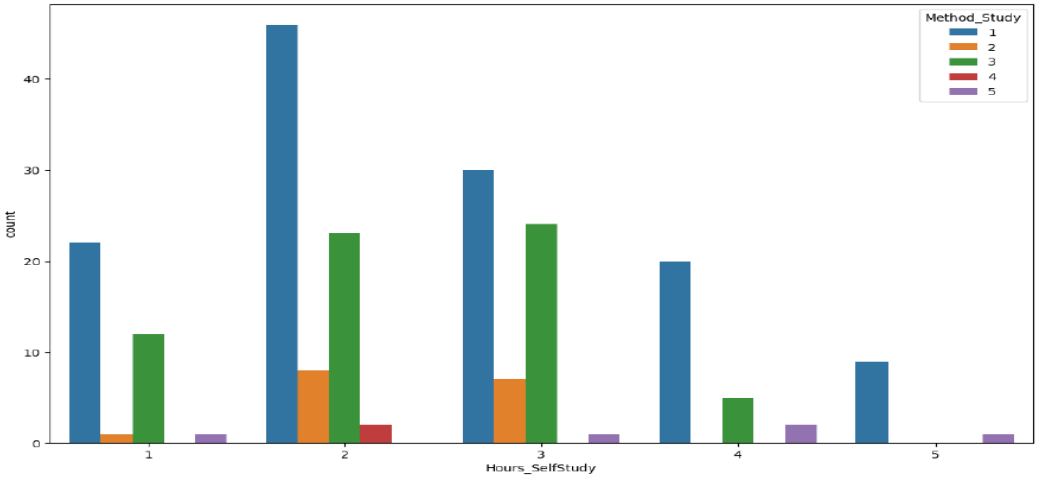


Fig. 10. Relationship between stable mind, optimistic, consistent learner and hours of self-study.

For effective learning, a quick and consistent learner devotes 1 to 3 hours daily. Studying beyond 3 hours does not add to their learning productivity. Research has indicated a favourable correlation between study duration and scholastic achievement [16]. A specific amount of study time will be reached before the grade no longer demonstrates a discernible shift. A quick and consistent learner experience less stress and anxiety and therefore has a stable mindset and is usually optimistic. Optimistic student prefers self-study using books or study material and is not dependent on peer group for learning. These numbers are evident from graphs given in Fig. 10.

V. Proposed ML Model to Predict Learning Capabilities

Using these conclusions, we can create a machine learning model (refer Fig. 11). The model divides the entire data set into two groups- training dataset and testing dataset. While training dataset has 70% of the data records chosen randomly from the original dataset, testing data set on the other hand has rest of the 30% records. The training dataset is utilized in the learning phase of the machine learning model. Correspondingly, the testing dataset is used in the evaluation phase to predict values for the 30% rows in the dataset. The ML model classifies a student either as a Quick Learner or a Slow Learner. Values predicted are then compared with actual values. The number of correctly predicted values are then compared with those predicted wrongly to determine the accuracy of the model. Higher the accuracy, higher is the probability that a new student will be accurately classified being a quick and consistent learner.

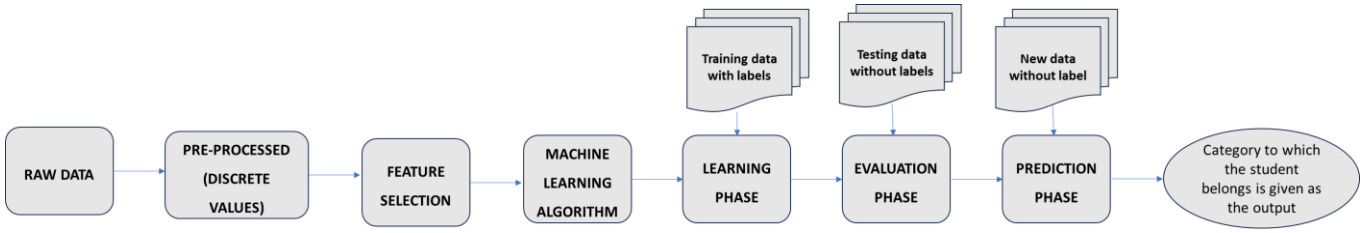


Fig. 11. Proposed ML model.

An accurate machine learning model will help the teachers to focus on inconsistent and slow learners and at the same time plan certain activities or assignments to further polish and accelerate the learning of quick and consistent learners.

The proposed machine learning model uses K Nearest Neighbour Classification algorithm. We also designed ML model using Naïve Bayes algorithm but accuracy of predictions was higher in case of knn algorithm [17]. For better and efficient results, we had done hyper parameter tuning. The confusion matrix of the k-nn algorithm is given in Fig. 12.

Accuracy (all correct / all) = (TP + TN) / (TP + TN + FP + FN).

Accuracy = (26 + 23) / (26 + 23 + 11 + 3) = 49 / 63 = 0.7777.

Misclassification (all incorrect / all) = (FP + FN) / (TP + TN + FP + FN).

Misclassification = 14 / 63 = 0.2222

Precision (true positives / predicted positives) = TP / (TP + FP)Precision = 26 / 37 = 0.7027.

Sensitivity aka Recall (true positives / all actual positives) = TP / (TP + FN).

Recall = 26 / 29 = 0.8965

Specificity (true negatives / all actual negatives) =TN / (TN + FP).

Specificity = 23 / 34 = 0.6764

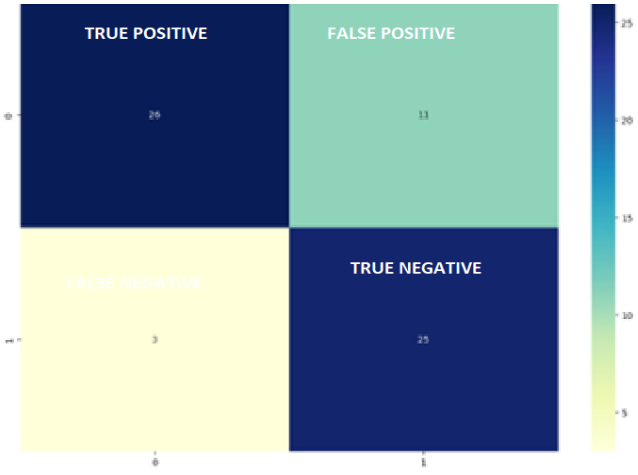


Fig. 12. Confusion matrix.

VI. Conclusion and Future Scope

The paper presents a study on 214 students in Higher Education. Several questions were asked to get an insight into their learning capabilities, behaviour, mental health, academic profile and their study habits. It was observed that students are facing issues including loneliness, stress, anxiety, resistivity to learn new things, lack of confidence, competence and proficiency. The students are aware about these challenges and are trying several techniques to deal with them and stay motivated as these issues are impacting their learning capabilities and success.

In the paper, we have proposed an AI Chatbot based on Natural Language Processing to interact with the students, guide them and motivate then whenever they feel low. The paper also proposes a machine learning model using knn algorithm that could classify a new student as a quick learner or a slow learner based on several factors related to mental health, hours of self-study, and preferred way to study. The model could predict data with an accuracy of 78%. In our subsequent study, we shall try to improve the accuracy by collecting more data about these factors affecting a student's academic performance. Since data is the fuel of AI applications, more the data better are the results.

References

[1] S. Slater, S. Joksimovic, V. Kovanovic, R.s Baker, and D. Gasevic (2016), "Tools for Educational Data Mining: A Review", Journal of Educational and Behavioral Statistics, Vol. 42, No. 1, 2016, pp. 88-106.

[2] Pooja Thakar, Anil Mehta, and Manisha (2015) "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue",International Journal of Computer Application, Vol. 100, No.12, January 2015, pp. 60-68.

[3] Buenano-fernandez D (2019) The use of tools of data mining to decision making in engineering education: a systematic mapping study. Comput Appl Eng Edu. https:// doi. org/ 10. 1002/ cae. 22100.

[4] Ahmed M, Najmul Islam AKM (2019) Deep learning: hope or hype. Ann Data Sci 7(3):427–432. https:// doi. org/ 10. 1007/ s40745- 019- 00237-0.

[5] Gil D, Fernandez-aleman JL, Trujillo J, Garcia-mateos G, Lujan-mora S, Toval A (2018) The effect of green software: a study of impact factors on the correctness of software. Sustainability. https:// doi. org/ 10. 3390/ su101 03471.

[6] Lu OHT, Huang AYQ, Huang JCH, Lin AJQ, Yang SJH (2018) Applying learning analytics for the early prediction of students ' academic performance in blended learning. Edu Technol Soc 21(2):220–232.

[7] Bydžovska H (2015) Are collaborative filtering methods suitable for student performance prediction?pp 425–430. doi: https:// doi. org/ 10. 1007/ 978-3- 319- 23485-4.

[8] Polyzou A, Karypis G (2016) Grade prediction with models specific to students and courses. Int J Data Sci Anal 2(3):159–171. https:// doi. org/ 10. 1007/ s41060- 016- 0024-z.

[9] Khan B, Sikandar Hayat Khiyal M, Daud Khattak M (2015) Final grade prediction of secondary school student using decision tree. Int J Comput Appl 115(21):32–36. https:// doi. org/ 10. 5120/ 20278- 2712.

[10] Shatrma N., Appukutti S., Garg U., Mukherjee J., Mishra S., (2023), Analysis of Student's Academic Performance based on their Time Spent on Extra-Curricular Activities using Machine Learning Techniques, I.J. Modern Education and Computer Science, Vol. 1, pp 46-57.

[11] Al-Azzam N, Elsalem L, Gombedza F. (2020) A cross-sectional study to determine factors affecting dental and medical students' preference for virtual learning during the COVID-19 outbreak. Heliyon. 2020;6(12): e05704.

[12] Jum'ah AA, Elsalem L, Loch C, Schwass D, Brunton PA. (2020). Perception of health and educational risks amongst dental students and educators in the era of COVID-19. Eur J Dent Educ.

[13] Matud, M. P., Díaz, A., Bethencourt, J. M., and Ibáñez, I. (2020). Stress and psychological distress in emerging adulthood: a gender analysis. J. Clin. Med. 9, 2859. doi: 10.3390/jcm9092859.

[14] Ekta Bhambri Marwaha. Subjective well-being among university students & physical activity patterns(2019). Int J Appl Res 2019;5(5):319-325. DOI: 10.22271/allresearch.2019.v5.i5d.11004.

[15] Yunusovich A. V., Ahmedov V., Norboyev K., Zakirov F., (2022), Analysis of Experimental Research Results Focused on Improving Student Psychological Health, I.J. Modern Education and Computer Science, Vol 2, pp 14-30.

[16] Mukun Liu (2022) The Relationship between Students' Study Time and Academic Performance and its Practical Significance.BCP Education & Psychology, 7:412-415. DOI:10.54691/bcpep.v7i.2696.

[17] Vidushi Singla, Rashi Thareja, Reema Thareja,(2022) " Psychometric Analysis Using Computational Intelligence for Smart Choices", International Journal of Modern Education and Computer Science(IJMECS), Vol.14, No.2, pp. 65-75, 2022.DOI: 10.5815/ijmecs.2022.02.06.

Sentiment Analysis Predictions in Digital Media Content using NLP Techniques

Abdulrahman Radaideh[1], Fikri Dweiri[2]
Engineering Management, University of Sharjah, UAE[1]
Industrial Engineering and Engineering Management, University of Sharjah, UAE[2]

***Abstract*—In the current digital landscape, understanding sentiment in digital media is crucial for informed decision-making and content quality. The primary objective is to improve decision-making processes and enhance content quality within this dynamic environment. To achieve this, a comprehensive comparative analysis of NLP for tweet sentiment analysis was conducted, revealing compelling insights. The BERT pre-trained model stood out, achieving an accuracy rate of 94.56%, emphasizing the effectiveness of transfer learning in text classification. Among machine learning algorithms, the Random Forest model excelled with an accuracy rate of 70.82%, while the K Nearest Neighbours model trailed at 55.36%. Additionally, the LSTM model demonstrated excellence in Recall, Precision, and F1 metrics, recording values of 81.12%, 82.32%, and 80.12%, respectively. Future research directions include optimizing model architecture, exploring alternative deep learning approaches, and expanding datasets for improved generalizability. While valuable insights are provided by our study, it is important to acknowledge its limitations, including a Twitter-centric focus, constrained model comparisons, and binary sentiment analysis. These constraints highlight opportunities for more nuanced and diverse sentiment analysis within the digital media landscape.**

***Keywords*—Sentiment analysis; digital media; decision-making; quality assurance; NLP**

I. Introduction

In the contemporary era, digital media has permeated nearly every facet of daily life, fundamentally altering communication, information consumption, and our interaction with the world. From the rapid expansion of social media platforms to the continuous accessibility of online news and entertainment, the digital media landscape has evolved into an omnipresent force shaping public discourse, impacting consumer behavior, and facilitating global connectivity[1].

Central to this profound digital shift is the role of sentiment—the collective emotional undercurrent that flows through the vast sea of digital content. Sentiment in digital media encompasses a spectrum of emotions, from jubilation and enthusiasm to anger and disillusionment [2]. It is the pulse that drives conversations, sparks movements, and dictates the success or failure of digital content, brands, and ideas.

It is also known as opinion mining, refers to the automated process of identifying, extracting, and evaluating sentiments and attitudes expressed within textual data. This process is integral for understanding public perception, consumer sentiment, political discourse, and market trends. It provides valuable insights that guide businesses in crafting effective marketing strategies, aids policymakers in gauging public sentiment on critical issues, and empowers researchers to delve deeper into the intricacies of human communication in the digital age.

The influence of sentiment in the realm of digital media cannot be overstated. Sentiments expressed by users on social media platforms can quickly escalate or deflate public interest in a topic, product, or event. News articles and opinion pieces, often dissected through sentiment analysis, can sway public opinion and even influence political decisions. Understanding and harnessing this dynamic interplay of sentiments within digital media are essential endeavors in an increasingly interconnected and digitized world [3]. Traditionally, sentiment analysis relied on rule-based and statistical approaches, which had their merits but often struggled to capture the subtleties of human language. These methods proved ill-suited for the ever-evolving digital discourse, where slang, context, and linguistic nuances abound. The limitations of conventional sentiment analysis methods have become increasingly evident in the face of the dynamic, multilingual, and culturally diverse nature of digital content [4].

In this context, artificial intelligence (AI) emerges as a transformative force. AI-powered sentiment analysis utilizes advanced Natural Language Processing (NLP) techniques, Machine Learning (ML), and Deep Learning (DL) to decipher the intricacies of language and context, providing unprecedented accuracy and adaptability. The advent of AI has paved the way for a new era in sentiment analysis, characterized by its ability to discern sentiments across various domains, languages, and platforms[5].

The research paper explores the symbiotic relationship between sentiment analysis and artificial intelligence, introducing a novel AI-powered approach that navigates the complexities of sentiment identification and classification in the digital age. The aim is to unravel the intricacies of sentiment analysis [6], elucidating its historical context, traditional methodologies, and the paradigm shift brought about by AI-driven solutions [7].

The aim of this research is to revolutionize decision-making processes and quality assurance within the dynamic realm of digital media by utilizing a groundbreaking AI-powered approach to sentiment analysis [8]. This comprehensive investigation delves into the integration of traditional statistical methodologies, such as the machine learning model, and cutting-edge deep learning architectures, including ML and DL, to discern their respective impacts and

potential synergies in the context of sentiment analysis. With a primary focus on real-time Twitter sentiment data, our aim is to unlock valuable insight that inform strategic decision-making and elevate the standards of quality assurance in digital media [9].

A. Objectives

1) To Evaluate Traditional and DL-Based Sentiment Analysis Models.

2) To Assess the Influence of Twitter Sentiment Analysis: Analyze the impact of real-time Twitter sentiment data on decision-making processes and quality assurance in digital media, emphasizing the added value it brings to sentiment analysis.

3) To Explore Feature Extraction and Sequence Retention: Investigate the ability of the ML and DL architecture to extract intricate features from digital media content while retaining essential temporal sequences, thereby enhancing the accuracy of sentiment analysis.

4) To Conduct a Comparative Analysis: Conduct a rigorous comparative analysis to discern the strengths and limitations of each approach, elucidating the trade-offs and opportunities they present in the context of decision-making and quality assurance.

5) To Provide Practical Recommendations: Offer concrete recommendations and insights for practitioners and stakeholders in digital media, guiding them in leveraging sentiment analysis as a potent tool for informed decision-making and robust quality assurance.

These objectives collectively drive our endeavor to advance the understanding of sentiment analysis's transformative potential and its pivotal role in shaping the future of decision-making and quality assurance within digital media.

In the contemporary landscape dominated by digital media, this research aims to redefine the boundaries of sentiment analysis by introducing a novel and comprehensive AI-powered approach. Unlike traditional sentiment analysis methodologies, which often struggle to adapt to the evolving nature of digital discourse, this work stands out through the integration of both established statistical techniques and cutting-edge deep learning architectures. This distinctive fusion allows for capturing the intricacies of sentiment in a dynamic, multilingual, and culturally diverse digital environment.

B. Contributions of the Research

1) Hybrid methodology integration: One of the primary novelties of this approach lies in seamlessly integrating traditional statistical methodologies, such as machine learning models, with advanced deep learning architectures. This unique hybrid methodology capitalizes on the strengths of both approaches, providing a more robust and adaptable sentiment analysis framework.

2) Real-time twitter sentiment analysis impact: Extending beyond conventional sentiment analysis, this research places particular emphasis on the real-time analysis of sentiment in the context of Twitter data. This distinctive focus allows for exploring and quantifying the immediate impact of sentiments on decision-making processes and quality assurance in digital media.

3) Temporal sequence retention in feature extraction: Addressing a critical gap in existing literature, this study delves into the temporal dynamics of sentiment by investigating the ability of ML and DL architectures to extract intricate features from digital media content while retaining essential temporal sequences. This novel approach enhances the accuracy of sentiment analysis, especially in capturing the temporal evolution of sentiments over time.

Through these novel contributions, this research advances the understanding of sentiment analysis's transformative potential, setting a new standard for decision-making processes and quality assurance within the ever-evolving realm of digital media.

II. Literature Reviews

The intersection of sentiment analysis and digital media has garnered significant attention in recent years, reflecting the growing recognition of the pivotal role sentiments play in shaping online discourse, content quality, and decision-making processes [10]. This literature review provides an overview of the historical context, key methodologies, and prior applications of sentiment analysis in the realm of digital media, highlighting the evolving landscape of this interdisciplinary field [11].

A. Historical Context of Sentiment Analysis in Digital Media

The roots of sentiment analysis can be traced back to the early days of NLP and ML [12]. Early sentiment analysis efforts primarily focused on binary sentiment classification, distinguishing between positive and negative sentiments in textual data. As digital media evolved, sentiment analysis adapted to accommodate the nuanced and multifaceted nature of sentiments ex-pressed in online content [13].

The advent of social media platforms in the early 2000s marked a significant turning point [14]. Researchers and organizations recognized the potential of sentiment analysis to extract valuable insights from the vast volumes of user-generated content on platforms like Twitter and Facebook. Since then, sentiment analysis has matured into a sophisticated field, incorporating advanced NLP techniques and machine learning models to capture sentiments in real time across a wide array of digital media sources [15].

B. Key Methodologies and Technologies in Sentiment Analysis

Sentiment analysis methodologies have evolved in tandem with advancements in NLP and AI technologies [16]. Early approaches relied on lexicon-based sentiment analysis, using predefined lists of words and phrases associated with positive and negative sentiments. While effective to some extent, these approaches struggled with con-text and sarcasm [17].

Machine learning techniques, particularly supervised learning, revolutionized sentiment analysis by enabling models to learn sentiment patterns from labeled datasets. Techniques such as SVM, Naive Bayes NB, and more recently DL models like RNN and Transformers, have significantly improved sentiment analysis accuracy [18]. Pre-trained language models, such as BERT and GPT-3, have further elevated sentiment analysis by capturing contextual nuances and domain-specific sentiment [19].

C. Prior Applications of Sentiment Analysis in Decision-Making and Quality Assurance within Digital Media

The applications of sentiment analysis within digital media are multifaceted and extend across various domains:

1) Content creation and optimization: Content creators use sentiment analysis to gauge audience reactions to their articles, videos, or social media posts. In-sights from sentiment analysis inform content optimization strategies, helping creators tailor content to audience preferences [20].

2) Engagement strategies: Organizations leverage sentiment analysis to identify viral content and trends. By understanding sentiment patterns, they can craft engagement strategies that resonate with their target audience and enhance brand loyalty [21].

3) Quality assurance: Sentiment analysis plays a crucial role in maintaining content quality. It assists in content moderation by identifying inappropriate or harmful content and helps fact-checkers and journalists identify misinformation and fake news [22].

4) Advertising and marketing: Marketers analyze sentiment to measure the effectiveness of advertising campaigns. They also use sentiment insights to personalize ad targeting and messaging.

5) News and journalism: Sentiment analysis aids news outlets in understanding public sentiment towards news stories, political events, and social issues. This information can influence editorial decisions and story selection.

6) Public opinion and policy making: Governments and policymakers monitor online sentiment to gauge public opinion on policy issues and to respond proactively to emerging trends or concerns [23].

The literature reviewed underscores the transformative potential of sentiment analysis in digital media [18]. It has evolved from a binary classification task to a sophisticated field empowered by AI and machine learning, offering valuable insights for decision-makers, content creators, and quality assurance processes [24].

D. Gap Analysis

In our research, notable gaps emerge, urging further exploration in the realm of sentiment analysis in digital media [25]. These include the integration of multi-modal data sources, the development of real-time decision support systems, cross-platform sentiment analysis, ethical considerations, and user-centric sentiment analysis [26]. Closing these gaps promises to enhance the depth and breadth of sentiment analysis applications, ensuring its ethical use, and fostering personalized, real-time decision-making in the dynamic digital media landscape [27] (see Table I).

TABLE I. GAPS ANALYSIS

Year	Technique	Dataset	Accuracy Achieved	Application	Pros	Cons
2017	Approaches Using Machine Learning and Lexicons	Twitter dataset	83.3%	Sentiment analysis of tweets	Machine learning can analyze text without feature engineering	Traditional methods require feature engineering
2018	Long Term Memory (LSTM) and Convolutional Neural Networks (CNN)	English language tweets	88.5%	Sentiment analysis of tweets	Efficient and reliable technique	-
2019	Convolutional, recurrent, neural networks, unsupervised, and mixed neural networks, as well as deep reinforcement learning	-	-	Sentiment analysis of texts	Can recognize new complex features	Less accurate than supervised techniques
2020	Algorithms for supervised machine learning, such as Support Vector Machines and Artificial Neural Networks	User-created texts	88.5%	Sentiment analysis of texts	Can extract users' feelings from their writing	Slow and take a long time to train
2021	Support vector machines (SVM) and Universal Language Model Fine-tuning (ULMFiT)	-	-	Sentiment analysis of texts	Powerful deep learning architecture	-
2022	ELMO and CNN	Twitter dataset	-	Clustering in service discovery	Effective discovery of the best service	-

III. METHODOLOGY

The study utilized a dataset obtained from the Kaggle website, comprising around 160,000 tweets from the Twitter blog categorized into three groups: positive, negative, and neutral. Initial analysis involved applying various pre-processing techniques to cleanse and prepare the tweets for feature extraction. Subsequently [28], the database was split into a training set and a test set. Features were extracted from tweets using diverse techniques, and machine learning algorithms were trained for tweet polarity classification, including support vector machine, naive Bayes, decision tree, and K-nearest neighbor approaches [29]**.** The performance of these classifiers was then assessed on the test set across all extraction techniques to compare their impact on the sentiment analysis process, using multiple performance evaluation measures. Following this evaluation, a model was proposed, specifically a recurrent neural network employing Long Short-Term Memory (LSTM). The results obtained from this model were then compared with the outcomes of the previous classifiers [30] (see Fig. 1).

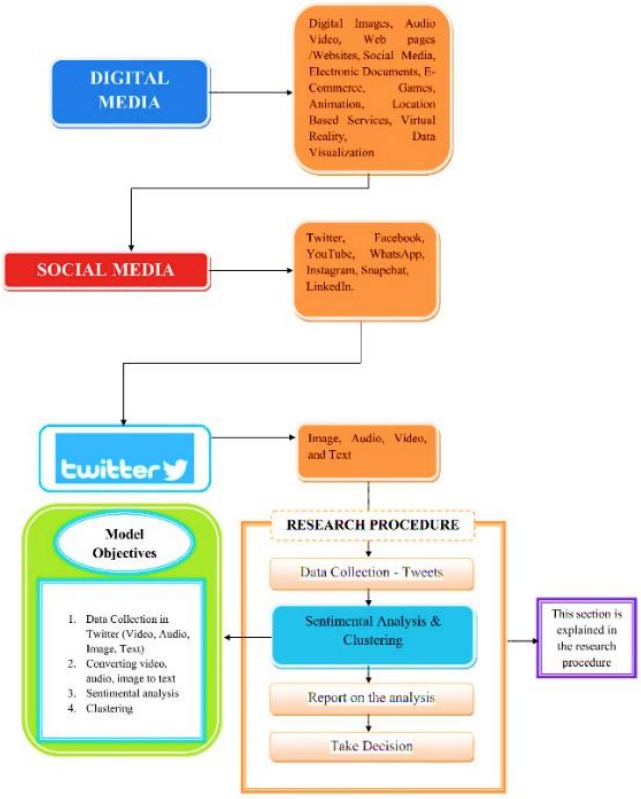


Fig. 1. General framework.

A. Tools and Resources

To accomplish the task of sentiment analysis, various tools and resources are necessary. The required tools and resources for analyzing the sentiment of tweets are discussed below:

1) Programming language: Python, a high-level general-purpose programming language, proves to be an excellent tool for artificial intelligence, machine learning, and deep learning. Python will be used for creating and training models. Several Python libraries will be employed for sentiment analysis, including Pandas, Numpy, Scikit Learn, NLTK, Re, Keras, PyTorch, and Transformers.

2) Software: Anaconda, a Python distribution platform, will be utilized, providing access to many built-in packages. Within Anaconda, Jupyter Notebook will serve as the primary environment for developing and training machine learning and deep learning models.

3) Twitter API: After developing and evaluating the machine learning and deep learning models, the Twitter API will be employed to extract new tweets and test the model's performance. Twitter allows the use of 3rd party Python packages like Tweepy to extract tweets based on query words and date range, facilitating the entire process.

4) Hardware: Given that various machine learning and deep learning models will be trained, the minimum system requirements are as follows: Core i5 Processor, 16 GB of RAM, Nvidia GPU with a minimum of 6 GB of V-RAM, and 100 GB of HDD space.

B. Machine Learning-Based Models

Through this model, we will initially process the data, extract the features, and then classify them based on machine learning algorithms. Fig. 2 shows the work steps.

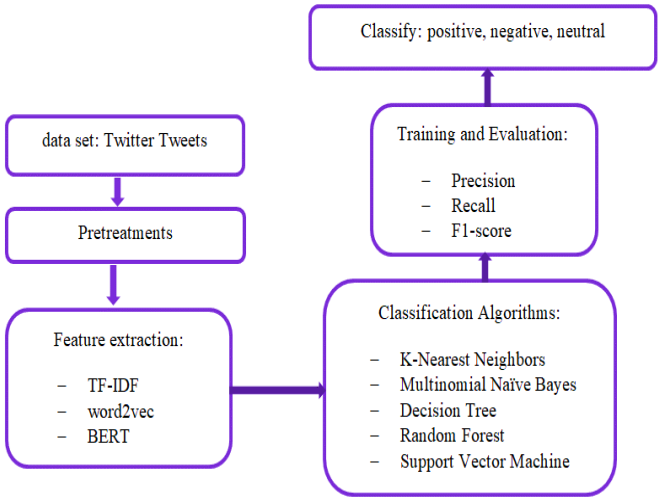


Fig. 2. Framework of model.

The model involves several steps, including pre-processing the data, feature ex-traction using techniques such as TF-IDF, word2vec, and BERT, and classifying the tweets using various algorithms such as k-nearest neighbors, multinomial Naïve Bayes, decision tree, random forest, and support vector machine. The model will be trained and evaluated using performance metrics such as precision, recall, and F1-score.

C. Features Selection

1) Data pre-processing: Since machines do not understand spoken or written natural language, data pre-processing is a very important step for sentiment analysis and a necessary process before training machine learning models. Data pre-processing aims to make it easier to train and test

classifiers by performing an appropriate set of transformations on the data. We did the pre-processing using the NLTK library in Python. Because we work with text data from Twitter. Tweets contain different parts that are not necessary or important to understanding the meaning of a tweet. Where we can extract the semantic meaning of the tweet by getting rid of all the unnecessary words and symbols by pre-processing the data. The tweets were pre-processed by selecting features that were likely to be relevant for sentiment analysis. In general, when selecting features for sentiment analysis, some common criteria researchers consider include relevance, in-formativeness, redundancy, computational efficiency, and interpretability. The pre-processing of the data follows the following steps: remove punctuation, remove stop words, remove URLs, remove emoji, remove hash marks, and drop all word wrapping, derivation, and markup. The data is now clean and ready for feature ex-traction. Fig. 3 shows the steps for data pre-processing.

2) Data preprocessing: The steps involved in pre-processing include converting letters to upper/lower case, tokenizing the text into individual units, removing unwanted characters and stopwords, normalizing the text, stemming to remove affixes, and lemmatization to reduce words to their base form. Finally, the pre-processed text is vectorized to convert the text data into numerical form that can be utilized by machine learning models.

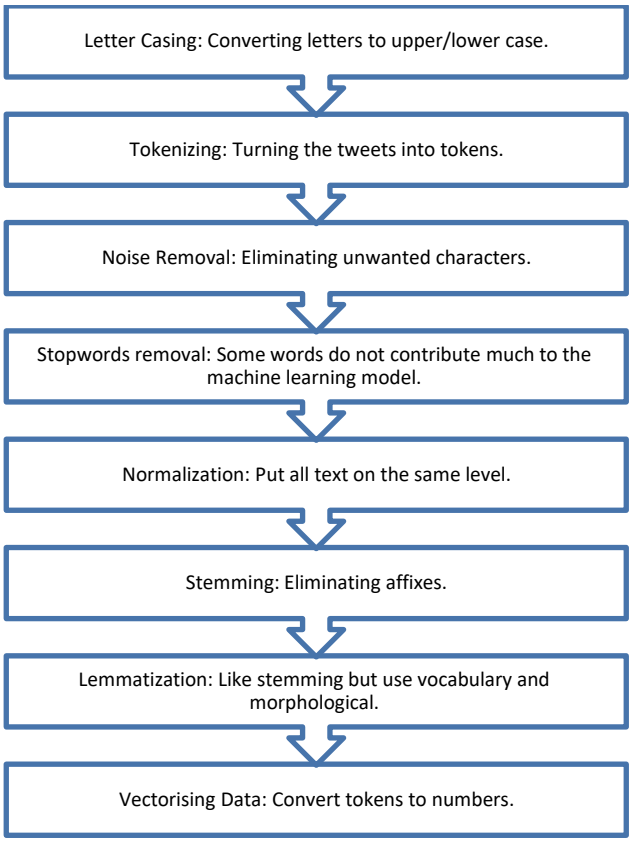


Fig. 3. Data pre processing.

3) Feature extraction: The study employed three techniques to extract features from text data:

a) TF-IDF algorithm (term frequency-inverse document frequency): is a statis-tical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document and the inverse document frequency of the word across a set of documents. To implement the TF-IDF technique, the implementation of the TF-IDF technique utilized a class from the sklearn library.

b) word2vec: is a predictive model for computing a continuous radial representation of quantities in large data sets. The given models use two alternative models to get a high dimensional vector for each word:

- PCA: A technique focused on reducing the dimensions of words that directly impact how the original set of vectors transforms into a new set.
- t-SNE: A technique for nonlinear dimensionality reduction and data visualization. It combines words from a higher dimension with ones from a lower dimension. The Gensim library was used to construct word vectors using word2vec, with parameters like tokenized words and mincount set accordingly.

c) BERT: is a pre-trained language model for deep, bidirectional representations of unlabeled text by co-adapting on both the left and right context in all layers. BERT can be used in a variety of language tasks, with only a small layer added to the base model. BERT was used in two ways:

- Use the hugging face BERT model to fine-tune our sentiment analysis.
- Use the BERT model for fine-tuning and training on our dataset.

D. Machine Learning Algorithms

Various machine learning can be used to accomplish the task of sentiment analysis. The following machine learning algorithms are used:

The sentiment analysis task involved the utilization of a diverse set of machine learning and deep learning algorithms:

1) K-Nearest Neighbors (KNN): Identify the group to which a new data point (tweet) belongs based on training data. If a new tweet is close to a negative group, it is classified as negative; if close to positive or neutral, the prediction is made accordingly.

2) Multinomial Naïve Bayes: Determine the probability of a tweet being positive, negative, or neutral based on its contents or words.

3) Decision tree: Classify a tweet based on its features, i.e., the words it contains.

4) Random forest: Constructed from multiple decision trees to provide a more accurate and stable prediction. Operates as an ensemble, potentially offering improved results compared to a single decision tree.

5) Support vector machine: Utilize a non-linear Support Vector Classification model to categorize a tweet into positive, negative, or neutral classes.

6) Voted classifier: Ensemble approach incorporating KNN, Multinomial Naïve Bayes, Random Forest, and Support Vector Machine to enhance sentiment analysis predictions.

7) LSTM: Leverage deep learning with LSTM, specifically effective for sequential data like text. LSTM's capacity to remember long-term dependencies in words contributes to its success in handling text sequences.

8) Transformer network: Implementation of an encoder-only transformer model and the use of a pre-trained transformer model for sentiment analysis.

9) K-Means clustering: Segregate groups with similar traits and assign them into clusters.

E. Model Training and Evaluation

- Training: Model training is an integral part of the whole process. It is very important to set the hyper-parameters of the models to the right ones to achieve good results.
- Evaluation: As we compare different machine learning models for sentiment analysis, various evaluation metrics will be employed to determine the model's performance.
- Confusion Matrix.

1) Precision.
2) Recall.
3) F1 Score.
4) AUC-ROC Curve.

F. Model Testing

After training and evaluating the model, the best-performing model will be selected for testing with new tweets from Twitter. This process can be implemented using the Twitter API to create a user-friendly web server. Users can enter a keyword and date range, and the server will display the corresponding tweets along with their polarity. This provides organizations or individuals with valuable insights into what people are tweeting about their products or themselves.

G. Comparative Analysis

A comparative analysis is provided for various sentiment analysis models and techniques applied to the digital media dataset. The objective is to assess and contrast the performance of these models in terms of their ability to accurately classify sentiment, computational efficiency, and practical applicability in real-world scenarios.

IV. Result and Discussion

All models underwent training on 80% of our dataset, with the remaining 20% reserved for validation. Accuracy was employed as the performance metric during training, focusing on the validation accuracy of the models. The results are presented in Table II.

From the table above, it's evident that the BERT pre-trained model significantly outperformed even the LSTM model. Among the machine learning models, the random forest exhibited superior performance compared to others, including the voting classifier.

TABLE II. Validation Accuracy

Model	Validation Accuracy
K-Nearest Neighbors	55.36%
Multinomial Naïve Bayes	65.18%
Decision Tree	66.55%
Random Forest	70.82%
Support Vector Machine	65.98%
Voted Classifier	69.86%
LSTM	81.12%
BERT	94.56%

A. Evaluation

Additional metrics, such as Confusion Matrix, Recall Score, Precision Score, and F1 Score, were employed to evaluate and compare the sentiment analysis models. The evaluation results for the aforementioned algorithms based on these criteria are presented in Fig. 4 and Table III.

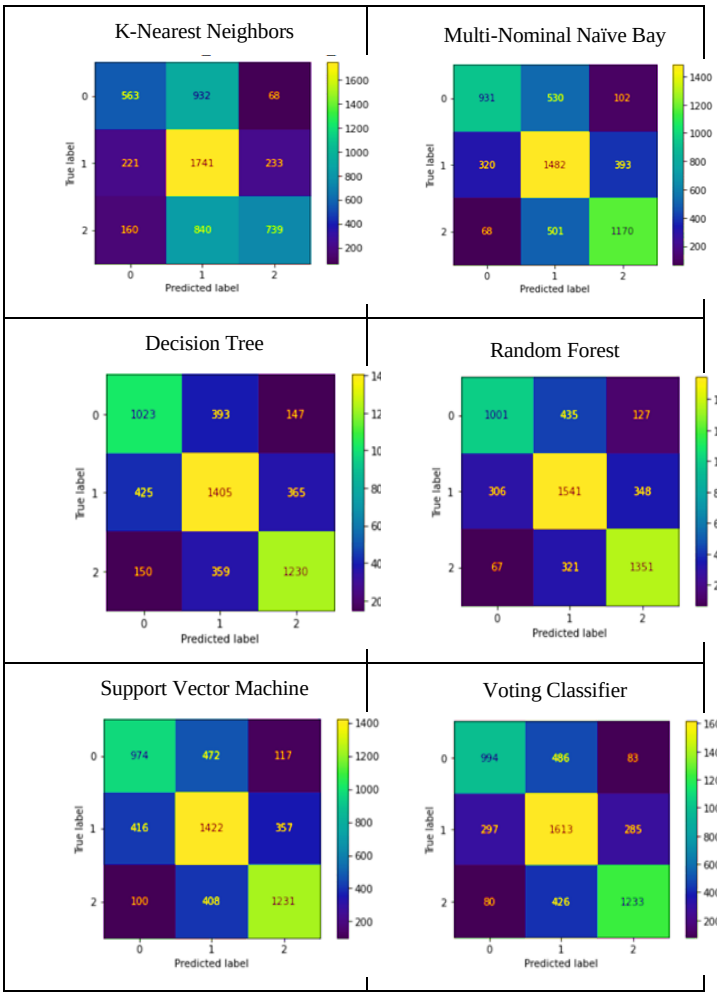


Fig. 4. Confusion matrix.

TABLE III. RECALL, PRECISION, F1 SCORE OF ALL ALGORITHMS

Model	Recall	Precision	F1 scores
K-Nearest Neighbors	0.5536	0.5923	0.5436
Multinomial Naïve Bayes	0.6518	0.6584	0.6321
Decision Tree	0.6655	0.6655	0.6509
Random Forest	0.7082	0.7091	0.6904
Support Vector Machine	0.6598	0.6609	0.6566
Voted Classifier	0.6986	0.7049	0.6921
LSTM	0.8112	0.8232	0.8012
BERT	0.9455	0.9551	0.9499

From the confusion matrix, it can be seen that most of the algorithm is best at classifying neutral tweets and they have a tendency to classify other tweets as neutral also.

The table shows how well different machine learning algorithms performed on a task, based on three metrics: recall, precision, and F1 score. The LSTM and BERT models had the highest scores, indicating they were the most effective algorithms.

B. Proposed Method-based Deep Learning Algorithm (RNN-LSTM)

This method aims to enhance the preceding approach through an exploratory analysis of data to extract features, followed by the application of a deep learning algorithm to classify tweets into positive, negative, or neutral categories. Fig. 5 illustrates the proposed framework for sentiment analysis. The improved method will be applied to the existing data for comparative evaluation with the previous approach, assessing the effectiveness of the proposed method.

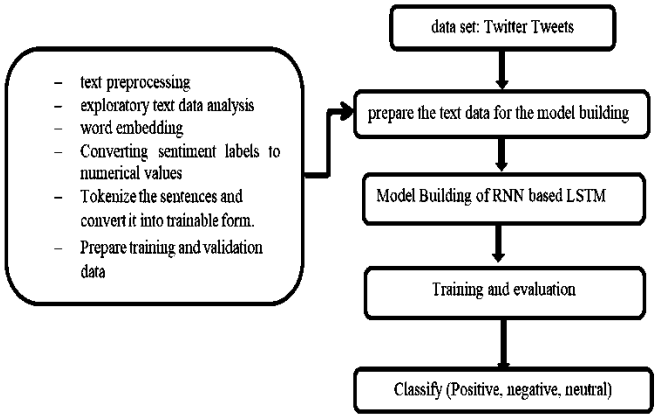


Fig. 5. Framework for proposed model.

C. Algorithms

Step 1: prepare the text data for the model building

1) Pre-processing of text. It represents the initial phase of NLP projects. Some of the pre-processing to use a text cleaning procedure to clean all the content: Stop words, URLs, and punctuation such as! $() *% @, lowercase, stemming, tokenization, and lemmatization have been removed.

2) Exploratory text data analysis is a straightforward yet incredibly informative method. In order to better understand the basic traits of the text data, it comprises (word frequency analysis, sentence length analysis, average word length analysis, distribution of the number of words, etc.). For instance:

- The distribution of the number of words for each sentiment. So that we will use these features in the model training process.
- Distribution of the number of letters for each sentiment.
- The count of the most frequent words in the entire text is essential for reinforcing the analysis in the feature extraction phase.

3) Apply word embedding to improve the model accuracy.

A word embedding is a type of learned representation of text in which words with the same meaning are represented similarly. It is regarded as one of the key developments in deep learning of document and word encoding for challenging natural language processing problems.

Word embedding is a method where individual words are represented as real-valued vectors in a specified vector space. Since each word is assigned to a distinct vector and the vector values are learned similarly to a neural network, the technique is frequently referred to as deep learning. Through the use of word embedding techniques, a corpus of literary works is used to learn a real-valued vector representation for a preset set-sized vocabulary. Other tasks, such as document categorization, include either unsupervised learning using document data or learning in conjunction with a neural network model. The word embedding method was developed using the Gensim library's keyed vectors library.

4) After loading the data, transform the sentiment into a numerical representation. All of the target categorical values must be converted to numerical format because the model will train on numbers and understand numbers better. Therefore, the model will be able to very effectively learn the target. The Python tools offer a number of methods that can be used to convert categorical data into appropriate numerical values; we utilized TensorFlow from the Gensim module**.**

5) Eliminating superfluous columns and compiling a text list with the target sentiment.

6) Tokenize the statements and modify them so that they can be trained.

- Tokenization is the process of breaking up a long block of text into tokens. Words, letters, or sub words can all be tokens in this context. So, there are three main categories of tokenization: word, character, and sub-word (n-gram characters). Take the phrase "Never give up" as an illustration.
- Tokens are most frequently created based on space. The tokenization of the statement yields three tokens, Never-give-up, assuming space as a delimiter. Every token is a word; hence it serves as an illustration of Word tokenization.

7) Prepare the embedding matrix as well as the training and validation data.

- A sample of data is used to unbiased evaluate how well a model fits a training dataset while modifying model hyperparameters is known as a validation dataset. The evaluation becomes more skewed when skill from the validation dataset is added to the model setup.
- A list of all words and their accompanying embeddings is called an embedding matrix. The embedding matrix is prepared by importing the train-test-split from Sklearn.

Step 2: Model Building

LSTMs employ a number of "gates" that regulate how data in a sequence enters, is stored in, and leaves the network. A typical LSTM has three gates: an output gate, an input gate, and a forget gate. Each of these gates is a separate neural network and may be thought of as a filter.

In order to avoid overfitting, the Dropout layer randomly sets input units to 0 with a frequency of rate at each step during training. Keep in mind that the Dropout layer only functions when the model's training is set to true, preventing any values from being dropped during inference.

A dense layer is densely connected to the layer above it is one in which every neuron in the layer is coupled to every other neuron in the layer above. The majority of artificial neural network networks employ this layer.

The values are unrolled starting with the last dimension when using the flatten operator.

Dropout layers: Since a dropout layer doesn't have any weights, it lacks parameters. A dropout layer only increases the likelihood that a neuron won't be tested by 1%. In a dropout layer, nothing more needs to be configured. We successively import layers, constants, dense, embedding, flatten, and initializers from keras in order to build the model.

Step 3: Model Training

The number of samples that must be processed before the internal model parameters are changed is determined by the hyperparameter known as batch size. A for-loop is a type of batch that makes predictions while iterating through one or more samples. At the end of the batch, the predictions are compared to the expected output variables, and an error is then calculated. This issue is fixed using the updated algorithm, for instance by lowering the gradient of the error. The number of epochs hyperparameter controls how many times the learning algorithm will run through the entire training dataset.

For every sample in the training dataset, the internal model parameters changed once throughout an epoch. An era is made up of one or more batches. For instance, a single-batch epoch is described by the batch gradient descent learning process.

Step 4: Model Accuracy

Model accuracy is a measure of the proportion of correct predictions made by a model out of the total number of predictions produced. This metric is commonly used to assess a model's performance, although other metrics may also be considered. We found that this model can get 96% accuracy and this is better than our previous analysis.

Step 5: Plotting and Displaying Results

From Fig. 6, during the sentiment analysis model's training phase, the loss and AUC metrics are shown for the training and validation sets. During the first 10 epochs, the loss is greatly reduced while the precision is noticeably improved.

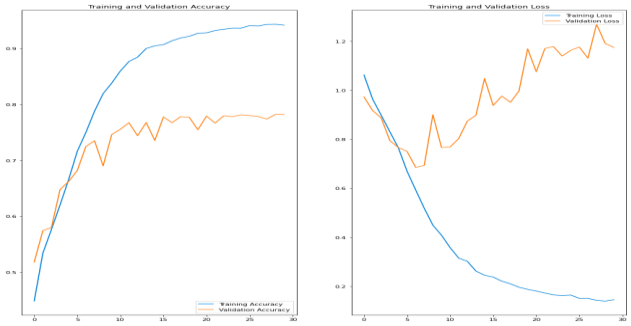


Fig. 6. Plot of results of RNN model.

Step 6: Based on their shared characteristics, various data subsets are segmented through clustering. Python offers a wide array of useful cluster analysis tools, and the choice of strategy depends on the specific task and the nature of the available data. Commonly utilized techniques in Python include Gaussian mixture models, spectral clustering, and K-means clustering. In this scenario, the K-means clustering method is employed. K-means clustering, a type of unsupervised machine learning, exclusively trains on inputs without generating outputs. It identifies distinct clusters of data points that are closest to each other. Once the data is partitioned into clusters, each point is assigned to the cluster whose mean is closest to that specific data point. We employ K-means clustering to create sentiment-based clusters from our data. For the implementation of sentiment analysis on the tweet data, a comprehensive pipeline has been developed. The model assigns different tweets to each of our clusters, encompassing three distinct sentiment labels (see Fig. 7 and 8).

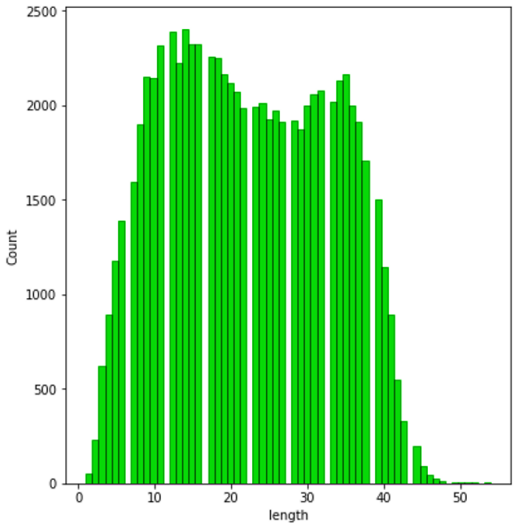


Fig. 7. Positive sentiments tweets.

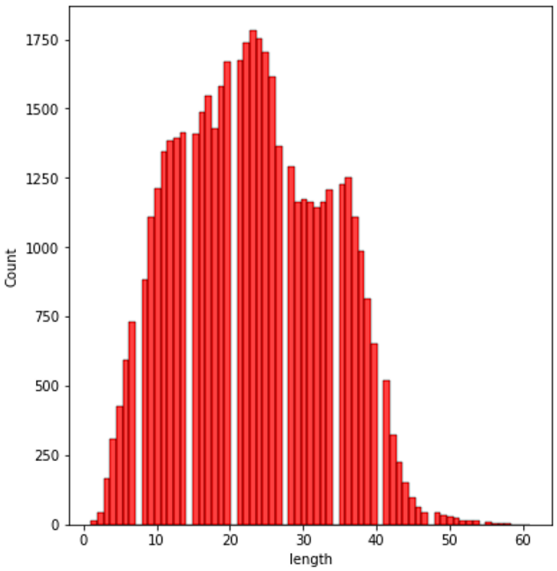


Fig. 8. Negative sentiments tweets.

So, the clustering algorithms assign a numerical value to each of the labels and assign a cluster based on its text context.

D. Summary of Results

The following are the main conclusions from the comparison of machine learning and deep learning algorithms for tweet sentiment analysis:

1) The BERT pre-trained model outperformed the LSTM model, achieving the greatest accuracy of 94.56%. This demonstrates that transfer learning for text classification problems can be quite successful when employing pre-trained language models.

2) With an accuracy rate of 70.82%, the Random Forest algorithm outperformed the other machine learning models. The method used in K Nearest Neighbours has the lowest accuracy, 55.36%.

3) The LSTM model received the greatest ratings for Recall, Precision, and F1, with values of 81.12%, 82.32%, and 80.12%, respectively, suggesting that it performed the best overall.

4) By achieving 96% accuracy, the suggested RNN-LSTM model with word embedding, dropout, and clustering proved the value of deep learning for sentiment analysis.

5) The model performance was evaluated thoroughly and rigorously utilizing confusion matrices, recall, precision, and F1 scores.

In conclusion, for tweet sentiment analysis, deep learning techniques, particularly the LSTM and BERT models, outperformed machine learning algorithms. To create more reliable and generalized models, more study is necessary.

V. Conclusions

Utilizing a range of machine learning and deep learning techniques on Twitter data, our aim was to enhance decision-making and content quality in the dynamic digital media landscape. Key findings highlight BERT's exceptional 94.56% accuracy, showcasing transfer learning's effectiveness. Noteworthy results include the Random Forest algorithm (70.82% accuracy) and the LSTM model, excelling in Recall, Precision, and F1 scores. Deep learning, exemplified by the RNN-LSTM model, demonstrated exceptional potential with a 96% accuracy, establishing LSTM and BERT as tweet sentiment analysis frontrunners. Future research should focus on refining models for real-world applications, exploring optimization, alternative architectures, and dataset expansion. Acknowledging study limitations points toward opportunities for a more nuanced approach within the digital media landscape.

In summary, this study lays the foundation for leveraging advanced sentiment analysis techniques, emphasizing the pivotal role of deep learning models, while recognizing the evolving nature of research in this domain.

References

[1] M. A. e. al, "Energy Choices in Alaska: Mining people's Perception and Attitudes from Geotagged Tweets,," Renewable and Sustainable Energy Reviews, vol. 124, p. 109781, May 2020, 2020.

[2] A. a. N. Elmitwally, "A Comprehensive Study for Arabic Sentiment Analysis (Challenges and Applications),," EgyptianInformatics Journal, vol.21,no.1,pp.7–12,Mar.2020,doi: https://doi.org/10.1016/j.eij.2019.06.001, 2020.

[3] G. A. e. al., "AI in the media and creative industries," New European Media , vol. 01, 2019.

[4] I. C.-P. J. F. S.-R. a. C. A. I. O. Araque, "Enhancing Deep Learning Sentiment Analysis with Ensemble Techniques in Social Applications," Expert Systems with Applications, vol. 77,, 2017.

[5] M. E. B. a. A. Kabiri, "HOMPer: A New Hybrid System for Opinion Mining in the Persian Language," Journal of Information Science, vol. 46, no. 1, pp. 101–117, , 2019.

[6] B. S. Y. X. a. C. H. E. Cambria, "New Avenues in Opinion Mining and Sentiment Analysis," IEEE Intelligent Systems, vol. 28, no. 2, 2018.

[7] Y. L. F. Z. X. S. P. a. K. K. E. Cambria, "SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis," Proceedings of the 29th ACM International Conference on Information & Knowledge Management, vol. 28,, 2020.

[8] N. C. a. P. Wang, "Advanced Combined LSTM-CNN Model for Twitter Sentiment Analysis," in Proceedings of 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), IEEE, 2018.

[9] M. T. A. A. S. M. A. M. E. B. a. X. Z. U. A. Chauhan, "A Comprehensive Analysis of Adverb Types for Mining User Sentiments on Amazon Product Reviews," Wide Web, vol. 23, no. 3, pp. 1811–1829, 2020.

[10] S. H. I. A. A. A. A. N. O. D. I. A. S. M. Y. I. S. A. e. a. Muhammad, "Afrisenti: A twitter sentiment analysis benchmark for african languages," arXiv preprint arXiv:2302.08956, 2023.

[11] S. S. S. J. C. a. M. S. Park, "Mind games: A temporal sentiment analysis of the political messages of the Internet Research Agency on Facebook and Twitter," New Media & Society 25, no. 3 , 2023.

[12] R. S. P. C. C. C. P. a. M. E. R. S. P. C. C. C. P. a. M. E. Catelli, "Lexicon-based sentiment analysis to detect opinions and attitude towards COVID-19 vaccines on Twitter in Italy.," Computers in Biology and Medicine 158, 2023.

[13] N. A. B. S. V. J. B. a. R. B. Braig, "Machine Learning Techniques for Sentiment Analysis of COVID-19-Related Twitter Data.," IEEE Access 11, 2023.

[14] O. M. C.-F. W. C.-F. a. K. I. Carvache-Franco, "Topic and sentiment analysis of crisis communications about the COVID-19 pandemic in Twitter's tourism hashtags," Tourism and Hospitality Research 23, no. 1 , 2023.

[15] S. R. J. a. H. Ramasangu, "Classification of Cognitive States using Task-Specific Connectivity Features," Eng. Technol. Appl. Sci. Res., vol. 13, no. 3, pp. 10675–10679, 2023.

[16] S. M. K. M. U. F. a. .. W. M. Anwer, "Attack Detection in IoT using Machine Learning," Eng. Technol. Appl. Sci. Res., vol. 11, no. 3, pp. 7273–7278, 2021.

[17] V. P. V. a. V. K. Yadav, "Long short term memory (LSTM) model for sentiment analysis in social data for e-commerce products reviews in Hindi languages," International Journal of Information Technology 15, no. 2 , 2023.

[18] O. A. A.-R. J. L. H. S. S. B.-C. J. Z.-P. J. A. Y. a. M. C.-C. Iparraguirre-Villanueva, "The public health contribution of sentiment analysis of Monkeypox tweets to detect polarities using the CNN-LSTM model," Vaccines 11, no. 2, 2023.

[19] A. M. E. P. a. A. S. Xu, "Sentiment Analysis On Twitter Posts About The Russia and Ukraine War With Long Short-Term Memory," Sinkron: jurnal dan penelitian teknik informatika 8, no. 2 , 2023.

[20] M. W. A. A. W. A. S. K. M. U. C. I. a. T. R. G. Bibi, "A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis.," Pattern Recognition Letters 158, 2022.

[21] N. S. J. A. P. S. B. K. S. A. L. A. K. Y. a. J. T. Leelawat, "Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning," Heliyon 8, 2022.

[22] M. S. a. R. R. Neethu, "Sentiment analysis in twitter using machine learning techniques.," n 2013 fourth international conference on computing, communications and Networking Technologies (ICCCNT), 2013.

[23] A. a. R. O. Alslaity, "Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions," Behaviour & Information Technology, 2022.

[24] P. Q. T. a. H. B. N. M. D. Nguyen, "An Application of Analytic Network Process (ANP) to Assess Critical Risks of Bridge Projects in the Mekong Delta Region"," Eng. Technol. Appl. Sci. Res., vol. 13, no. 3, pp. 10622–10629, 2023.

[25] S. J. A. M. B. R. S. a. R. T. Mann, "Twitter sentiment analysis using enhanced BERT.," In Intelligent Systems and Applications: Select Proceedings of ICISA 2022,, 2022.

[26] G. a. A. S. Kaur, "A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis," Journal of Big Data 10, no. 1, 2023.

[27] S. K. Shrivastav, "Exploring the application of analytics in supply chain during COVID-19 pandemic: a review and future research agenda.," Journal of Global Operations and Strategic Sourcing 16, no. 2, 2023.

[28] H. S. S. S. M. T. M. a. P. M. Madhu, "Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments.," Expert Systems with Applications 215, 2023.

[29] F. Sufi, "Algorithms in low-code-no-code for research applications: a practical review.," Algorithms 16, no. 2, 2022.

[30] S. C. M. F. G. A. a. E. N. Fanni, "Natural language processing.," In Introduction to Artificial Intelligence, pp. 87-99. Cham: Springer International Publishing, 2023.

An Enhanced Approach for Realizing Robust Security and Isolation in Virtualized Environments

Rawan Abuleil[1], Samer Murrar[2], Mohammad Shkoukani[3]
Department of Computer Science, Philadelphia University, Amman, Jordan[1]
Department of Computer Science, Applied Science Private University, Amman, Jordan[2, 3]

***Abstract*—Transitioning into the next generation of supercomputing resources, we're faced with expanding user bases and diverse workloads, increasing the demand for improved security measures and deeper software compartmentalization. This is especially pertinent for virtualization, a key cloud computing component that's at risk from attacks due to hypervisors' integration into privileged OSs and shared use across VMs. In response to these challenges, our paper presents a two- pronged approach: introducing secure computing capabilities into the HPC software stack and proposing SecFortress an enhanced hypervisor design. By porting the Kitten Lightweight Kernel to the ARM64 architecture and integrating it with the Hafnium hypervisor, we substitute the Linux-based resource management infrastructure, reducing overheads. Concurrently, SecFortress employs a nested kernel approach, preventing outerOS from accessing mediator's memory, and creating a hypervisor box to isolate untrusted VMs' effects. Our initial results highlight significant performance improvements on small scale ARM-based SOC platforms and enhanced hypervisor security with minimal runtime overhead, establishing a solid foundation for further research in secure, scalable high- performance computing.**

***Keywords*—Virtual Machine (VM); High-Performance Computing (HPC); cybersecurity; hypervisor security**

I. INTRODUCTION

Advances in computing technology have ushered in a paradigm shift in how computational resources are deployed and utilized in recent decades. The rapid growth and adoption of virtualization technologies are at the forefront of this transition [1]. By abstracting the physical hardware from the software, virtualization enables the creation of multiple isolated Virtual Machines (VMs) that can run concurrently on a single physical machine, resulting in significant improvements in resource utilization and cost efficiency [1].

However, as with any technology, virtualization brings with it new challenges, most notably in the area of security [1]. The hypervisor, the abstraction layer that allows the creation of VMs, is a lucrative target for attackers [2]. If an attacker successfully compromises the hypervisor, they may gain control of all VMs running on the system, resulting in a significant security breach [3]. Furthermore, while the isolation of VMs from each other and the host system is beneficial for security, it can also be used by attackers to conceal malicious activity [3].

This necessitates the development of enhanced security solutions capable of effectively protecting the hypervisor and the virtual machines that run on it [4]. Several technologies have been developed to this end, providing various mechanisms for securing virtualized environments [4]. Hafnium, ARM TrustZone, and SecFortress, for example, provide unique security solutions that can significantly improve the security of virtualized environments [1, 2, 5].

Hafnium is a microkernel-based VM monitor that provides secure isolation between virtual machines [5]. It enables each VM to run in its own isolated environment, memory-separated from other VMs [5]. This means that even if an attacker gains control of one VM, they cannot access the memory of other VMs [5]. Hafnium also ensures control flow integrity (CFI), which prevents unauthorized changes to a VM's control flow [5]. Fig. 1 depicts the default hafnium system architecture. Hafnium, as illustrated in the diagram, relies on a primary VM to make scheduling decisions and explicitly invoke context switches to secondary VMs via a privileged hyper-call interface.

ARM TrustZone, on the other hand, provides a secure execution environment within a processor that allows sensitive tasks to be run in a separate environment from the rest of the system [1]. ARM TrustZone can protect the system from both external and internal threats by ensuring that only authenticated and verified code is allowed to run within this secure environment [1].

SecFortress approaches security differently. Its goal is to protect the hypervisor by isolating it from the rest of the operating system [2]. SecFortress ensures that an attack on one VM or the outer operating system does not compromise the hypervisor or any other VMs by providing each VM with its own dedicated hypervisor box and preventing direct interaction between the hypervisor and the outer operating system [2]. Fig. 2 presents the architecture of SecFortress.

Each of these technologies offers a distinct solution for securing virtualized environments, but their full potential may be realized only when they are integrated into a unified security framework. The purpose of this paper is to investigate the integration of these technologies into a multi-layered security solution for virtualized environments, with the goal of providing comprehensive protection against both external and internal threats [3].

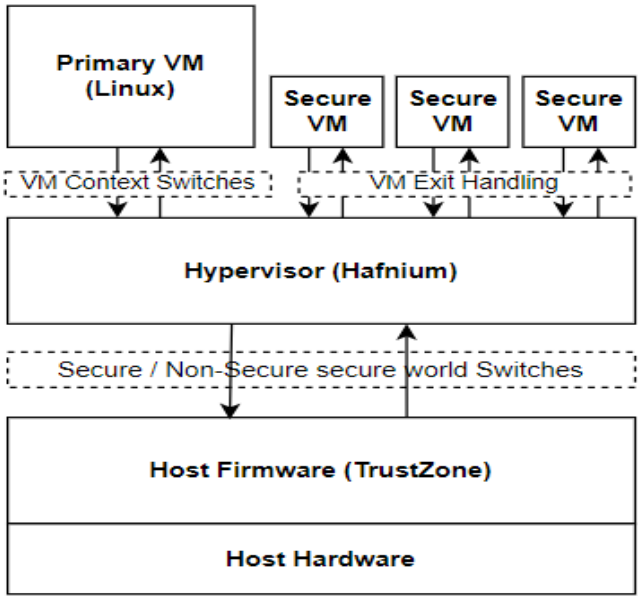


Fig. 1. Hafnium VM configuration.

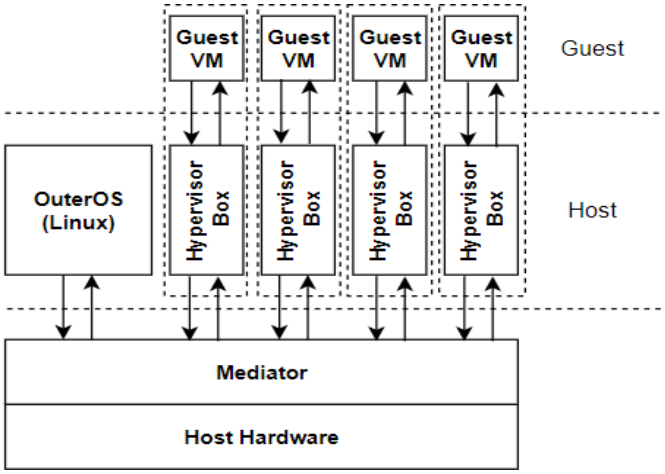


Fig. 2. SecFortress architecture.

The complexities of integrating various security mecha isms, the potential performance overheads of multiple security layers, and the requirement for ongoing updates to address new security vulnerabilities all present significant challenges to the implementation of such a framework. These challenges, however, can be addressed with careful planning and design, paving the way for a more robust and secure computing environment in the virtualization era.

II. Background

Virtualization, enabled by lightweight kernels or hypervisors, has transformed computing by allowing multiple virtual machines (VMs) to run on a single physical host [5]. This operation is enabled and managed by a hypervisor, a key component of the virtualization stack. Despite playing an important role in resource management and VM isolation, hypervisors are vulnerable to security threats from both guest VMs and the host environment [5]. As a result, the design and implementation of secure, lightweight hypervisors are critical to protecting VMs and ensuring system security [5].

The ARM TrustZone is a hardware-based security extension for ARM processors [1], and Hafnium is a lightweight security isolation layer for virtual machines on ARM platforms [5]. The TrustZone technology creates a Trusted Execution Environment (TEE) by partitioning the system into secure and non-secure worlds, whereas Hafnium isolates VM memory and provides a unique security-focused virtualization solution [1]. Hafnium's minimalist design reduces potential attack surfaces, allowing it to be a lightweight hypervisor focused on memory isolation between VM instances while leaving performance and availability guarantees to the host OS [5].

The nested kernel concept takes a different approach to security by embedding a small, lightweight, and isolated kernel within a larger one [2]. This strategy achieves logic isolation by tracking all changes to the virtual-to-physical mapping and removing sensitive instructions from untrusted components, protecting physical memory and lowering the Trusted Computing Base (TCB) in complex systems [2].

Securing our digital infrastructure remains critical in the era of virtualization and cloud computing. Building upon the strengths of Hafnium, ARM TrustZone, and SecFortress, this paper proposes a new multi-layered security strategy to fortify security at both the VM and hypervisor levels, thereby protecting against both internal and external threats. While promising, the combination of these technologies presents certain challenges, including the complexity of managing the various security mechanisms, potential performance overhead due to multiple security layers, and the necessity for continuous updates to counter newly discovered security vulnerabilities [3, 4].

The successful integration of these security solutions into an organization's infrastructure necessitates careful design, comprehensive implementation strategy, and meticulous planning. Despite the obstacles, such an approach holds the potential to significantly enhance the security posture of virtualized and cloud computing environments, making them more resilient against potential attacks [6, 7].

As we forge ahead, extensive testing, performance optimization, and continuous updates will play pivotal roles in overcoming these challenges. By harnessing the unique advantages provided by Hafnium, ARM TrustZone, and SecFortress, we can lay the groundwork for a more secure, robust, and resilient virtualized environment. The path to achieving this goal is strewn with difficulties, but by working collaboratively, we can hope to stay one step ahead of evolving cybersecurity threats and secure our virtualized infrastructure effectively [7].

III. Related Works

Various approaches have been proposed to secure the run-time of hypervisors. HyperLock and DeHype, for instance, deconstruct KVM by assigning a separate isolated hypervisor instance to each VM, similar to the isolated context each VM has in SecFortress [4]. However, unlike SecFortress, they don't protect the hypervisor against a compromised host OS. SecFortress also differs from systems like MultiHype, which supports running multiple hypervisors on a single physical

platform, in that it ensures a smaller Trusted Computing Base (TCB) and stronger isolation by creating a single hypervisor box for each VM [8].

Nexen, SecVisor, and SeL4, along with SecFortress, utilize the nested kernel to reconstruct the virtualization platform [9]. However, they have their limitations; Nexen doesn't consider the security vulnerabilities in its shared service domain, SecVisor focuses on kernel integrity protection within guest VMs rather than isolation among different VMs, and SeL4, being a formally verified microkernel, doesn't include common OS components [9]. Another memory isolation implementation, Hyper Wall, requires support from specific hardware like FPGA, contrasting with SecFortress's ability to be deployedon commercial x86 platforms [9].

Hardware-based defense technologies have also emerged. Intel TDX, for instance, isolates VMs from the hypervisor by adding a secure arbitration mode [10]. AWS Nitro Enclaves and Arm CCA offer different approaches to creating isolated VM execution environments [11]. However, these systems either focus on protecting the guest from an untrusted hypervisor or applications rather than the entire VM, which differs from SecFortress's target of bidirectional isolation protection between VMs and the hypervisor.

In this paper, we have furthered the research into secure HPC OS/Rs by presenting preliminary results of our approach and an initial proof-of-concept implementation [12]. We have identified several potential research directions, such as evaluating our approach on more realistic systems and workloads, designing I/O mechanisms that maintain secure system isolation without imposing significant performance overheads, and investigating dynamic partitioning approaches for secure partitions and VM images [12].

Also, while we have used the Hafnium hypervisor as a starting point for secure virtualization in HPC, we are still evaluating its long-term suitability [12]. The necessary modifications to support HPC workloads, the need for a potential new hypervisor architecture tailored to HPC environments, and the upcoming ARM platform (ARMv9) which introduces significant security, isolation, and trusted computing features, are all factors that could impact the direction of future researchin this area [11, 12].

IV. Proposed Model

A. Design

The TrustZone-Assisted SecFortress solution is a pain stinkingly designed architecture aimed to improve security in a virtualized environment. The innovative design combines TrustZone technology's robust isolation capabilities with the SecFortress hypervisor's flexible and comprehensive security services. As a result, hypervisors and their associated virtual machines benefit from a powerful combination of hardware and software-enforced security mechanisms.

The secure boot mechanism is at the heart of our design. We ensure a trustworthy startup process by utilizing Trust- Zone technology, which allows only authenticated software to launch. This significantly reduces potential threats from unauthorized or malicious software, allowing the system to operate in a trusted state from the start.

Our solution embeds a trust anchor within TrustZone's secure world to establish a root of trust within the system. This provision has important implications for improving hypervisor security and supporting other security functions such as cryptographic key management and secure storage. It ensures a higher level of trust in the system, essentially laying the groundwork for all subsequent security protocols to operate on.

Our design addresses the difficult challenge of securely handling interrupts and I/O operations. The SecFortress solution includes a mediator component that acts as a go-between for the hypervisor and the rest of the system. The mediator significantly reduces potential vulnerabilities that could be exploited during these operations by managing and securely handling interrupts and I/O operations.

Inter-VM communication can be a security risk if not properly managed. Our design ensures secure communication between VMs via the SecFortress solution's mediator. This intermediary validates each communication request to ensure it comes from a reliable source before it reaches its intended destination. This feature prevents unauthorized access and potential data leaks, thereby strengthening the system's overall security.

To provide a secure environment for sensitive processes and applications, our solution design makes use of Trust- Zone's hardware-based isolation capabilities to create a secure execution environment within the hypervisor. This strategic inclusion effectively insulates these processes from potential threats in the 'normal' world, thereby protecting the integrity of our secure world.

The comprehensive isolation of hardware and software components is a critical aspect of our design. This is accomplished through TrustZone's hardware-level isolation, which creates two distinct environments: one for the hypervisor and one for the VMs and outerOS. This hardware isolation is supplemented by the SecFortress's mediator's software-level isolation. As a result, our design fortifies the hypervisor's shell, effectively isolating it from the rest of the system and potentialattack vectors.

Our design also prioritizes system integrity and resilience in the face of a variety of potential attacks. TrustZone technology protects the integrity of the hypervisor by preventing unauthorized changes to its code. SecFortress' security services, which validate the integrity of data and communication channels within the system, add to this. In terms of resilience, the secure boot feature ensures that the system starts up in a trusted state, while the isolation provided by TrustZone and the mediator makes the system difficult to compromise, increasingits resistance to potential threats.

The way we design prioritizes secure communication and data handling. SecFortress' mediator is at the heart of all communication between the VMs, the outerOS, and the hypervisor, validating the origin and destination of each communication request. TrustZone technology protects data integrity and confidentiality by isolating sensitive data within a

secure world. This two-pronged approach significantly reduces the risk of data exposure or tampering from malicious processes.

The TrustZone-Assisted SecFortress solution design promotes adaptability. The solution is designed to be compatible with a wide range of hardware platforms and hypervisors with minimal modifications, allowing it to be widely deployed across a wide range of systems. The design's scalability, which is further enhanced by its modularity, enables it to protect systems of various sizes, from individual servers to large data centers, without compromising security.

The TrustZone-Assisted SecFortress solution is also optimized for efficiency and performance. With the SecFortress solution's mediator minimizing overhead in handling communication between the VMs, the outerOS, and the hypervisor, and TrustZone ensuring efficient use of hardware resources, optimal system performance is guaranteed.

The solution design incorporates built-in fail-safe mechanisms to ensure that the system remains secure even if a component fails. For example, if the mediator component detects an error, it activates built-in fail-safe routines to prevent a system-wide failure. Similarly, TrustZone technology ensures that any breaches in the normal world do not impact the secure world, adding an extra layer of security.

The solution is designed to work seamlessly with a wide range of systems, taking into account both modern and older systems that may not have built-in security measures. It is compatible with various system architectures and hypervisor types, giving it versatility in securing various systems. The design also includes recovery and maintenance measures capable of detecting potential security threats, initiating protective actions, and allowing for easy system updates and patches. This not only ensures long-term security but also improvesthe system's overall security capabilities.

The TrustZone-Assisted SecFortress solution is designed for the future, combining scalability, robustness, compatibility, efficiency, and fail-safe mechanisms. Its comprehensive and adaptable design supports a variety of system architectures and hypervisors, as well as a wide range of virtual machines and can handle increasing load and traffic. Because of its adaptability, it is an effective security solution for complex virtualized environments. Furthermore, the solution reduces costs while maintaining security by leveraging TrustZone's existing hardware security features and SecFortress's minimalistic design. Also, our design considers future developments in the cybersecurity landscape. It is easily upgradable to handle new types of security threats or to incorporate advancements in hypervisor and virtualization technology. This foresight distinguishes our solution, making it a comprehensive, robust, and scalable option for securing hypervisor environments today and in the future.

The TrustZone-Assisted SecFortress solution is divided into several layers or levels, as illustrated in Fig. 3, and are as follows:

Level 1 - Hardware Layer: This includes the actual hardware platform. At this level, the TrustZone technology creates two distinct environments: the secure world' and the 'normal world'. The most sensitive processes and data reside in the secure world.

Level 2 - TrustZone Technology: TrustZone technology primarily operates at this level, providing fundamental hardware- based isolation capabilities and establishing a root of trust within the system. It also manages the secure boot mechanism, which ensures that only authenticated software can run.

Level 3 - Mediator Component: At this level, the mediator component operates, providing a layer of software-based isolation on top of TrustZone's hardware-based isolation. It also handles interrupts, I/O operations, and inter-VM communications securely, acting as a liaison between the hypervisor and the rest of the system.

Level 4 - Hypervisor: At this level, the SecFortress hypervisor operates. It communicates directly with TrustZone technology to take advantage of hardware-based isolation capabilities for enhanced security. It is also the location of the secure execution environment for sensitive processes and applications.

Level 5 - Virtual Machines (VMs) and outer OS/Primary VM using Kitten lightweight kernel: At this level, virtual machines and the outer / Primary VM operating system operate. They communicate with the hypervisor and the mediator, enabling secure communication and operations.

These components interact in a variety of ways to form a comprehensive, secure, and adaptable system. TrustZone technology, for example, serves as the foundation for system security by creating a secure execution environment and managing the secure boot process. The SecFortress hypervisor, in turn, relies on this secure environment to run VMs and manage their communication via the mediator component.

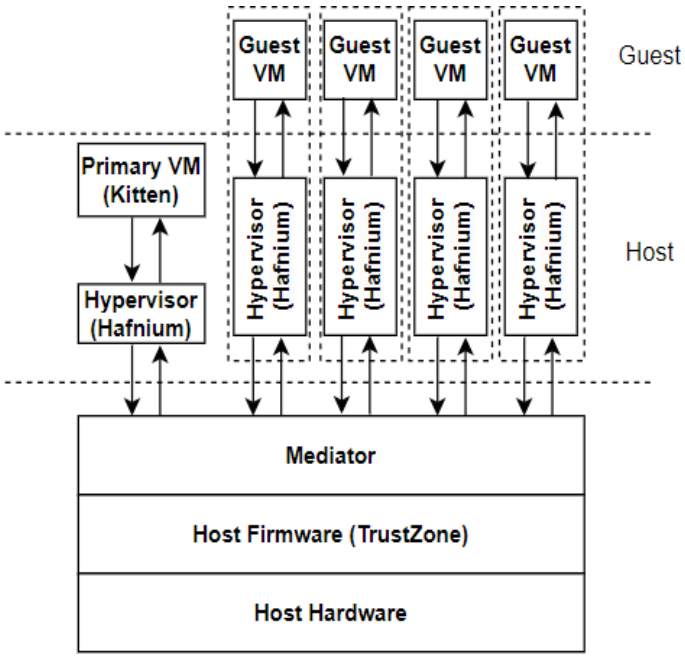


Fig. 3. Proposed model trustzone-assisted secfortress solution architecture.

B. Implementation

The implementation of the TrustZone-Assisted SecFortress solution begins with configuring the ARM-based System-on-a-Chip (SoC) to use TrustZone technology. This process begins with the system boot-up, during which the Secure World is configured before any other system components are initialized.

To establish a clear demarcation between the Secure and Non- Secure Worlds, the memory layout and IRQ controllers must be carefully adjusted. TrustZone's Monitor Mode is set up as an extra execution level within the Secure World to host the mediator software, which manages secure transitions between the two worlds.

After properly configuring the Secure and Non-Secure Worlds, we proceed to integrate the Hafnium hypervisor into the system's Non-Secure World. The isolation capabilities of Hafnium are critical to the system's security. It is intended to create isolated partitions for each virtual machine (VM), effectively isolating them from one another and from the hypervisor itself. The configuration of Hafnium is meticulously adjusted during this phase to manage VMs, enforce memory access policies, and control inter-VM communication. Fig. 4 depicts the memory access view of each SecFortress component. For example, the left-most large box represents the VM's memory access view. That is, each VM has complete access to its own memory but no access to the memory of others. The mediator has access to all memory and controls the page tables to determine the memory view of each component. Illegal memory accesses across components will result in pagefaults, which will be detected by the mediator.

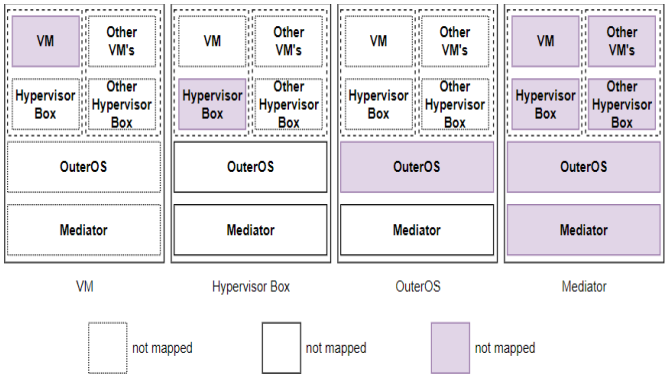


Fig. 4. Memory access view of VM, Hypervisor Box, outerOS, and mediator.

We integrate the Kitten lightweight kernel into the Hafnium hypervisor after it has been established. The Kitten kernel is the foundation of our hypervisor layer. It is designed to be extremely efficient in high-performance computing environments. The integration of the kernel entails prioritizing the optimization of memory management, task scheduling, and I/O operations, which improves overall system performance.

We begin development of the mediator software after successfully integrating the Hafnium hypervisor and Kitten kernel. The use of programming languages compatible with the ARM architecture and the TrustZone environment is required for this task. Memory protection, instruction protection, and control flow management are among the critical security services enforced by the mediator software. It is also intended to intercept and manage VM exits, ensuring safe context switchesbetween system components.

The addition of VM image integrity checks and encryption measures improves system security even further. Before each startup, VM image integrity checks are performed to ensure that the VM images have not been tampered with. Furthermore, the VM images are encrypted with strong encryption algorithms, ensuring the data stored within the VMs' security. The cloud provider's key management service secures the encryption keys.

We chose to implement software updates for the TrustZone- Assisted SecFortress virtualization layer via offline install packages in order to maintain system security. This method gives us more control over the software versions that are running on the system. System administrators install these updates manually, lowering the risk of online threats.

We employ several strategies to reduce performance overhead, ensuring that the system remains operational and does not impair overall performance. The use of a nested MMU extension to create memory protection domains, efficient memory management with the mediator's internal allocator, hard coding of sensitive instruction entry points during system bootup, and the minimization of memory mapping updates within the hypervisor box are among these strategies [13].

Implementing the TrustZone-Assisted SecFortress solution is a meticulous process that includes system initialization, hypervisor setup, kernel integration, mediator software development, VM security measures, software update procedures, and performance overhead reduction strategies. The end result of this comprehensive implementation process is a robust and secure virtualization environment suitable for secure cloudcomputing applications [13, 14].

V. Results and Evaluation

This section describes the TrustZone-Assisted SecFortress solution's results thorough evaluation and security analysis, including protection against mediator tampering, the level of isolation and confidentiality, denial-of-service (DoS) mitigation, performance assessment, and the results of practical attack scenarios.

A. Security Analysis

The TrustZone-Assisted SecFortress solution's comprehensive security analysis begins with protection against mediator tampering. Because the mediator plays such an important role in the system, its integrity is critical. As a result, its security is ensured by secure boot and code integrity checks during system boot, which is a process that protects the mediator's code by write-protecting it, preventing attackers from tampering with it. This writes protection extends to dynamic checks within non-TCB components, protecting sensitive instructionsand code from bypass attacks.

Effective isolation and confidentiality are critical security objectives of the TrustZone-Assisted SecFortress solution, which it achieves by isolating hypervisor boxes from the

outerOS and other hypervisor boxes. Memory access control and paging mechanisms are used to restrict access to sensitive memory regions. OuterOS is prevented from accessing the memory of the hypervisor boxes by unmapping hypervisor box-related memory regions in the kernel master page table. Additionally, zeroing physical pages before assigning them to hypervisor boxes strengthens the integrity defense against attacks.

In terms of Denial-of-Service (DoS) attacks, the TrustZone-Assisted SecFortress solution defends itself by meticulously checking the memory and registers of guest VM states before returning control to them. While this comprehensive measure reduces the possibility of crashes or interference with other VMs by addressing any state mishandling, it does not completely prevent DoS attacks from originating from the host.

B. Performance Evaluation

A performance evaluation was carried out to determine the overhead and efficiency of the combined solution. The solution was tested using various performance benchmarks on Pine A64-LTS SBC and Ubuntu 18.04.5 with Linux 5.2 for Intel VT platforms. The benchmarks focused primarily on CPU and memory performance, with consistent results demonstrating minimal overhead for virtualization and secure isolation. Even when lightweight kernels and ARM TrustZone-based mechanisms were used, there was no significant performance degradation.

A series of tests, including Stream and Random-Access micro-benchmarks, were used to assess memory and I/O performance. These tests revealed that Hafnium and ARM TrustZone introduced minor overhead in some scenarios, but overall performance was satisfactory.

The combined solution was then put through its paces with application performance benchmarks like the HPCG mini-app and a subset of the NAS Parallel Benchmark suite. These tests demonstrated that the solution was capable of running a full mini-app benchmark with minimal overhead and of providing secure isolation for HPC applications and workloads.

C. Practical Attack Evaluation

The TrustZone-Assisted SecFortress solution was subjected to realistic attack scenarios in order to validate the effectiveness of the security measures. CVE analysis for Linux/KVM vulnerabilities was included in these scenarios. The evaluation revealed that the security mechanisms of the solution were effective in preventing privilege escalation, information leakage, memory corruption, and denial-of-service attacks from compromised outerOS or malicious VMs. In essence, the combined solution's isolation was critical in mitigating the impact of compromised components and safeguarding sensitive data andmemory regions.

Finally, the TrustZone-Assisted SecFortress solution combines lightweight kernels, Trusted Execution Environments (TEEs), the Hafnium hypervisor, and ARM TrustZone to achieve robust security isolation in virtualization environments. It effectively addresses both security and performance concerns, thereby assisting in the development of more secure and efficient virtualized systems. Comprehensive security and performance evaluations confirm its resistance to common attacks and ability to handle high-performance computing workloads with minimal overhead. As virtualization technology evolves, it promises to be a solid foundation for secureand high-performance virtualization environments.

VI. Conclusion

In this paper, we presented an enhanced to secure hypervisor runtime and a new use case for Lightweight Kernels as resource management services in securely isolated HPC systems. As part of this effort, we integrated the ARM64-ported Kitten LWK with the Hafnium hypervisor in our SecFortress solution to support secure virtual machine instances on a compute node. SecFortress partitions the virtualization platform strategically into a trusted mediator, an isolated outerOS, and multiple restricted hypervisor box instances, improving security isolation in high-performance computing platforms. The new approach prevents the outerOS from accessing the hypervisor's memory, with each hypervisor box instance limited to the least amount of memory access. As a result, even if one instance is compromised, the integrity and confidentiality of other instances are not jeopardized. We have provided an initial proof of concept implementation and preliminary evaluation, which show that our approach has no significant performance overheads on a variety of HPC benchmarks. SecFortress' experimental results show that it can defeat exploits against the host OS and VMs with negligible performance overhead, implying that SecFortress could be an effective solution for improving both virtual machine security and performance. This work has also identified a number of future challenges and made a compelling case for security isolation and trusted computing as key features of next-generation HPC platforms. Fully supporting security isolation in a scalable and performant manner will most likely pose a significant challenge for HPC OS/R architectures, necessitating future research. Our integrated solution, we believe, provides a promising foundation for the evolution of more secure and efficient virtualized systems.

References

[1] "Arm trustzone technology." https://developer.arm.com/technologies/trustzone, accessed August 2023.

[2] Q. Zhou, X. Jia, S. Zhang, N. Jiang, J. Chen, and W. Zhang, "Secfortress: Securing hypervisor using cross-layer isolation," in *2022 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 212-222, IEEE, 2022.*

[3] H. Nguyen, Y. Tan, and X. Gu, "Pal: P ropagation-aware a nomaly l ocalization for cloud hosted distributed applications," in *Managing Large-scale Systems via the Analysis of System Logs and the Application of Machine Learning Techniques*, pp. 1–8, 2011.

[4] W. Shi, J. Lee, T. Suh, D. H. Woo, and X. Zhang, "Architectural support of multiple hypervisors over single platform for enhancing cloud computing security," in *Proceedings of the 9th conference on Computing Frontiers*, pp. 75–84, 2012.

[5] "Hafnium hypervisor." https://www.trustedfirmware.org/projects/hafnium, Accessed August 2023.

[6] C. Meyers, "The biggest cloud breaches of 2019 and how to avoid them for 2020." Lacework Editorial, December 2019.

[7] M. Gontovnikas, "The 11 biggest data breaches of 2020." Auth0 Blog, 2020.

[8] N. Dautenhahn, T. Kasampalis, W. Dietz, J. Criswell, and V. Adve, "Nested kernel: An operating system architecture for intra-kernel privilege separation," in *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and*

Operating Systems, pp. 191–206, 2015.

[9] Z. Mi, D. Li, H. Chen, B. Zang, and H. Guan, "(mostly) exitless vm protection from untrusted hypervisor through disaggregated nested virtualization," in *Proceedings of the 29th USENIX Conference on Security Symposium*, pp. 1695–1712, 2020.

[10] Intel, "Intel trust domain extensions." https://software.intel.com, 2020.

[11] Amazon, "AWS Nitro Enclaves." https://docs.aws.amazon.com/enclaves/, 2020.

[12] M. Boubakri, F. Chiatante, and B. Zouari, "Open portable trusted execution environment framework for risc-v," in *2021 IEEE 19th International Conference on Embedded and Ubiquitous Computing (EUC)*, pp. 1–8, 2021.

[13] Y. Wu, Y. Liu, R. Liu, H. Chen, B. Zang, and H. Guan, "Comprehensive vm protection against untrusted hypervisor through retrofitted amd memory encryption," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 441–453, IEEE, 2018.

[14] A. M. Azab, P. Ning, Z. Wang, X. Jiang, X. Zhang, and N. C. Skalsky, "Hypersentry: enabling stealthy in-context measurement of hypervisor integrity," in *Proceedings of the 17th ACM conference on Computer and communications security*, pp. 38–49, 2010.

Efficient Evaluation of SLAM Methods and Integration of Human Detection with YOLO Based on Multiple Optimization in ROS2

Hoang Tran Ngoc*, Nghi Nguyen Vinh, Nguyen Trung Nguyen, Luyl-Da Quach
FPT University, Can Tho 94000, VietNam

***Abstract*—In the realm of robotics, indoor robotics is an increasingly prominent field, and enhancing robot performance stands out as a crucial concern. This research undertakes a comparative analysis of various Simultaneous Localization and Mapping (SLAM) algorithms with the overarching objective of augmenting the navigational capabilities of robots. This is accomplished within an open-source framework known as the Robotic Operating System (ROS2) in conjunction with additional software components such as RVIZ and Gazebo. The central aim of this study is to identify the most efficient SLAM approach by evaluating map accuracy and the time it takes for a robot model to reach its destinations when employing three distinct SLAM algorithms: GMapping, Cartographer SLAM, and SLAM_toolbox. Furthermore, this study addresses indoor human detection and tracking assignments, in which we evaluate the effectiveness of YOLOv5, YOLOv6, YOLOv7, and YOLOv8 models in conjunction with various optimization algorithms, including SGD, AdamW, and AMSGrad. The study concludes that YOLOv8 with SGD optimization yields the most favorable outcomes for human detection. These proposed systems are rigorously validated through experimentation, utilizing a simulated Gazebo environment within the Robot Operating System 2 (ROS2).**

***Keywords*—Indoor robotic; SLAM; ROS2; Robot model; Human detection; YOLO**

I. Introduction

With the rapid advancement of Artificial Intelligence (AI) and the continuous evolution of sensor technologies, coupled with the introduction of the Robot Operating System (ROS) and its latest iteration, ROS 2 [1], the development of indoor robots has become more accessible than ever before. Mobile robots have yielded substantial economic benefits across various sectors, including industry, warehousing, and logistics [2]. Notably, robots are no longer confined to industrial applications; they are increasingly being deployed in the realm of mental healthcare, where they assist therapists in caring for the elderly and children struggling with depression [3]. In light of these developments, our objective is to create a pet robot designed to serve the purposes mentioned above. Within the realm of developing such a robot, we consider two fundamental tasks of paramount importance: Navigation and Human Detection. This article aims to compare different methods and put forth the most optimal approaches for accomplishing these critical tasks.

Sensors play a crucial role in how robots and autonomous vehicles perceive their surroundings. The choice and installation of sensors have a significant impact on the specific results of observation and also influence the complexity of SLAM problems. Based on the primary type of sensor used, SLAM can be categorized into Visual-SLAM and LiDAR-SLAM. Visual SLAM remains a particularly challenging task due to inherent difficulties. Moreover, vision cameras struggle to extract features from texture less areas, which limits the applicability of Visual SLAM. On the other hand, LiDAR SLAM primarily relies on LiDAR technology for environmental sensing. The relative movement and pose changes of the laser radar are determined by comparing point clouds captured at different moments. LiDAR SLAM offers advantages in terms of stability, simplicity, precise map data, and lower computational requirements compared to Visual SLAM. Numerous investigations have explored LIDAR-SLAM techniques in the literature [4]-[9].

This research primarily centers on evaluating Slam techniques through the utilization of ROS2, with a primary emphasis on assessing their performance based on the resulting maps. The evaluated methods fall within the category of 2D LIDAR Slam techniques designed for indoor settings, including GMapping, Cartographer SLAM, and Slam-Toolbox. In addition, we have integrated a LIDAR-SLAM technique with a human detection system. The methods we evaluated and tested for this integration include YOLOv5 [10], YOLOv6 [11], YOLOv7 [12] and YOLOv8 [13], using multiple collected datasets. Ultimately, this research aims to develop an optimized system for indoor robotic operations, ensuring precise localization and robust human detection capabilities within indoor spaces. This system is designed to facilitate avoidance maneuvers and ensure safety during robot operations.

II. Related Work and Our System

A. Related Work

Recent advancements in the field of mobile robot navigation have enabled robots to operate effectively in various environments, including warehouses, retail stores, and crowded pedestrian areas. A plethora of navigation solutions have been proposed to address these challenges [14]. One of these solutions introduced a navigation system and environmental representation that utilized 3D data obtained from tilting 2D laser scanners for navigation. Initially, conventional methods such as A* and DWA were employed

*Corresponding Author.

for path planning and obstacle avoidance. However, the prevalence of cost-effective 3D depth cameras and laser scanners has gradually overshadowed the use of tilting 2D laser scanners. Some aspects of these approaches were customized to account for the specific geometry, limited accuracy, and characteristics of 2D tilting laser setups. Furthermore, the emergence of sparse multi-beam laser scanners has rendered traditional raycasting methods ineffective in clearing free space, especially in dynamic environments. In response to this challenge, Sparse Traversability Volume (STVL) has emerged as a scalable alternative suitable for various types of sparse and long-range sensors, replacing traditional techniques effectively [15].

The Robot Operating System (ROS) Navigation has historically been one of the most popular navigation solutions built on top of ROS. However, with the introduction of ROS2, Navigation2 was developed as a successor to build upon the success of ROS Navigation. Navigation2 incorporates a behavior tree for orchestrating navigation tasks and utilizes novel methods designed to handle dynamic environments, making it applicable to a wider range of modern sensors. SLAM is a critical technology in mobile robotics, allowing robots to map unknown environments while simultaneously determining their own position based on the created map. Several SLAM algorithms have been proposed, categorized into two groups: earlier algorithms employing Bayes-based filter approaches like GMapping [16], and newer ones using graph-based methods such as Cartographer [17], Karto SLAM [18], and Slam Toolbox [19].

For our human detection task, we have chosen to leverage the YOLO (You Only Look Once) framework for several compelling reasons. Among the numerous object detection algorithms available, YOLO stands out due to its exceptional combination of speed and accuracy. It excels in rapidly and accurately identifying objects within images, making it an ideal choice for real-time applications. YOLO has gained prominence as a central system for real-time object detection in various domains, including robotics, autonomous vehicles, and video monitoring, demonstrating its reliability and versatility in a wide range of applications [20]. The YOLO family of object detection models has undergone a series of iterations, evolving from the original YOLOv1 to the most recent YOLOv8. Each iteration has built upon the foundation of its predecessors, aiming to address limitations and enhance overall performance in object detection tasks. However, in our research paper, we will specifically focus on evaluating human detection using YOLO versions ranging from YOLOv5 to YOLOv8. This selective approach allows us to assess the advancements and capabilities of these more recent YOLO iterations in the context of human detection, which is a pivotal aspect of our study.

B. Our System

In this study, we are using an open-source robotic middleware ROS 2. Inheriting from ROS, the libraries provided by ROS 2 used in this project make robotic programs and development more flexible and easier. ROS 2 also offers the capability to obtain hardware abstractions of the robot model, encompassing sensors, motors, and actuators which can be used for navigation and are accessible through URDF and XACRO files. Additionally, we utilize other open-source software in conjunction with ROS 2, namely Gazebo and RVIZ 2. Gazebo functions as simulation software, enabling the virtual simulation of robots through the use of plugins. This allows us to construct a simulated environment represented as URDF and WORLD files, facilitating the simulation of the robot within the Gazebo world, as depicted in Fig. 1. In this article, the robot model used is a mobile robot with two wheels, equipped with several sensors including a 360° 2D LIDAR and a camera. These sensors play a crucial role in tasks like navigation and human tracking. The detailed illustration of this robot model is shown in Fig. 2. RVIZ 2 serves as a data visualizer tool for robot data, presenting various data types, including laser scans, maps retrieved from the map server, and grid displays. As well as ROS, RVIZ 2 also provides navigation goals and poses estimation functionalities, which are inherited from RVIZ in order to accomplish the Robot's navigation in the ROS 2 virtual environment.

In order to make the autonomous robot navigate in its environment safely without encountering collisions with either static or moving obstacles, the assistance of SLAM is required. Different SLAM algorithms can be used for mapping such as Karto SLAM, Cartographer SLAM, Gmapping SLAM, and Hector SLAM, but in this paper, we want to compare the package provided by ROS 2 (SLAM_toolbox).

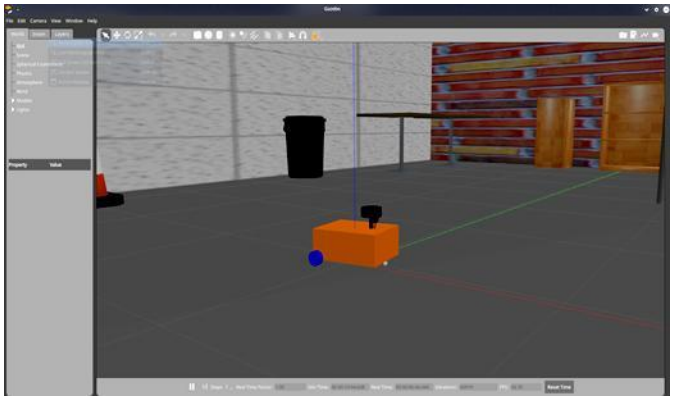


Fig. 1. Robot model in Gazebo virtual world.

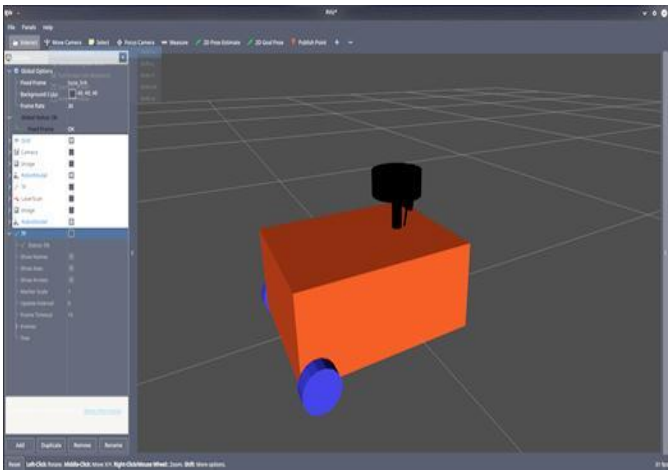


Fig. 2. Robot description on Rviz 2.

III. METHODOLOGY

A. LIDAR 2D SLAM Algorithms

In this section, an analysis of SLAM algorithms in the ROS2 environment will be conducted. Among the available open-source laser scanner SLAM algorithms, Cartographer is a graph-based algorithm that manages a graph representing robot poses and features. It offers resource efficiency, especially for constructing large-scale maps. It consists of a front-end responsible for tasks like scan matching, trajectory building, and submap generation, as well as a back-end that handles loop closure procedures, using the Google Ceres graph solver. Cartographer provides a pure localization mode for users with existing maps and supports data serialization for storing processed sub-maps. However, it has encountered challenges, such as the discontinuation of maintenance and support by Google, leading to its abandonment. This algorithm divides a large single map into smaller sub-maps by integrating two separate 2D SLAM techniques. One method focuses on local operations, while the other deals with global aspects, and both employ a LiDAR sensor. These two methods are optimized independently. In the local SLAM process, sub-maps are created by collecting and arranging data, involving the alignment of multiple scans relative to the initial position. These sub-maps are grids with defined resolutions, with each grid point indicating occupancy probability based on prior measurements updated as new sub-maps are generated. An algorithm optimizes sub-map positioning for alignment, aiding extrapolation. The second part, global SLAM, leverages feedback from these sub-maps, which are associated with robot positions. This enhances maps and reduces accumulated SLAM errors, a process known as loop closure [24]. The well-known optimization technique called Spare Pose Adjustment (SPA) [23] [38] is utilized in Cartographer SLAM, and a map-scanner is activated whenever a sub-map is generated to close the loop and incorporate that sub-map into the graphic. Two formulas are provided to determine whether a cell is classified as busy, empty, or transitioning to an empty state within a map cell, enhancing comprehension.

$$M_{new}(cell) = F^{-1}(F(M_{old}(cell)\ F(f_{hit}))) \quad (1)$$

where: $M_{old}(cell)$ is the old probability of the cell which could be an error, f_{hit} is the probability function that represents a map cell is busy, and $F = \frac{F}{1-F}$.

Scan matching process goes through a minimization of the following function:

$$\arg_{\delta}\min \sum_{k=1}^{K} (1\text{-}M_{smoothen}(T_{\delta}s_k))^2 \quad (2)$$

where, $M_{smoothen}$ represents the value of a cell that has been smoothed using its neighboring values, s_k denotes the laser scan reading involves to the cell, T_{δ} is the matrix transformation that displaces the point h_k to δ, and δ is the posture vector $(\delta_x, \delta_y, \delta_\theta)$.

Additionally, Cartographer may struggle to create suitable maps for annotation and localization when integrated with other robotic platform localization software that lacks exceptional odometry. Its complexity can hinder modifications and resolution of seemingly straightforward issues, limiting its suitability for many applications.

GMapping Slam is widely used within the Robot Operating System (ROS) and stands out for its frequent adoption. GMapping employs the Rao Blackwellized Particle Filter (RBPF) [21] technique for map generation. However, it's important to note that GMapping has limitations when applied to large environments and struggles with precise loop closure in industrial-scale spaces. The idea of the RBPF for SLAM, first introduced by Murphy [22] in 1999, is to estimate the joint posterior $p(x_{1:t}, m \mid z_{1:t}, u_{1:t-1})$ of the map where m is the map and the trajectory $a_{1:t} = a_1, \ldots, a_t$ of the robot. The $z_{1:t} = z_1, \ldots, z_t$ is the given observations and the odometry measurements is $o_{1:t-1} = o_1, \ldots, o_{t-1}$. These both can be obtained by the robot's data. The fully RBPF factorization for SLAM is representing below:

$$p(m, a_{1:t} \mid z_{1:t}, o_{1:t-1}) = p(m \mid a_{1:t}, z_{1:t})\ p(a_{1:t} \mid z_{1:t}, o_{1:t-1}) \quad (3)$$

Nonetheless, filter-based approaches such as GMapping encounter difficulties when attempting to achieve seamless reinitialization across multiple sessions.

The SLAM Toolbox is a versatile mapping solution that efficiently covers large areas using standard mobile Intel CPUs commonly found on robots. It simplifies space mapping through automation and supports session serialization, enabling users to easily improve existing maps. What makes it unique is its preservation of complete raw data and pose-graph, enabling various innovative tools like manual pose-graph manipulation and kinematic map merging. The SLAM Toolbox provides three primary operational modes: synchronous mapping for high-quality maps, asynchronous mapping for real-time performance, and pure localization for adapting to dynamic environments.

Significant enhancements to the OpenKarto SLAM library [19], [38] have boosted its speed and adaptability, making it a valuable tool for robot navigation and mapping tasks. The SLAM Toolbox has been effectively integrated, tested, and utilized on diverse robotic platforms around the world by both professionals in the industry and researchers. It serves as the default SLAM solution in ROS 2, replacing GMapping. Incorporated into the ROS 2 Navigation2 project, it enables real-time positioning in changing environments, facilitating autonomous navigation. Its user-friendly interface empowers both experts and non-experts to map extensive spaces in real time, establishing its significance in robotics and autonomous system development.

To evaluate a Slam algorithm, two criteria such as the accuracy of the generated map based on comparing the ground truth and the map generated or its feature locations, and the time taken for the robot to navigate and get to its destination for each environment are considered. In order to investigate the map's quality, we created three different simulation environments on the Gazebo simulator, and we added the robot model to those simulation worlds that we created. The robot model has laser scanner sensors (LIDAR 360) that supply data crucial for map creation when implementing three

SLAM techniques in RVIZ 2. The data were transmitted to the scan topic for the map, to generate and then visualize it on RVIZ 2; afterward, map topics were activated. The higher the map fits the ground truth the better the map quality. Fig. 3 shows RVIZ 2 and Gazebo when launching our virtual world from their respective directories. In Fig. 4, we illustrate three maps generated by three different SLAM methods within a simulated environment.

These maps are saved in PGM file format and have been compared with ground truth data. To create these maps, the robot navigated through the simulated environment using a control framework called "ros2_control," which is a reimplementation of the "ros_control" framework used in ROS.

After mapping each environment using different SLAM techniques with their default parameters, the resulting maps were saved as YAML files. Notably, we observed that the SLAM_toolbox method exhibited higher accuracy. The map generated by this SLAM method had less noise and better alignment with the ground truth. Detailed evaluations of these SLAM methods will be discussed further in the experimental section.

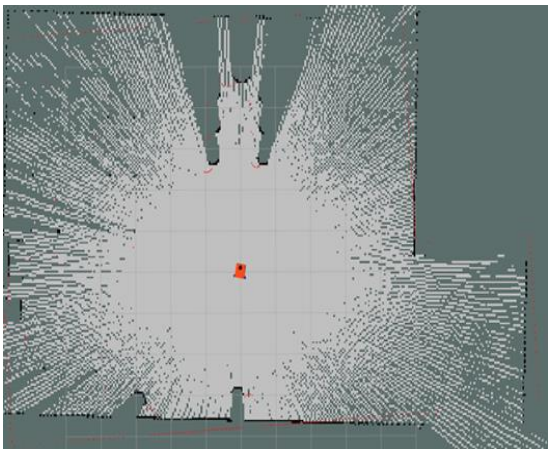


Fig. 3. Robot model creating the map of the environment.

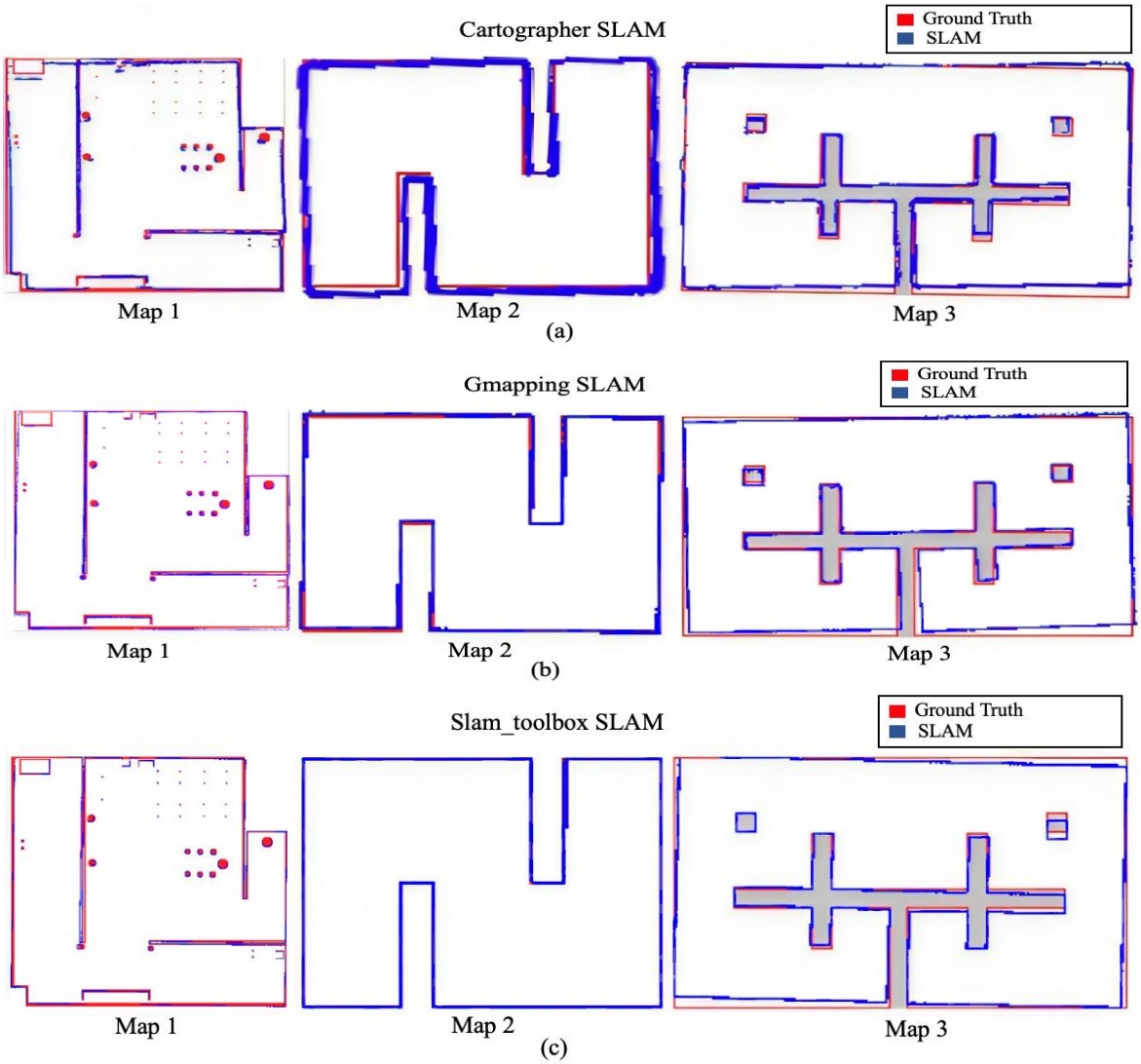


Fig. 4. Generated map results compared to ground truth with: (a) Cartographer SLAM, (b) Gmapping SLAM, (c) Slam_toolbox SLAM.

To enable autonomous navigation of the robot, the robot received a generated map of each algorithm as an input. The AMCL (Adaptive Monte Carlo Localization) algorithm employs a probabilistic localization model that aids the movement of the robot between different positions using a "2D Navigation Goal" provided by RVIZ 2 to reach each destination; and "2D Pose Estimator" that allows the position of the robot model to be set within an environment.

B. Human Detection Algorithm

In this section, we scrutinize and assess the human detection algorithm, aiming to propose an optimal model for indoor robots. This model is designed to help robots in tasks like obstacle avoidance and ensuring operational safety. Among the many algorithms for object detection, the YOLO (You Only Look Once) framework has gained recognition for its exceptional blend of speed and accuracy, enabling rapid and reliable object identification in images. Over time, the YOLO family has gone through multiple iterations, with each new version building upon the previous ones to address shortcomings and enhance performance.

In 2020, Ultralytics introduced YOLOv5 [10]. Unlike YOLOv4, which utilized Darknet, YOLOv5 was developed using Pytorch and incorporated various improvements. YOLOv5 also integrated an AutoAnchor algorithm, a pre-training tool that evaluates and adjusts anchor boxes to better suit the dataset and training parameters, including image size. Initially, it applies a k-means function to dataset labels to establish starting conditions for a Genetic Evolution (GE) algorithm.

In September 2022, the Meituan Vision AI Department introduced YOLOv6 [11]. Its network design features an efficient backbone PAN topology neck, which utilizes RepVGG or CSPStackRep blocks. It uses an efficient decoupled head with a hybrid-channel strategy. Furthermore, a new quantization technique was proposed to achieve faster and more accuracy.

YOLOv7 [12] was published in July 2022. At the time of its release, it outperformed all known object detectors in terms of speed and accuracy, achieving frame rates ranging from 5-160 FPS. It was trained only on the MS COCO dataset without pre-trained backbones. YOLOv7 brought forth numerous architectural modifications and a range of improvements, which boosted accuracy without affecting inference speed, albeit increasing training time.

In January 2023, Ultralytics, the organization responsible for YOLOv5, released YOLOv8 [13]. It has the capability to handle multiple vision-related tasks, including classification, object detection, segmentation, pose estimation, and tracking.

In Table I, we compare the structure and loss functions employed in various versions of YOLO. Fig. 5 illustrates the structure of the human detection model using YOLOv8-N, which we propose for adoption. With its compact size and high accuracy, YOLOv8-N is well-suited for deployment in resource-constrained hardware robots. Our evaluations indicate that YOLOv8-N outperforms other models. The results of our indoor detection capabilities will be presented in the experimental section.

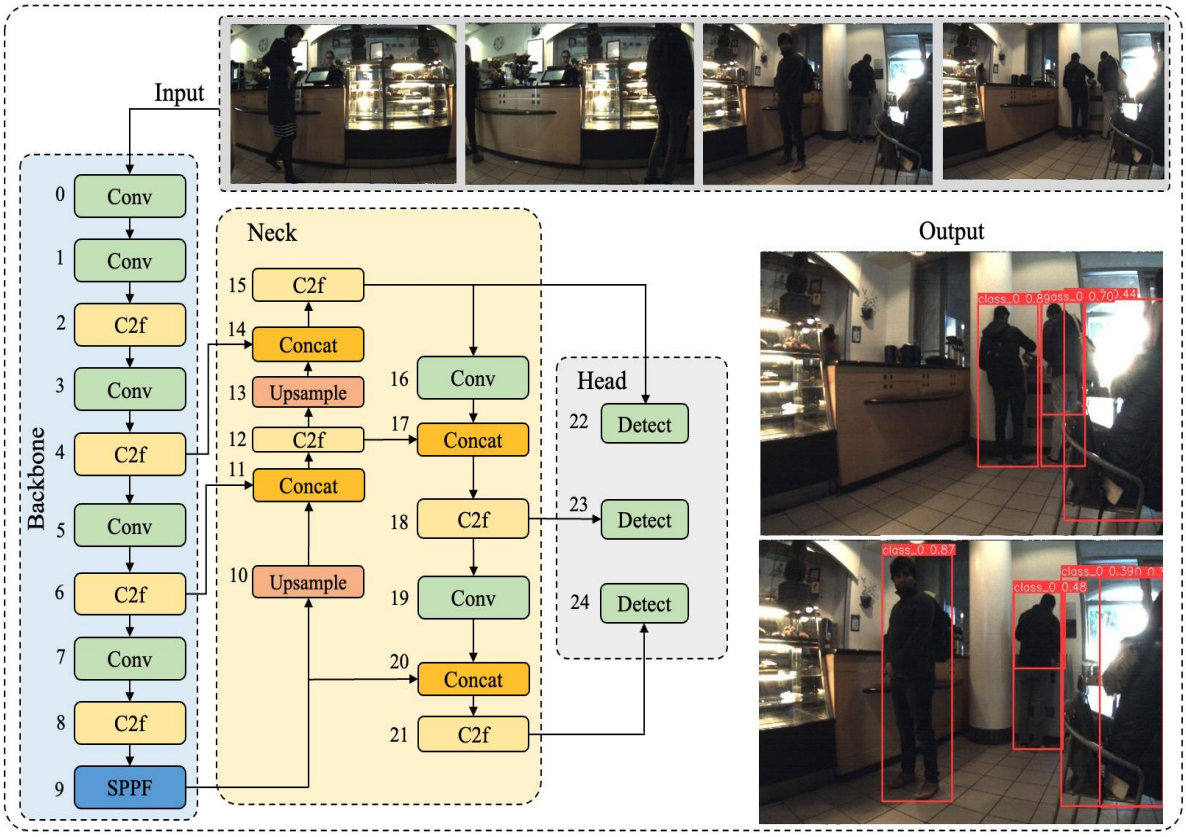


Fig. 5. The structure of the human detection model using YOLOv8-N.

TABLE I. ARCHITECTURE OF YOLO'S VERSION

	YOLOv5-N	YOLOv6-N	YOLOv7-Tiny	YOLOv8-N
Backbone	CSP Darkent53 [30]	RepVGG [31] and CSPRepStack [32]	EELAN [12]	CSPDarkent53 [30]
Neck	PANet	RepPAN	PANet	PAN-FPN
Head	B x (5 +C)	Decoupled Classification, and Detection Head	Lead Head	Decoupled Head
Loss Function	Binary Cross Entropy (BCE) [25], and Logit Loss Function (LLF)[26]	Varifocal Loss (VFL) [26], and Distribution Focal Loss (DFL) [27]	BCE with Focal Loss, and IoU [28]	VFL, DFL Loss, and CIOU [29]
Parameters	1.9M	4.3M	6.2M	3.2M

IV. EXPERIMENTAL RESULTS

A. SLAM Performance and Results

After utilizing the three SLAM techniques, Cartographer, GMapping, and SLAM_toolbox, the map was successfully generated with the highest level of precision using the SLAM_toolbox method, as depicted in Fig. 4. To assess the impact of map accuracy, we recorded the time taken by the robot to navigate and reach its destination in various environments. We divided the time measurement process into three segments for each map, conducted multiple test runs, and subsequently computed the average duration.

The first two phases of testing were conducted on Map_2 and Map_3, as illustrated in Fig. 6. In each of these navigation scenarios, a map generated by one of the three SLAM algorithms was utilized. Table II presents the time taken by the robot to reach its destination on the maps generated by Cartographer SLAM, GMapping, and SLAM_toolbox. For map 2, the goal point coordinates were set as (x = 4.0, y = 2.0), while for Map_3, they were set as (x = 0.0, y = 6.0), with the z-axis being maintained at 0.

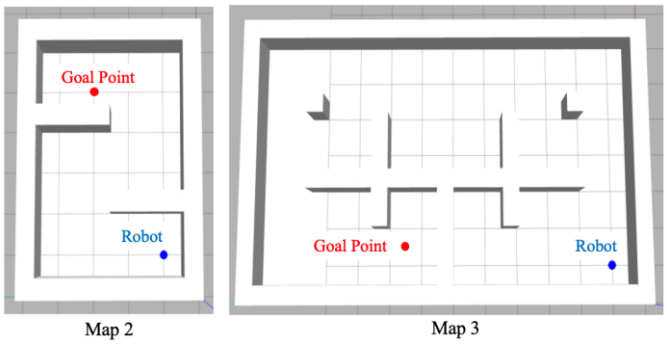


Fig. 6. Map 2 and 3 with their goal point.

TABLE II. THREE SLAM METHOD TRAILS FOR THE SECOND AND THIRD MAP

	Cartographer (s)		Gmapping (s)		Slam_toolbox (s)	
	Map_2	Map_3	Map_2	Map_3	Map_2	Map_3
Test 1	25.90	72.64	25.73	70.72	**24.29**	**70.11**
Test 2	26.10	70.15	26.10	72.17	**24.45**	**70.12**
Test 3	26.24	72.45	26.12	72.66	**24.71**	**70.07**
Average	26.08	71.75	25.983	71.85	**24.483**	**70.1**

Upon examination of Table II, it becomes apparent that when the robot is in operation with the SLAM_toolbox map, it achieves a quicker trajectory completion compared to the two other methods. This phenomenon is attributed to the SLAM_toolbox's capacity to produce maps characterized by a higher degree of precision and reduced noise levels. Consequently, the robot's operational stability is significantly improved.

A wide and intricate map with numerous obstacles was tested, as shown in Map_1. Fig. 7 illustrates the map when the robot navigates to its destination, with the green line indicating the path that the robot model must follow to reach its destination. The destination is set using RVIZ's 2D Goal Pose tool and is represented by a green arrow.

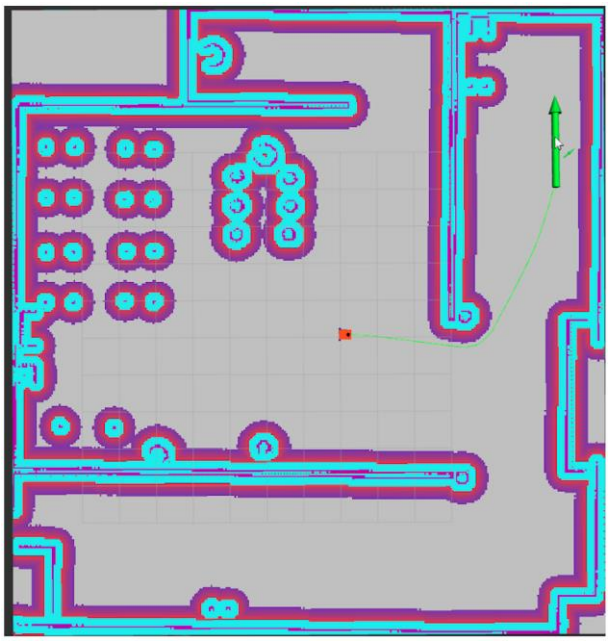


Fig. 7. Robot navigates to its destination in Map_1.

The last part was tested on Map_1 with three goal points. Each of these navigation tasks is executed using a map generated by its respective SLAM algorithm. The coordinates for point 1 is (x = 7.5, y = -2.0, z = 0.0); point 2: (x = 4.0, y = -8.0, z = 0.0); point 3: (x = -5.0, y = 0.0, z = 0.0). Fig. 8 shows Map_1 and the three goal points where the robot will move to, each represented by specific coordinates.

Fig. 8. The Map_1 with 3 different goal points.

Table III presents the duration it took for the robot to reach its destinations on the maps created by each method. It was observed that the map created using Slam Toolbox is more accurate than GMapping and Cartographer SLAM. To calculate the accuracy percentage of the point on the created map compared to the point we give before, we use the following Eq. (4):

$$Accuracy\ (\%) = [1 - (\frac{Error}{Range})] * 100 \quad (4)$$

where: Error is the Euclidean distance between actual goal points and the robot location after navigating, Range is the maximum possible distance between the two points, which is the distance between the actual point and the origin (0, 0) on Gazebo virtual world. Although the travel times to the destination points being relatively close for the use of maps in all three methods, Table IV shows the robot movement of the SLAM_toolbox method is more accurate to the virtual world than the other. The reason is Gmapping uses a particle filter to build grid maps from 2D lidar data and primarily relies on LIDAR data for mapping which is not well suited for expansive environments and struggles to accurately complete loop closures on an industrial scale, and the Cartographer relies on incorporating LIDAR and IMU data for highly accurate mapping because of its complexity in algorithm (IMU is not used in this robot model). Based on the empirical evidence from our conducted experiments, we readily infer that the utilization of the SLAM_toolbox for SLAM mapping is demonstrably superior.

B. Human Detection Results

The experiments were conducted to train and evaluate YOLO's versions using two distinct datasets. The training model was executed on an Ubuntu 20.04 platform, utilizing an Intel Core i7 processor and 16 GB of RAM. Firstly, we train YOLO's versions 20 epochs with the same hyparameters on a small dataset to choosing up the best optimal version for our work. After that we train chosen version 50 epochs on larger dataset with different optimizations such as SGD, AdamW [33], and AMSGrad [34]. Performance was assessed based on metrics such as Recision, Recall, and mAP(0.5-0.95).

1) Datasets and Evaluation Metrics:

a) Datasets: The research paper ultilizes three datasets created by using framework FiftyOne [35] to spliting human images, bounding box from the MSCOCO dataset [36], and JRDB dataset [39]. The small dataset contains 12.6K images with 10K for train and 2.6K for validation. The large dataset ontains 66.6K images with 64K for train and 2.6K for validation. Moreover, we also use augmentation methods on each of those dataset before training. Augment methods used are Blur, MedianBlur, ToGray, CLAHE.

TABLE III. THREE SLAM METHOD TIME RECORDED FOR THE FIRST MAP

	Cartographer (s)			**Gmapping (s)**			**Slam Toolbox (s)**		
	1st destination	2nd destination	3rd destination	1st destination	2nd destination	3rd destination	1st destination	2nd destination	3rd destination
Test 1	35.76	41.32	55.31	35.98	43.39	54.29	**35.05**	**41.35**	**53.58**
Test 2	35.90	42.35	54.91	35.74	43.51	54.77	**35.22**	**41.62**	**53.79**
Test 3	35.04	42.38	55.39	35.20	43.28	54.42	**35.01**	**41.84**	**53.45**
Average	35.56	42.01	55.20	35.64	43.39	54.49	**35.09**	**41.60**	**53.60**

TABLE IV. DISTANCE ACCURACY OF THREE SLAM METHODS FOR THE FIRST MAP

	Cartographer (m)			**Gmapping (m)**			**Slam Toolbox (m)**		
	Goal point 1	Goal point 2	Goal point 3	Goal point 1	Goal point 2	Goal point 3	Goal point 1	Goal point 2	Goal point 3
Test 1	x: 7.73867 y: -1.93425	x: 4.11695 y: -8.24249	x: -4.9892 y: -0.4127	x: 7.4873 y: -2.1837	x: 4.0047 y: -7.838	x: -4.988 y: 0.210	**x: 7.583 y: -2.088**	**x: 4.036 y: -8.046**	**x: -5.007 y: -0.027**
Test 2	x: 7.73895 y: -1.93484	x: 4.11672 y: -8.2428	x: -4.9896 y: -0.4125	x: 7.4933 y: -2.1807	x: 4.0123 y: -7.8801	x: -4.874 y: 0.1199	**x: 7.484 y: -2.018**	**x: 4.06252 y: -8.13627**	**x: -5.0832 y: -0.0174**
Test 3	x: 7.7385 y: -1.9343	x: 4.1163 y: -8.24216	x: -4.9890 y: -0.4122	x: 7.4854 y: -2.1858	x: 4.0098 y: -7.8823	x: -5.003 y: -0.173	**x: 7.5276 y: -2.092**	**x: 4.04852 y: -8.07527**	**x: -4.968 y: -0.027**
Average	x: 7.7387 y: -1.9344	x: 4.11668 y: -8.24248	x: -4.9893 y: -0.41254	x: 7.48874 y: -2.1834	x: 4.0089 y: -7.8668	x: -5.1035 y: -0.173	**x: 7.532 y: -2.066**	**x: 4.04952 y: -8.08527**	**x: -5.0198 y: -0.0242**
Accuracy	96.81%	97.0%	91.74%	97.63%	98.5%	96.04%	**99.05%**	**98.89%**	**99.37%**

b) Evaluation metrics: In object detection challenges and scientific research, various annotated datasets are used, and the primary metric for assessing the accuracy of object detections is the Average Precision (AP). AP relies on four fundamental concepts:

- True positive (TP): Accurately identifying a real bounding box in line with the ground truth.
- False positive (FP): Incorrectly identifying a non-existent object or inaccurately placing an existing object's detection.
- False negative (FN): Failing to detect a bounding box that matches the ground truth.
- True Negative (TN): Nevertheless, in the realm of object detection, the concept of True Negative (TN) is not relevant.

To address this, the intersection over union (IOU) metric is commonly employed. IOU is a measurement based on the Jaccard Index, which quantifies the similarity between two sets of data. In object detection, IOU calculates the overlap between the predicted bounding box (Bp) and the ground-truth bounding box (Bgt), dividing it by the area of their union. This provides a more suitable and informative metric for evaluating object detection performance, particularly in scenarios where True Negatives are not relevant [37].

In our research, we will primarily evaluate the performance of our model using the mean Average Precision (mAP) metric over a range of IoU thresholds (0.5 to 0.95). This mAP score considers the integral of mAP across these different IoU thresholds, providing a comprehensive assessment of detection accuracy. Additionally, we will also utilize the Precision metric to evaluate the percentage of correct positive predictions made by our model and the Recall metric to assess the percentage of correct positive predictions among all the ground truth annotations. These metrics collectively offer a well-rounded evaluation of our model's performance in object detection tasks.

The equation of those metrics is shown below:

$$\text{IoU} = \frac{\text{area of overlap}}{\text{area of union}} \tag{5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{FP+TP}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{FN+TP}} \tag{7}$$

2) Results and discussion: Fig. 9 to Fig. 11 illustrate the mAP (0.5-0.95), precision, and recall, while Table V presents the final results of training models on our dataset. In this set of results, two models, YOLOv8 and YOLOv6, outperform YOLOv5 and YOLOv7, with mAP@0.5:0.95 scores of 0.438 and 0.503, compared to 0.424 and 0.407, respectively. Notably, YOLOv6 demonstrates superiority across these metrics, as evidenced by its leading performance in Fig. 9 with the highest mAP score.

However, despite YOLOv6's impressive performance, we decided to exclude it from further consideration due to two critical factors. Firstly, it boasts a substantial number of parameters, totaling 4.3 million, which can strain computational resources. Secondly, its convergence appears to plateau within the first 20 epochs, suggesting that other models might potentially achieve even higher performance with prolonged training. Thus, we have opted for a trade-off between processing speed and accuracy in our model selection process.

Furthermore, YOLOv8, while not particularly outstanding in the initial 20 epochs, is our preferred choice. This decision is rooted in its consistent convergence, indicating that with more training time, it can potentially outperform other models. Additionally, as referenced in [13], YOLOv8 currently boasts superior processing speed compared to its counterparts, making it well-suited for real-time object detection applications.

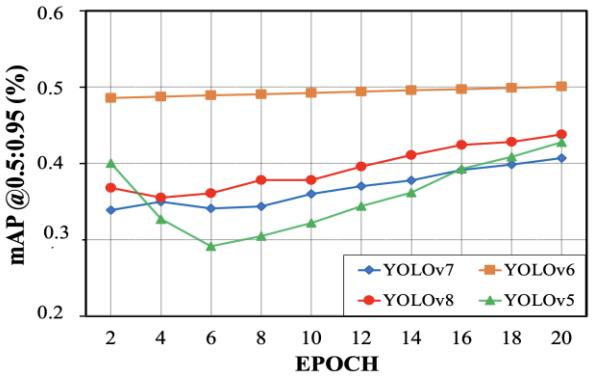


Fig. 9. mAP@0.5:0.95 of YOLO's versions.

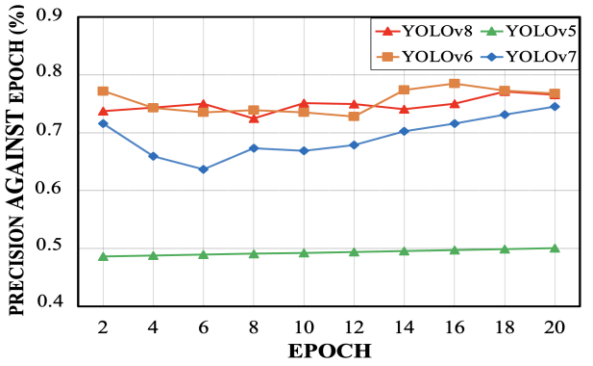


Fig. 10. Precision of YOLO's versions.

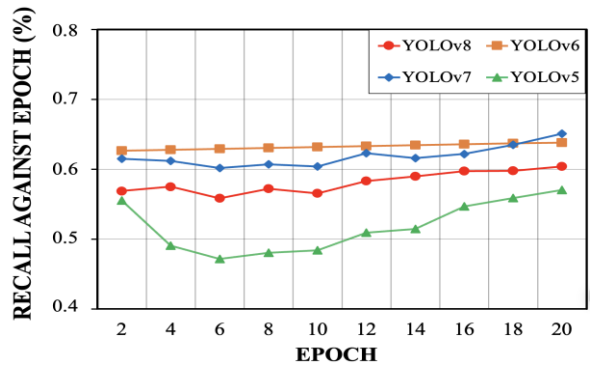


Fig. 11. Recall of YOLO's versions.

TABLE V. LAST RESULT OF FIRST EXPERIMENT

	Precision	Recall	AP (IoU=0.5)	AP (IoU=0.5:0.95)
YOLOv5	0.503	0.58	0.696	0.424
YOLOv6	0.768	0.651	0.791	0.503
YOLOv7	0.725	0.679	0.664	0.407
YOLOv8	0.766	0.604	0.728	0.438

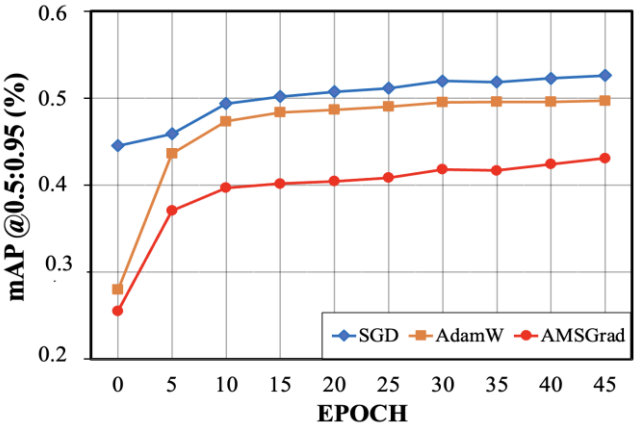


Fig. 12. mAP@0.5:0.95 of YOLOv8 with optimizers.

In the second experiment, we conducted training for 50 epochs using YOLOv8 on a dataset larger than the first with 66.6K images, employing various optimizers, including SGD, AdamW, and AMSGrad. The results from the last epoch are presented in Table VI. Notably, the SGD optimizer demonstrated superior performance across precision, recall, and mean Average Precision (mAP) with 0.794 Precision, 0.657 Recall, and 0.527_mAP@0.5:0.95. It has shown an increase over the original YOLOv8 model in [13] with only smaller than 0.4 mAP@0.5:0.95, which is obvious because here the parameters are trained to focus exclusively on human detection compared to numbers of class up to 80 in the original model. In Fig. 12, 13, and 14, we provide insights into the mAP@0.5:0.95, precision, and recall results throughout the 50-epoch training process. Remarkably, SGD consistently outperformed the other optimizers. Nonetheless, it's worth highlighting that even at the 50th epoch, all three methods exhibited ongoing improvement. This suggests that they have not yet reached their maximum potential, indicating that AdamW and AMSGrad might still have untapped strengths.

In Fig. 15, we conducted another evaluation to assess the human detection capability of the YOLOv8-SGD framework that we propose for the task of detecting individuals, particularly humans, within the JRDB dataset. This evaluation involved detecting individuals in various indoor settings, such as shopping centers, train stations, and cinemas.

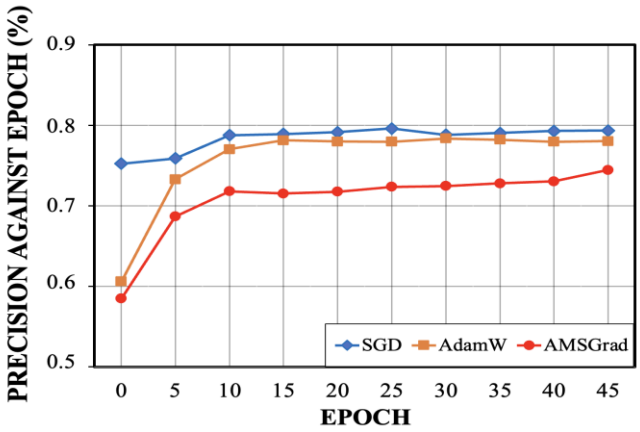


Fig. 13. Precision of YOLOv8 with optimizers.

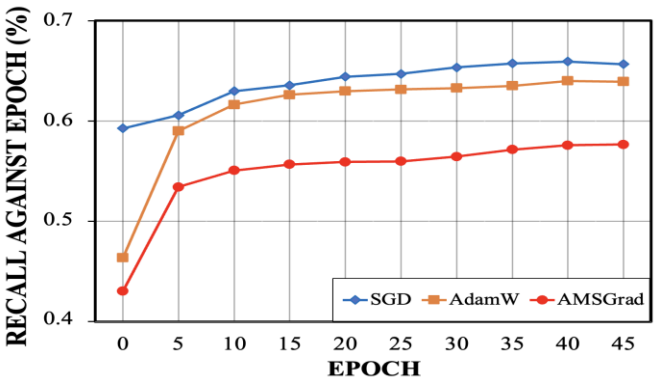
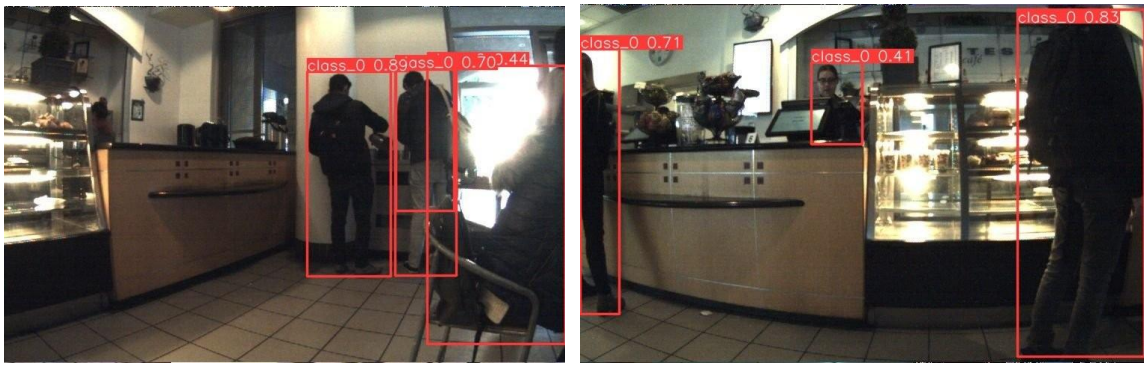


Fig. 14. Recall of YOLOv8 with optimizers.

TABLE VI. LAST RESULT OF THE SECOND EXPERIMENT

	Train /Box _loss	**Val /Box _loss**	**Val /dfl _loss**	**(Pre)**	**(Re)**	**AP IoU =0.5**	**AP IoU= 0.5:0.95**
YOLO v8-SGD	1.024	1.029	1.08	0.794	0.657	0.757	0.527
YOLO v8-ADAMW	1.082	1.087	1.15	0.782	0.639	0.735	0.498
YOLO v8-AMSGrad	1.205	1.2	1.28	0.754	0.577	0.674	0.438



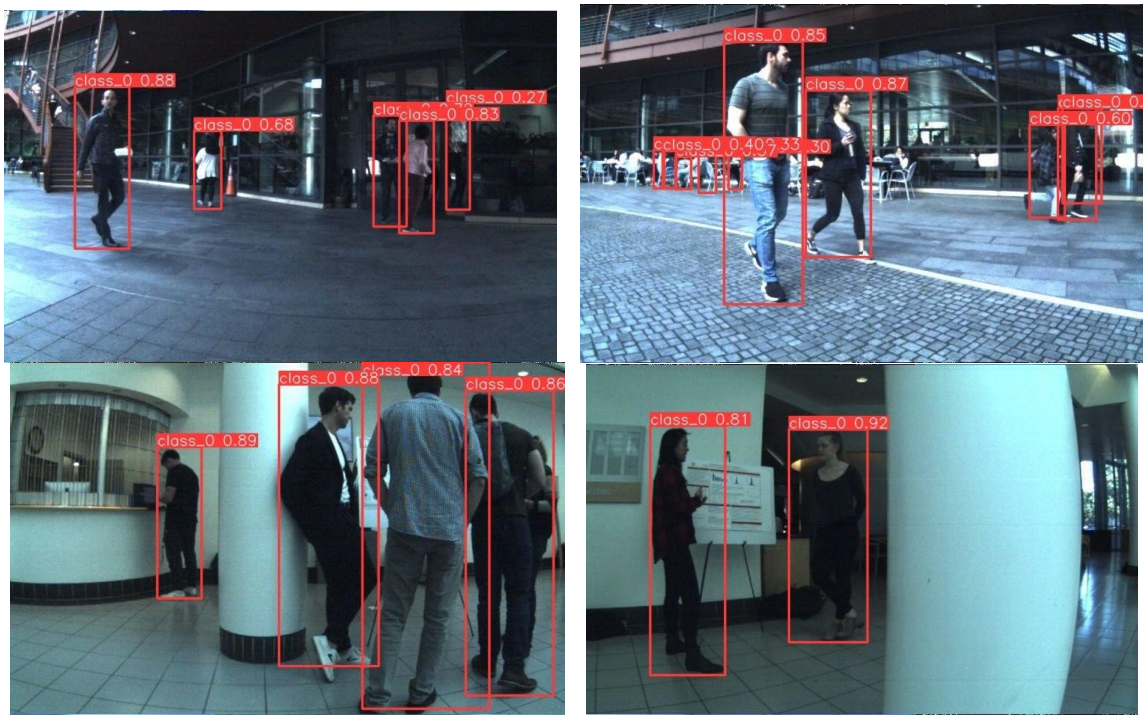


Fig. 15. The human detection capability of the YOLOv8-SGD framework with JRDB dataset.

V. Conclusion

We have successfully compared currently commonly used state-of-the-art methods for each task such as SLAM, Navigation, and Human detection. Furthermore, we have also provided analysis and evaluation of the experimental results.

Our findings align with our expectations, and we have identified potential areas for enhancement. We utilized the maps generated by each algorithm in both RVIZ 2 and Gazebo for guiding the robot to three distinct destinations. We repeated this process in five trials. The average values obtained from these tests were then used to create graphs that depict the time taken for various destinations, as shown in the table. The comparison revealed that the map generated with the Slam Toolbox exhibits greater precision compared to GMapping and Hector SLAM. The robot's motion aligns more accurately with the virtual world when using Slam Toolbox, in contrast to the others. However, the choice between these SLAM solutions depends on your specific project requirements, familiarity with the tools, hardware capabilities, and the complexity of your robot's operating environment. Each of these options has its strengths and weaknesses, and the better choice will vary depending on the context of your application.

As for human detection, the experimental results returned were quite surprising as the YOLOv8 model did not have outstanding results compared to older models; however, we see it still has potential for development, so we still choose it. Furthermore, because our initial goal was to apply to mobile robots with small microprocessors, this trade-off between processing speed and accuracy was extremely reasonable.

After this article, our next direction is to research more deeply into other more advanced tasks such as tracking, determine people's location, movements and behaviors of using 3D point cloud, etc. It will have some challenges such as privacy, security, fairness, scalability, and interdisciplinary collaboration will be addressed to ensure ethical and impactful innovation. But we hope that we can further apply modern technology to contribute to improving human happiness in today's society.

References

[1] M. Steve, M. Tom, L. David, M. Alexey, F. Michael. “From the Desks of ROS Maintainers: A Survey of Modern & Capable Mobile Robotics Algorithms in the Robot Operating System 2”, Robotics and Autonomous Systems, vol. 168, 2023. (https://doi.org/10.1016/j.robot.2023.104493)

[2] Huang, Jiahao, Steffen Junginger, Hui Liu, and Kerstin Thurow, "Indoor Positioning Systems of Mobile Robots: A Review" Robotics 12, no. 2: 47, 2023. https://doi.org/10.3390/robotics12020047

[3] Shibata T, Wada K. Robot therapy: a new approach for mental healthcare of the elderly - a mini-review. Gerontology. 2011;57(4):378-86. doi: 10.1159/000319015. Epub 2010 Jul 15. PMID: 20639620.

[4] Saputra, M.R., Markham, A., & Trigoni, A. (2018). Visual SLAM and Structure from Motion in Dynamic Environments: A Survey. ACM Comput. Surv., 51, 37:1-37:36.

[5] Zhou, H., & Zhang, T. (2014). The combination of SfM and monocular SLAM. The 26th Chinese Control and Decision Conference (2014 CCDC), 5282-5286.

[6] Pyo, Y., Jo, H., Jeong, R. and Im, T., (2017). ROS Robot Programming. Bucheon: Rubi Peipeo, pp.313:359.

[7] B. Abhishek, S. Gautham, D. Varun Rufus Raj Samuel, K. Keshav, U. P. Vignesh and S. R. Nair, "ROS based stereo vision system for autonomous vehicle," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, 2017, pp. 2269-2273.

[8] Oajsalee, S., Tantrairatn, S., & Khaengkarn, S. (2019). Study of ROS Based Localization and Mapping for Closed Area Survey. 2019 IEEE 5th International Conference on Mechatronics System and Robots (ICMSR), 24-28.

[9] Filipenko, M., & Afanasyev, I. (2018). Comparison of Various SLAM Systems for Mobile Robot in an Indoor Environment. 2018 International Conference on Intelligent Systems (IS), 400-407.

[10] Zhu, Xingkui, et al. "TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios." arXiv (Cornell University), Cornell University, Aug. 2021, https://doi.org/10.48550/arxiv.2108.11539.

[11] Li, Chuyi, et al. "YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications." arXiv (Cornell University), Cornell University, Sept. 2022, https://doi.org/10.48550/arxiv.2209.02976.

[12] Wang, Chien-Yao, et al. "YOLOv7: Trainable Bag-of-freebies Sets New State-of-the-art for Real-time Object Detectors." arXiv (Cornell University), Cornell University, July 2022, https://doi.org/10.48550/arxiv.2207.02696.

[13] Reis, Donald J., et al. "Real-Time Flying Object Detection With YOLOv8." arXiv (Cornell University), Cornell University, May 2023, https://doi.org/10.48550/arxiv.2305.09972.

[14] E. Marder-Eppstein, E. Berger, T. Foote, B. Gerkey and K. Konolige, "The Office Marathon: Robust navigation in an indoor office environment," 2010 IEEE International Conference on Robotics and Automation, Anchorage, AK, USA, 2010, pp. 300-307, doi: 10.1109/ROBOT.2010.5509725.

[15] Macenski S, Tsai D, Feinberg M. Spatio-temporal voxel layer: A view on robot perception for the dynamic world. International Journal of Advanced Robotic Systems. 2020;17(2). doi:10.1177/1729881420910530

[16] R P Guan, B Ristic and L Wang, "KLD sampling with Gmapping proposal for Monte Carlo localization of mobile robots", Information Fusion, no. 49, pp. 79-88, 2019.

[17] Dwijotomo, A.; Abdul Rahman, M.A.; Mohammed Ariff, M.H.; Zamzuri, H.; Wan Azree, W.M.H. Cartographer SLAM Method for Optimization with an Adaptive Multi-Distance Scan Scheduler. Appl. Sci. 2020, 10, 347. https://doi.org/10.3390/app10010347

[18] K. Konolige, G. Grisetti, R. Kümmerle, W. Burgard, B. Limketkai, andR. Vincent, "Efficient sparse pose adjustment for 2D mapping," IEEE/RSJ2010 Int. Conf. Intell. Robot. Syst. IROS 2010 - Conf. Proc., no. October,pp. 22–29, 2010.

[19] Macenski, Steve & Jambrecic, Ivona. (2021). SLAM Toolbox: SLAM for the dynamic world. Journal of Open Source Software. 6. 2783. 10.21105/joss.02783.

[20] Terven, Juan R., and Diana Cordova-Esparza. "A Comprehensive Review of YOLO: From YOLOv1 and Beyond." arXiv (Cornell University), Cornell University, Apr. 2023, https://doi.org/10.48550/arxiv.2304.00501.

[21] Murphy, K., Russell, S. (2001). Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks. In: Doucet, A., de Freitas, N., Gordon, N. (eds) Sequential Monte Carlo Methods in Practice. Statistics for Engineering and Information Science. Springer, New York, NY. https://doi.org/10.1007/978-1-4757-3437-9_24

[22] K. Murphy, "Bayesian map learning in dynamic environments," in Proc. Conf. Neural Inf. Process. Syst., Denver, CO, 1999, pp. 1015–1021.

[23] Trejos K, Rincón L, Bolaños M, Fallas J, Marín L. 2D SLAM Algorithms Characterization, Calibration, and Comparison Considering Pose Error, Map Accuracy as Well as CPU and Memory Usage. Sensors. 2022; 22(18):6903. https://doi.org/10.3390/s22186903

[24] W. Hess, D. Kohler, H. Rapp and D. Andor, "Real-time loop closure in 2D LIDAR SLAM," 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 2016, pp. 1271-1278, doi: 10.1109/ICRA.2016.7487258.

[25] Ma Yi-de, Liu Qing, and Qian Zhi-Bai. Automated image segmentation using improved pcnn model based on cross-entropy. In Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004., pages 743–746. IEEE, 2004.

[26] Zhang, Haoyang, et al. "VarifocalNet: An IoU-aware Dense Object Detector." arXiv (Cornell University), Cornell University, Aug. 2020, https://doi.org/10.48550/arxiv.2008.13367.

[27] Hossain, Sazzad, et al. "Dual Focal Loss to Address Class Imbalance in Semantic Segmentation." Neurocomputing, vol. 462, Elsevier BV, Oct. 2021, pp. 69–87. https://doi.org/10.1016/j.neucom.2021.07.055.

[28] Rezatofighi, Hamid, et al. "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression." arXiv (Cornell University), Cornell University, Feb. 2019, https://doi.org/10.48550/arxiv.1902.09630.

[29] Zheng, Zhaohui, et al. "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression." arXiv (Cornell University), Cornell University, Nov. 2019, https://doi.org/10.48550/arxiv.1911.08287.

[30] Bochkovskiy, Alexey, et al. "YOLOv4: Optimal Speed and Accuracy of Object Detection." arXiv (Cornell University), Cornell University, Apr. 2020, arxiv.org/pdf/2004.10934v1.

[31] Ding, Xiaohan, et al. "RepVGG: Making VGG-style ConvNets Great Again." arXiv (Cornell University), Cornell University, Jan. 2021, https://doi.org/10.48550/arxiv.2101.03697.

[32] Wang, Chien-Yao, Hong-Yuan Mark Liao, et al. "CSPNet: A New Backbone That Can Enhance Learning Capability of CNN." arXiv (Cornell University), Cornell University, Nov. 2019, https://doi.org/10.48550/arxiv.1911.11929.

[33] Loshchilov, Ilya, and Frank Hutter. "Decoupled Weight Decay Regularization." International Conference on Learning Representations, Sept. 2018, openreview.net/pdf?id=Bkg6RiCqY7.

[34] Reddi, Sashank J., et al. "On the Convergence of Adam and Beyond." arXiv (Cornell University), Cornell University, Feb. 2018, arxiv.org/pdf/1904.09237.

[35] Voxel. "GitHub - Voxel51/Fiftyone: The Open-source Tool for Building High-quality Datasets and Computer Vision Models." GitHub, github.com/voxel51/fiftyone.niversity), Cornell University, Apr. 2020, arxiv.org/pdf/2004.10934v1.

[36] Lin, Tsung-Yi, et al. "Microsoft COCO: Common Objects in Context." Lecture Notes in Computer Science, 2014, pp. 740–55. https://doi.org/10.1007/978-3-319-10602-1_48.

[37] R. Padilla, S. L. Netto and E. A. B. da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niteroi, Brazil, 2020, pp. 237-242, doi: 10.1109/IWSSIP48289.2020.9145130.

[38] K. Konolige, G. Grisetti, R. Kümmerle, W. Burgard, B. Limketkai and R. Vincent, "Efficient Sparse Pose Adjustment for 2D mapping," *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Taipei, Taiwan, 2010, pp. 22-29, doi: 10.1109/IROS.2010.5649043.

[39] Ehsanpour, M., et al., JRDB-Act: A Large-scale Dataset for Spatio-temporal Action, Social Group and Activity Detection. 2021, arXiv.

Optimizing Network Intrusion Detection with a Hybrid Adaptive Neuro Fuzzy Inference System and AVO-based Predictive Analysis

Dr. Sweety Bakyarani.E[1], Anil Pawar[2], Sridevi Gadde[3],
Mr. Eswar Patnala[4], Dr. P. Naresh[5], Prof. Ts. Dr. Yousef A. Baker El-Ebiary[6]
Department of Computer Science-Faculty of Science and Humanities, SRM Institute of Science and Technology, Kattankulathur, India 603203[1]
Professor, Department of Computer Engineering-Sanjivani College of Engineering, Savitribai Phule Pune University, Pune, India[2]
Assistant professor, Raghu Engineering College, Department of Computer Science and Engineering, AP, India[3]
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India[4]
Assistant Professor, Electrical and Electronics Engineering Department, Gandhi Institute of Technology and Management Visakhapatnam, Andhra Pradesh, India[5]
Faculty of Informatics and Computing, UniSZA University, Malaysia[6]

Abstract*—Protecting data and computer systems, as well as preserving the accessibility, integrity, and confidentiality of vital information in the face of constantly changing cyberthreats, requires the vital responsibility of detecting network intrusions. Existing intrusion detection models have limits in properly capturing and interpreting complex patterns in network behavior, which frequently leads to difficulties in robust feature selection and a lack of overall intrusion detection accuracy. The drawbacks of current methods are addressed by a unique approach to network intrusion detection presented in this paper. This framework discusses the difficulties presented by changing cyberthreats and the critical requirement for efficient intrusion detection in a society growing more networked by the day. Using a Hybrid Adaptive Neuro Fuzzy Inference System and African Vulture Optimization model with Min-Max normalization and data cleaning on the NSL-KDD dataset, the methodology outlined here overcomes issues with complex network behavior patterns and improves feature selection for precise identification of potential security threats. This approach meets the need for an effective intrusion detection system. Python software is used to implement the suggested model since it is flexible and reliable. The results show a notable improvement in accuracy, with the Hybrid Adaptive Neuro Fuzzy Inference System and African Vulture Optimization model surpassing previous approaches significantly and obtaining an exceptional accuracy rate of 99.3%. The accuracy of the proposed model was improved by African Vulture Optimization, rising from 99.2% to 99.3%. When compared to Artificial Neural Network (78.51%), Random Forest (92.21%), and Linear Support Vector Machine (97.4%), this amazing improvement is clear. When compared to other techniques, the suggested model exhibits an average accuracy gain of about 20.79%.

Keywords*—Network intrusion; cyberthreats; normalization; African vulture optimization; data cleaning

I. Introduction

The Internet has smoothly merged into the framework of everyday life in the rapidly changing digital age, acting as a vital resource for both people and businesses. It now serves as the foundation for keeping records, company operations, and connectivity. The safety and confidentiality of online transactions, nevertheless, are an increasing worry brought on by this previously unheard-of dependence on the World Wide Web. Because of this, cybersecurity has become a crucial area of concern for both business and academics, motivating the commitment of significant funds to protect contemporary web-based networks from possible dangers and abnormalities [1]. Several cyber security issues have emerged, posing several potential hazards to the online lives [2]. Attackers frequently use the flaws in well-known software to target computer systems on networks. These attackers' damage may result in significant issues like service interruptions or even substantial financial losses [3]. In modern linked world, NIDSs are imperative for protecting the accessibility, security, and reliability of data [4]. The two primary types of methods used by NIDSs to do this are signature-based detection and anomaly-based detection [5]. In order to be very effective in recognizing assaults using widely recognized signatures and structures, signature-based NIDSs rely on predefined attack patterns [6]. Nevertheless, their susceptibility to novel attacks is a serious obstacle because they are unable to adjust to new dangers without foreknowledge. Fig. 1 shows the Network intrusion detection system.

There is still much space for enhancement of NIDS efficiency, despite major improvements. The enormous amount of information about network traffic produced, the rapidly changing technical environment, the sizeable collection of features that make up data sets for training and the requirement for actual intrusion detection provide difficulties [7]. The efficacy of NIDS and the speed of training models can both be hampered by unnecessary or redundant characteristics [8]. Because of this, improving the effectiveness of machine learning -based detection models require careful features subset selection and tuning. By defining and explaining what makes up usual network behavior, anomaly-based detection systems, on the other hand, can be used to discover assaults that are unknown or novel.

Both methods have advantages, but to handle the always changing network security concerns, constant advancements are needed [9]. Thus, in order to improve the accuracy of machine learning (ML)-based detection models, careful data subset selection and parameter optimization are required [10].

Different methods were used to fortify network defenses, each with their own advantages and disadvantages. By detecting known attacks using predetermined sequences of attacks, signature-based detection systems act as the initial line of protection [11]. It is impossible to overstate how successful they have historically been at identifying popular hazards. Nevertheless, because they rely on historical signatures for identification, these tools fail when faced with fresh, never-before-seen threats [12]. Anomaly-based Identification systems are skilled at identifying unidentified dangers based on models created around typical actions since they focus on the recognition of abnormalities from standard procedures [13]. As they may be able to identify new threats that signature-based systems overlook, these systems add a key layer of security [14]. Systems for detecting intrusions now depend heavily on machine learning techniques [15]. The focus of conventional methods is on implementing feature engineering and selection, which can be computationally demanding and could only collect deep characteristics, leading to subpar recognition rates. With the promise for higher precision and fewer positive results, artificial intelligence techniques have become widely used to detect fraudulent network traffic [16]. RNNs, CNNs and DRL are a few examples of deep learning techniques that have shown promise in overcoming the drawbacks of conventional machine learning.

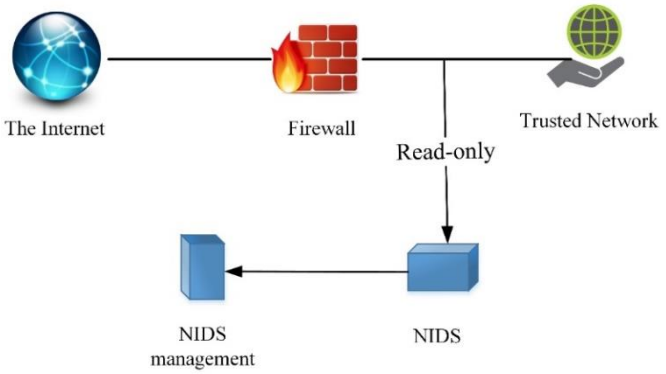


Fig. 1. NID system.

Ongoing improvements in NIDS are required due to the explosive growth of the Internet and the resulting rise in security threats [17]. To develop reliable and effective techniques and systems that precisely detect and react to unwanted or harmful actions within a computer network is the overarching problem statement for network intrusion detection. To safeguard the network's resources from many types of assaults, like as intrusions, infections with viruses, compromises of data, and denial-of-service attacks, this entails differentiating between normal network traffic and suspicious or malicious behavior. Key contributions of this work as follows:

- Data cleaning and Min-Max normalization are included in the study's thorough pre-processing procedure. These methods guarantee the dataset's dependability and quality, which improves the accuracy of subsequent intrusion detection analysis.
- African Vulture Optimization (AVO) provides a novel feature selection method based on the scavenging habits of vultures. By shrewdly determining the most pertinent features for network intrusion detection, it greatly improves model performance and interpretability.
- The Hybrid ANFIS (Adaptive Neuro-Fuzzy Inference System) Model is the central component of the research. The effective capturing of complex patterns in network behavior is made possible by the Hybrid ANFIS. In order to achieve accurate and exact intrusion detection, the model adapts and fine-tunes fuzzy rules through backpropagation.

An outline of the research is given in Section I. The Section II analyses the material that has already been written and highlights the need to handle particular modifications in Network Intrusion detection. Concerning the complexity of Network Intrusion detection, Section III defines the main research issue. Data collection, preprocessing, feature selection, and Hybrid ANFIS are described in Section IV of the paper. The research's importance in detecting Network Intrusion Detection is emphasized in Section V, which gives empirical findings, compares classifier performance, examines consequences and Section VI summaries the conclusion of the paper.

II. Related Works

An essential component of ensuring cybersecurity is intrusion detection and the capability to identify attacks. The aim of this study is on defending against assaults on cyberattack detectors built using machine learning. Here, an approach for adversarial machine learning detection was developed by Pawlicki, Choras, and Kozik [18]. In actuality, the confrontational nature of the surroundings in which they are used has not been taken into consideration while designing modern machine learning algorithms. Therefore, a variety of attacks are currently being directed towards machine learning systems. By constructing adversarial attacks using the four currently proposed approaches, this work assesses the potential of degrading the effectiveness of an intrusion recognition process at test time and then provides a mechanism to identify such assaults. Both artificial neural networks and four techniques for creating adversarial attacks have the necessary historical data. The five separate classifiers' outputs are contrasted, and a thorough explanation of the new detecting technique is provided. In this work intrusion detection systems have not yet been extensively studied in terms of identifying confrontational attacks on false neural networks. And also, it has high false positive rate.

Detecting network intrusions is crucial for maintaining digital security on the network. The primary analysis technique employed in the area of NID is the identification and evaluation of aberrant traffic by extracting statistical

aspects of flow. However, because these features must be created and retrieved individually, the original flow data is frequently lost, which reduces the effectiveness of the detection process. In this study, instead of explicitly designing the process's features directly collected the flow's actual data details for evaluation. The deep hierarchical network, which incorporates the enhanced LSTM and LeNet-5 neural network is developed by Zhang et al. [19]. Instead of retraining a pair of networks independently, a feasible network cascade approach was designed to train the suggested hierarchical network simultaneously. A unique traffic collection system can need a lot of resources, including specialized hardware, software, and committed employees, to build and maintain. Managing large amounts of real-world traffic data might put a burden on the facilities available due to computational and storage needs. This strategy could cost a lot to implement in terms of infrastructure setup and ongoing maintenance.

The advancement of network infrastructure and technology has advanced quickly in the past few decades, and Internet services have extended throughout all industries. The prevalence of infringement has increased, and many contemporary systems have been breached, making the advancement of technology for information security to identify new attacks essential. An IDS that uses deep and machine learning algorithms to identify irregularities in network traffic is the greatest vital security-related technologies. Employing a deep neural network techniques and an outstanding network efficiency for network intrusion detection was proposed by Maithem and Al-sultany [20]. The primary purpose of this study is to use advanced IDS to find unidentified attack packages. In this model, detection of attacks is carried out in two different ways (binary categorization and several classes' classification). With regards to the high accurateness with multiclass categorization and with dual classification), the suggested system has demonstrated interesting results. It suggests that the research used deep learning approaches to detect network assaults with excellent classification accuracy. A number of significant flaws and restrictions demand attention. In order to make sure that these networks fail to biased in a specific dataset like the KDD Cup 99. For a viable implementation, it is also essential to handle data imbalance and address computational scalability difficulties.

The potential attack area for cyber hackers is expanding as additional gadgets with internet access come online. Many intrusion detection systems look for identified breaches using network communications characteristics. Despite depending on these fingerprints, investigators have recently employed a machine learning techniques to identify network threats. For an intrusion detection system that is ready for the market, these methods often have a high false-positive rate. Atefinia and Ahmadi [21] proposed a deep neural network model in this study to decrease the incidence of intrusion detection systems that overreact in response to anomalies. The trials make use of the CSE-CIC-IDS2018 dataset, and the models that emerge from them can be included into IDS to produce alerts or thwart upcoming assaults. Though seeking to reduce alarms that are unfounded, the creation of a modular deep neural network for intrusion detection has a number of possible downsides and difficulties. Complexity and interpretability problems can be introduced by modular neural networks. It might be demanding of resources and difficult to apply custom feature extractors and datasets to various contexts. Increasing the amount of training with large data systems carries the danger of over fitting, and there may be issues with data privacy and security.

The reviewed research papers bring up some of the shortcomings and difficulties that intrusion detection systems face, especially when dealing with large false-positive rates and aggressive attacks on neural networks. Handling large amounts of real-world traffic data can put a burden on resources and result in high maintenance and infrastructure expenditures. The reliability of these systems depends on the careful handling of notable problems such bias, data imbalance, and computing scalability. When working with big datasets, modular neural networks can add complexity and interpretability issues, requiring resources and putting data privacy at risk. The aforementioned studies highlight the necessity of continued investigation to surmount these constraints and augment the efficacy of intrusion detection systems with regard to the dynamic landscape of cyber threats.

III. Problem Statement

From the above literature reviews, the necessity to improve intrusion detection system's effectiveness as well as efficacy in order to safeguard computer networks and data against cyberattacks is the central issue in the arena of intrusion detection. Identifying earlier unidentified and changing attack strategies, maintaining sizable and imbalanced data sets, assuring real-time detection and response, and enhancing the comprehensibility of detection models are some of the current issues [21].

IV. Network Intrusion Detection using Hybrid ANFIS and AVO-based Predictive Analysis

The main measures made to improve the precision and effectiveness of the intrusion detection system was described in the methodology section of the network intrusion detection study. The NSL-KDD dataset, a labeled network intrusion detection dataset with a wide variety of network traffic scenarios, was the primary dataset used in this study. It commenced a thorough pre-processing step to guarantee the dataset was prepared for analysis. In order to standardize the data and prepare it for further analysis, this required data cleaning and Min-Max normalization. The process of feature selection, a crucial factor in determining the effectiveness of intrusion detection systems, forms the basis of the methodology. It used the ground-breaking African Vulture Optimization (AVO) method to address this. AVO provides a clever way to choose the most pertinent characteristics, optimizing the feature subset for better model performance and interpretability. It is inspired by the scavenging activity of vultures. After selecting features, the Hybrid Adaptive Neuro-Fuzzy Inference System (ANFIS) was put into use to detect intrusions. Fuzzy logic and neural networks are used with ANFIS to capture complex patterns in network behavior. It achieves accurate intrusion detection by fine-tuning fuzzy rules through backpropagation. By lowering false positives and increasing detection precision, this methodology seeks to

strengthen network security in the end. Compared to alternative methods, this thorough framework exhibits an exceptional accuracy rate of 99.3%, proving its efficacy through rigorous experimentation. Fig. 2 shows the Overall Framework for Hybrid ANFIS and AVO-based Network Intrusion Detection.

A. Data Collection

The labeled network intrusion detection dataset known as NSL-KDD was utilized in numerous assignments to test various deep learning-based methods for developing various IDS techniques. Four types of characteristics—basic, time-based, content-based and host-based traffic features—can be distinguished among the 41 characteristics included in the NSL-KDD dataset. These traits' worth is mainly determined by their constant, separate, and symbolic nature. Five bout classes, including Normal Denial-of-Service (DoS), Root to Local (R2L), Probe, and Unauthorized to Root (U2R), are included in the NSL-KDD dataset. The attributes associated with each NSL-KDD data can be used to determine these attack classifications [22].

B. Pre-processing with Data Cleaning and Min-Max Normalization

Data preparation modifies the data ranges in an NSL-KDD dataset to improve information gathering and operation. The dataset's maximum and minimum ranges exhibit high contrast variance. The normalization of data during this phase lessens an algorithm's challenges.

Data cleaning procedures are used to eliminate data redundancy, noise, mistakes, and undesirable information from the dataset. Only the pertinent data may be processed subsequently in this process.

The normalization role, which contains a least and extreme algorithm and transforms the remaining data value between [-1, 1] and [0, 1], heavily relies on data scalability. The normalization formula is given in Eq. (1).

$$I'^{*} = \frac{D*'-D*'_{min}}{D*'_{max}-D*'_{min}} \quad (1)$$

The phrase I' in Eq. (1) denotes the value of the input that has been transformed (or, more specifically, normalized). The terms D'_{max} *and* D'_{min} stand for the supreme and lowest values of the input variable D', respectively, whereas the term D' stands for actual value [23].

C. Feature Selection using African Vulture Optimization

African Vulture Optimization, an innovative metaheuristic strategy based on the behavior of African Vultures, has become a popular choice for feature selection in network intrusion detection. The world's most common bird species, vultures, are usually carnivorous. They are dependent on scavengers to dissect carcasses because they are incapable of doing so themselves. African vultures have been reported to soar above 11,000 meters and to circle over great distances in search of possible food sources. However, once they locate a food supply, they have trouble getting to it quickly. The stronger vultures are frequently encircled by the less powerful ones, which delays the feeding process. The less dominant vultures gradually start looking for food when the dominant ones get tired. This distinctive scavenging behavior has prompted the creation of a novel metaheuristic method designed to tackle feature selection problems in the field of network intrusion detection [24].

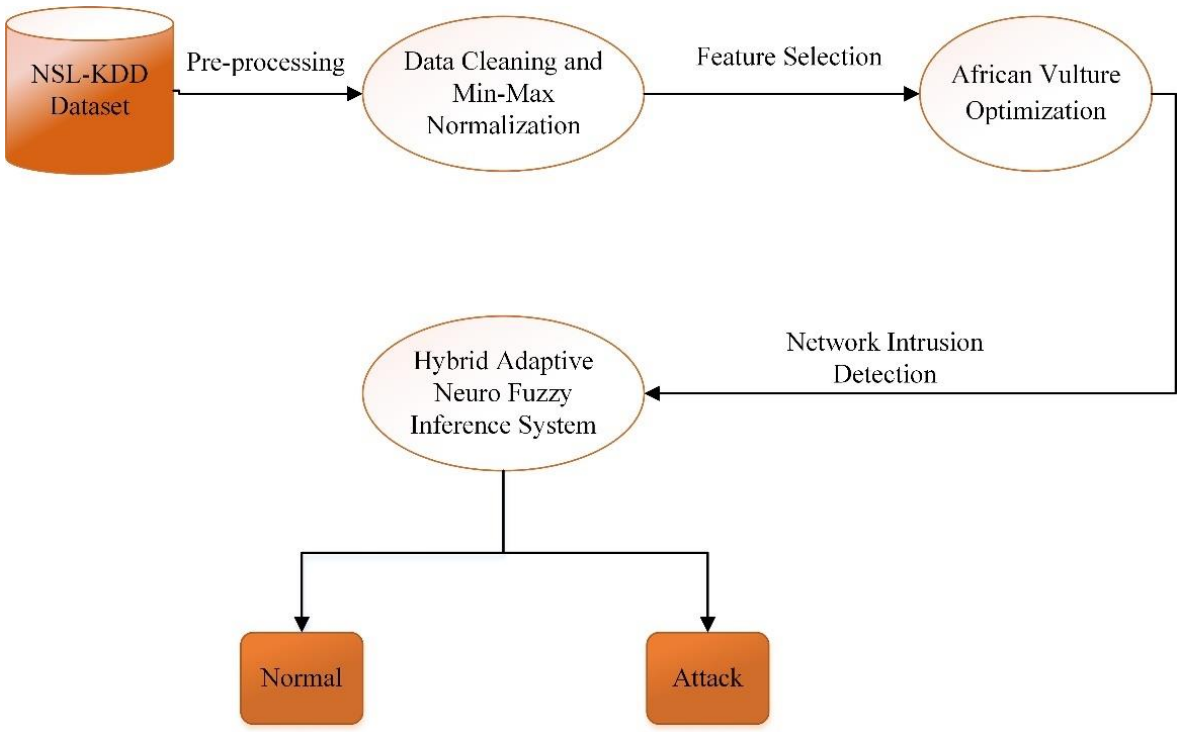


Fig. 2. Overall framework for hybrid ANFIS and AVO-based network intrusion detection.

Step 1: To ensure that every feature is given equal weight in the first stage of feature selection for network intrusion detection, try to increase the number of possible characteristics. Finding the highest-performing characteristics within a particular group is the main goal. Eq. (2) gives the description of this operation.

$$S(I) = \begin{cases} best\ vul_1, if\ Z_l = F_1 \\ best\ vul_2, if\ Z_l = F_2 \end{cases} \quad (2)$$

Here, F_1 and F_2 are parameters that were assessed before to optimization; they both have to be between 0 and 1, with the requirement that $F_1 + F_2 = 1$.

Step 2: Finding the features' "famine rate" is the focus of this step. Characteristics are rated according to how well they can advance the optimization, just like actual vultures fly in pursuit of food. Eq. (3) accurately depict this process mathematically:

$$V = (2 \times \tau + 1) \times n \times \left(1 - \frac{iter_l}{max_{iter}}\right) + I \quad (3)$$

In Eq. (3), τ denotes a random number between 0 and 1. The current iteration is indicated by $iter_l$, the total number of iterations is shown by max_{iter}r, and n is a constant that directs the optimization process through its stages of investigation and processing. There are restrictions on the variable d that fall between -2 and 2. If n is less than zero, it means that a feature is scarce, like vulture hunger, and if n is greater than one, it means that a trait is abundant, like a well-fed vulture.

Step 3: In order to facilitate feature selection, vultures are outfitted with random feature subsets that provide two possible configurations and a variable r_1 that has a range of zero to one. Eq. (4) and Eq. (5) provide the following mathematical breakdown of the technique used to choose the best feature set:

$$If\ r_1 \geq randr_1$$
$$S(l+1) = BV(l) - T(l) \times S \quad (4)$$
$$If\ r_1 < randr_1$$
$$S(l+1) = BV(l) - S + rand_2 \times ((ub - Ib) \times rand_3 + Ib \quad (5)$$

S here stands for the unique feature subsets that were selected at random when searching for the best features. The best-performing feature subsets are stored in BV, while the lower and upper bounds for feature values are represented by Ib and ub. Two random variables, rand2 and rand3, have values ranging from 0 to 1.

Step 4: The selected features are represented by the value of |S|, which splits this step into two halves. Both segments include rotational flights when 0.5 < |S| < 1. An active feature can be identified if |S| is greater than or equal to 0.5. During this stage, less significant features try to use the stronger ones to improve their performance. The position of the new feature set, designated as S(k+1), is found by applying Eq. (6), Eq. (7), and Eq. (8).

$$S(l+1) = \frac{best_1 + best_2}{2} \quad (6)$$

$$b_1 = best\ vul_1(l) - \frac{best\ vul_1(l) \times S(l)}{best\ vul_1(l) - S(l)^2} \times S \quad (7)$$

$$b_2 = best\ vul_2(l) - \frac{best\ vul_2(l) \times S(l)}{best\ vul_2(l) - S(l)^2} \times S \quad (8)$$

In the field of network intrusion detection, African Vulture Optimization (AVO) feature selection is a new method. Using a dynamic optimization process that imitates vulture foraging activity, AVO enhances intrusion detection systems' effectiveness by selecting pertinent features. The technique improves detection accuracy and reduces false positives by utilizing AVO, which guarantees that the most important characteristics are taken into account. This adds to the general stability and dependability of the network security configuration.

D. Employing Hybrid Adaptive Neuro-Fuzzy Inference System for Intrusion Detection

A hybrid ANFIS system for intrusion detection combines the features of fuzzy logic and neural networks to produce a flexible and adaptive solution for identifying network intrusions while reducing false positives. ANFIS contains of five layers, as shown in Fig. 3.

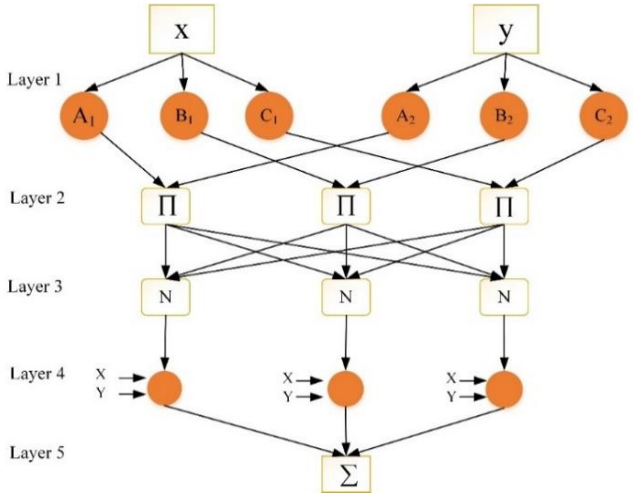


Fig. 3. ANFIS architecture.

The fuzzification layer is represented by Layer 1. It computes the membership function as follows in Eq. (9) and Eq. (10):

$$\lambda_{ai}(y) = \frac{1}{1 + \left[\left(\frac{y - c_{ai}}{\alpha_{ai}}\right)^2\right]^{\beta_{ai}}} \quad (9)$$

$$\lambda_{bi}(z) = \frac{1}{1 + \left[\left(\frac{z - c_{bi}}{\alpha_{bi}}\right)^2\right]^{\beta_{bi}}} \quad (10)$$

Where, the bell function parameters are c_{ai}, c_{bi}, α_{ai}, α_{bi}, β_{ai}, and β_{ai}.

Layer 2 establishes the rules layer. Each node's firing power is represented by the output in Eq. (11):

$$\omega_i = [\lambda_{ai}(y)] \times [\lambda_{bi}(z)] \quad (11)$$

The normalization layer is designated as Layer 3. It adjusts the computed firing strength to be normal by Eq. (12).

$$\overline{\omega_i} = \frac{\omega_i}{\omega_1+\omega_2} \quad (12)$$

Hybrid ANFIS and AVO Algorithm	
Load the input data	
Perform preprocessing operation	//Data Cleaning and Min-Max Normalization
Select feature using AVO	
Calculate the fitness of vultures	
If s ≥ 1then	
Upgrade vulture location using Eq. (5)	
Otherwise	
Upgrade vulture location using Eq. (6)	
Network Intrusion Detection	//Hybrid ANFIS
Calculate RMSC of each particle	
If goal achieved	
Apply Optimal Particle Position	
End	

The consequent layer is represented by Layer 4. This layer's output is the result of multiplying the polynomial resulting from fuzzy rules by the normalized firing strength of Eq. (13):

$$\overline{\omega_i}F_i = \overline{\omega_i}(p_i y + q_i z + r_i) \quad (13)$$

where, p_i, q_i and r_i are the consequent parameter sets.

The defuzzification layer was designated as Layer 5. The ANFIS output as a whole is what it produces by Eq. (14).

$$F = \sum_i \overline{\omega_i}F_i = \frac{\sum_i \omega_i F_i}{\sum_i \omega_i} \quad (14)$$

The learning algorithm must adjust each adjustable parameter in order for the initial training data to match the output of the ANFIS. The RMSE (Root Mean Square Error) between projected values and actual measurements is trained into the ANFIS model to get the lowest possible value. RMSE is characterized by Eq. (15):

$$RMSE = \sqrt{\frac{1}{N*}\sum_{i=1}^{N*}(y-z)^2} \quad (15)$$

In Eq. (15), N* is the number of samples, y represents the actual measurement, z the forecasted value.

When optimizing Network Intrusion Detection with a Hybrid Adaptive Neuro-Fuzzy Inference System, the parameters of the African Vulture Optimization algorithm play a crucial role. In feature selection, AVO is essential to the intrusion detection system's overall efficacy. The optimization procedure is directly impacted by the particular parameters of the AVO algorithm, such as population size, convergence criterion, and number of iterations. The Hybrid ANFIS and AVO-based system's accuracy and efficiency are greatly impacted by fine-tuning these parameters, which improves predictive analysis. Through their interaction, AVO and the Hybrid ANFIS model provide a synergistic approach that successfully addresses the difficulties associated with network intrusion detection and highlights the significance of parameter tuning for obtaining better outcomes in cybersecurity applications. Overall process of the proposed model is given in Fig. 4.

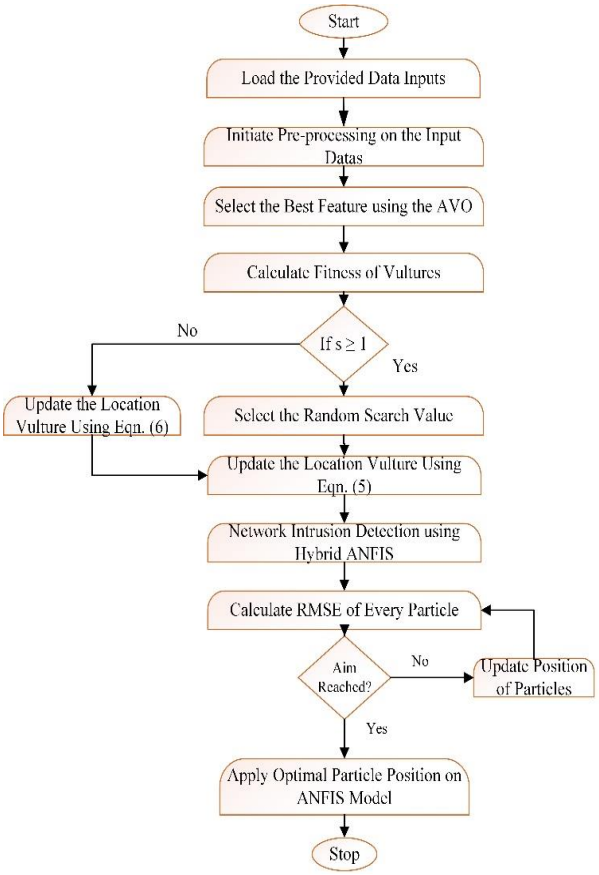


Fig. 4. Overall flowchart of proposed hybrid ANFIS and AVO.

V. Results and Discussions

The results of this extensive technique should be presented in the network intrusion detection study's results section. First, describe in detail how we trained and tested our intrusion detection system using the NSL-KDD dataset. Examine the pre-processing findings after the first round of data collection, emphasizing how much data cleaning and Min-Max normalization improved the dataset's analytical fit. Then the key feature selection procedure, where the African Vulture Optimization (AVO) approach was used. This data shows how well AVO performs in terms of identifying the most pertinent characteristics, refining the feature subset, and enhancing interpretability and performance of the model. The Hybrid Adaptive Neuro-Fuzzy Inference System (ANFIS) is an effective tool for intrusion detection. Its ability to capture complex patterns in network behavior is one of its most important applications. ANFIS optimizes fuzzy rules to provide extremely accurate intrusion detection through its backpropagation technique. The results show that this approach significantly improves detection precision while also lowering false positives, which strengthens network security in the end. Robust testing has generated empirical confirmation, demonstrating the framework's improved performance at a remarkable 99.3% accuracy rate—quite an accomplishment in comparison to other well-established methodologies. This method's effectiveness is confirmed by these results, which also highlight how much better network intrusion detection systems could become.

A. Training and Validation Accuracy of Hybrid ANFIS and AVO

Using 80% of the data for training and 20% for validation was the proposed strategy. The accuracy level and loss rate fluctuation graphs for the complete Hybrid ANFIS and AVO model procedure are shown in Fig. 5 and Fig. 6. When the training intervals of the Hybrid ANFIS and AVO model reach 100, the overall graph of the accurateness ratio and loss ratio stabilizes.

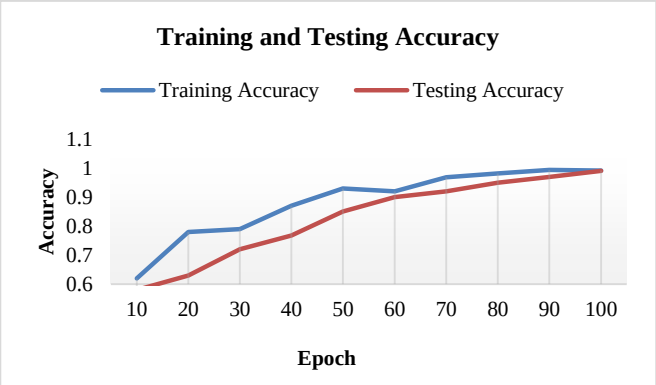


Fig. 5. Accuracy of training and testing values.

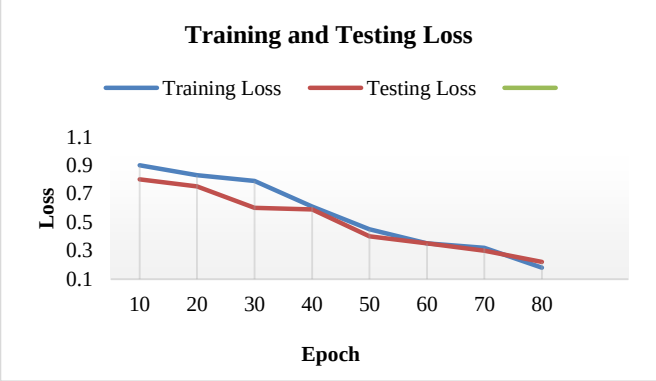


Fig. 6. Loss of training and testing values.

B. Evaluation of Performance

Recall, F1-score, precision, and accuracy were employed as comparison evaluation criteria. The model was evaluated using these parameters. They are shown below:

Accuracy: The prediction accuracy used to evaluate classification performances, as given in Eq. (16), is used to evaluate the classifier's overall performance.

$$A = \frac{TP\prime+TN\prime}{TP\prime+TN\prime+FP\prime+FN\prime} \quad (16)$$

Precision: The degree to which a collection of outcomes approve with one another is referred to as precision. The example is in Eq. (17).

$$P = \frac{TP\prime}{TP\prime+FP\prime} \quad (17)$$

Recall: To determine, below a convinced set of molds, a particular dependent variable, recall analysis, as illustrated in Eq. (18). This process is carried out within predetermined bounds that depend on one or more factors in the incoming data.

$$R = \frac{TP\prime}{TP\prime+FP\prime} \quad (18)$$

True positive pixels are denoted by TP', true negative pixels by TP', false positive pixels by FP', and false negative pixels by FN'.

F1-score: Recall and accuracy are related in the classification task. The F1-score definition is shown in Eq. (19).

$$F1-score = 2 * \frac{Pre*Re}{Pre+\mathrm{Re}} \quad (19)$$

The assessment results of the created Network intrusion detection system employing the combined strategy are shown in Table I.

TABLE I. PERFORMANCE METRICS OF HYBRID ANFIS AND AVO MODEL

Metrics	Values (%)
Accuracy	99.3
Precision	96.8
Recall	98.5
F1-Score	99.1

The performance indicators demonstrate the efficiency of the proposed model in detecting network intrusions, which is quite encouraging. With an astounding accuracy rate of 99.3%, the classification of network activity is classified with a high degree of overall accuracy. The system's precision in reducing false positives is demonstrated by the impressive score of 96.8% obtained by precision, a metric that gauges the model's ability to properly categorize incursions. Recall, which measures how well the model can identify real incursions, is also quite significant at 98.5%. A robust 99.1% is reached by the F1-Score, which balances recall and precision and highlights the system's well-balanced performance.

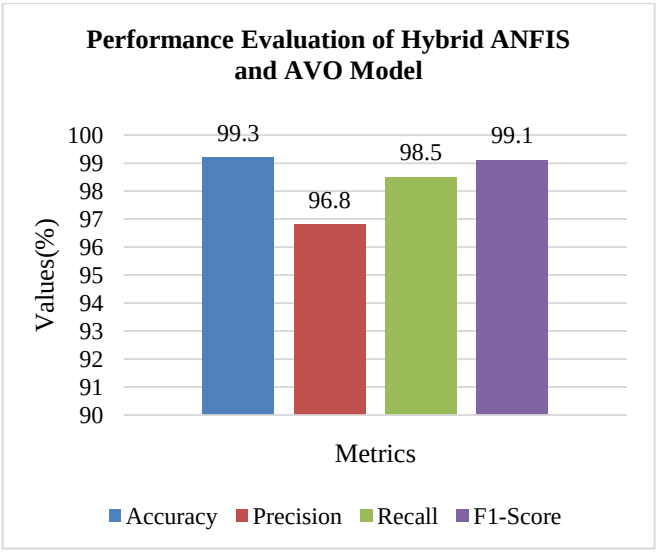


Fig. 7. Performance evaluation of hybrid ANFIS and AVO model.

These results demonstrate accurate and trustworthy anomaly detection with high values across key performance measures. The performance evaluation of the proposed Hybrid ANFIS and AVO model is shown in Fig. 7.

Table II shows the Accurateness, Recall, Precision and F1-score of the proposed approach with existing methods. The accuracy of the suggested method Hybrid ANFIS and AVO model (99.3%) is higher than the existing approaches ANN (78.51%), Random Forest (92.21%) and Linear SVM (97.4%).

Fig. 8 depicts the graphic depiction of the performance metrics of proposed with existing approaches. The precision of the suggested method Hybrid ANFIS and AVO model (96.8%) is higher than the existing ANN (96.6%), Random Forest (96.8%) and Linear SVM (97%). The recall of the suggested method Hybrid ANFIS and AVO model (98.5%) is higher than the existing approaches ANN (62.05%), Random Forest (61.5%) and Linear SVM (97%). The F1-score of the suggested method Hybrid ANFIS and AVO model (99.1%) is higher than the existing approaches ANN (75.5%), Random Forest (75.2%) and Linear SVM (97%).

The suggested model's graph shows how, as it reached a point of consistency using the provided hyperparameters, the accuracy of the training and validation sets rose quickly over a shorter period of time.

Before optimization the value of accuracy for proposed Hybrid ANFIS and AVO model is 99.2%. The accuracy achieved after optimization using Hybrid ANFIS and AVO model is 99.3%. The fitness of Hybrid ANFIS and AVO model is depicted in Fig. 9.

C. Discussion

This paper highlights the relevance of the results through a thorough analysis, including a comparison with existing methods, by combining African Vulture Optimization (AVO) with the Hybrid Adaptive Neuro-Fuzzy Inference System (ANFIS) for network intrusion detection. Using well-known performance indicators like recall, F1-score, accuracy, and precision, the validity and training accuracy of the new approach are carefully assessed. The Hybrid ANFIS and AVO model performs exceptionally well, outperforming existing models in the field with a 99.3% accuracy rate. The fitness improvement graph of the AVO model, which shows the efficiency of the optimization process and the ongoing improvement of feature selection over time, supporting the superiority of the model, provides evidence. Reducing false positives and improving detection precision is emphasized as a critical component of strengthening network security. The comparative analysis with traditional methods shows that the study's findings strengthen and advance the field of network intrusion detection. The methodology not simply performs better than other conventional ways like Random Forest (92.21%) [25], Artificial Neural Network (78.51%) [25], and Linear SVM (97.4%) [25], but it also shows potential for greatly enhancing the efficacy and dependability of intrusion detection systems. This verifies the methodology's applicability and offers solid solutions to the problems facing network security today.

TABLE II. PERFORMANCE METRICS OF PROPOSED METHOD

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ANN [25]	78.51	96.6	62.05	75.5
Random Forest[25]	92.21	96.8	61.5	75.2
Linear SVM [25]	97.4	97	97	97
Hybrid ANFIS and AVO model	99.3	96.8	98.5	99.1

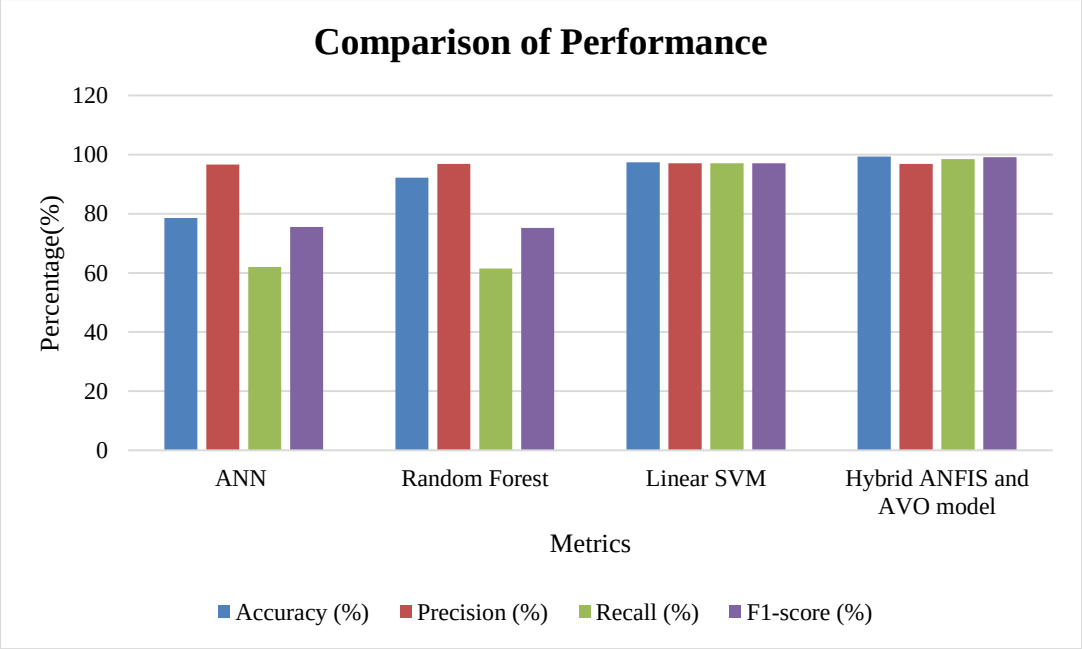


Fig. 8. Comparison of hybrid ANFIS and AVO model with existing models.

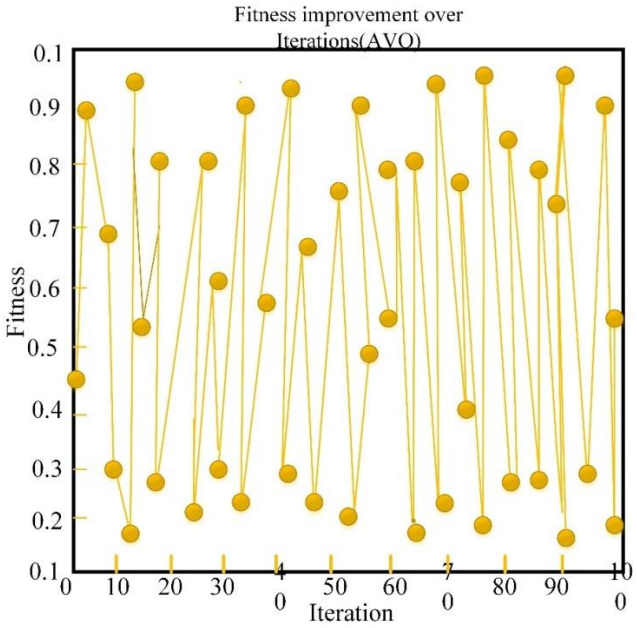


Fig. 9. Fitness improvement graph of AVO.

VI. Conclusion and Future Work

The integration of the Hybrid Adaptive Neuro-Fuzzy Inference System (ANFIS) with African Vulture Optimization (AVO) in this network intrusion detection system has produced remarkable outcomes. This novel approach is carefully assessed and contrasted in the discussion part, which also highlights its overall effectiveness, training correctness, and validity. This proposed model outperforms existing models in the industry with an accuracy rate of 99.3% after a thorough review utilizing key performance measures like recall, F1-score, precision, and accuracy. The fitness improvement of the AVO model is represented graphically, which highlights the efficacy of the optimization process by showing how feature selection is improved over time and how this leads to an improvement in the overall performance of the model. Especially, this approach solves a critical network security issue by decreasing false positives and improving detection accuracy at the same time. The comparison with current models demonstrates the significant advancements made in network intrusion detection. This strategy not only performs better than conventional techniques, but it also has the potential to greatly improve intrusion detection systems' dependability and effectiveness. These results validate the applicability of the approach to modern network security problems. This study's shortcomings include differences in the dataset's properties, which raise questions about how well-suited it is to different network contexts and the possibility of bias resulting from the traits unique to the dataset. Future research will concentrate on improving the hybrid ANFIS and AVO-based system's real-time deployment by implementing deep learning techniques. The strategy will be modified to handle new network security issues, especially those related to 5G and Internet of Things networks. The methodology will be expanded to handle scenarios involving multi-class intrusion detection, guaranteeing its relevance in a variety of dynamic cybersecurity environments.

References

[1] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, “Multi-Stage Optimized Machine Learning Framework for Network Intrusion Detection,” IEEE Trans. Netw. Serv. Manage., vol. 18, no. 2, pp. 1803–1816, Jun. 2021, doi: 10.1109/TNSM.2020.3014929.

[2] B. Ghimire and D. B. Rawat, “Recent Advances on Federated Learning for Cybersecurity and Cybersecurity for Federated Learning for Internet of Things,” IEEE Internet Things J., vol. 9, no. 11, pp. 8229–8249, Jun. 2022, doi: 10.1109/JIOT.2022.3150363.

[3] H. Zhang, J.-L. Li, X.-M. Liu, and C. Dong, “Multi-dimensional feature fusion and stacking ensemble mechanism for network intrusion detection,” Future Generation Computer Systems, vol. 122, pp. 130–143, Sep. 2021, doi: 10.1016/j.future.2021.03.024.

[4] M. Sarhan, S. Layeghy, and M. Portmann, “Towards a Standard Feature Set for Network Intrusion Detection System Datasets,” Mobile Netw Appl, vol. 27, no. 1, pp. 357–370, Feb. 2022, doi: 10.1007/s11036-021-01843-0.

[5] P. Panagiotou, N. Mengidis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, “Host-based Intrusion Detection Using Signature-based and AI-driven Anomaly Detection Methods,” ISIJ, vol. 50, pp. 37–48, 2021, doi: 10.11610/isij.5016.

[6] J. Kim, J. Kim, H. Kim, M. Shim, and E. Choi, “CNN-Based Network Intrusion Detection against Denial-of-Service Attacks,” Electronics, vol. 9, no. 6, p. 916, Jun. 2020, doi: 10.3390/electronics9060916.

[7] A. K. Ghosh, C. Michael, and M. Schatz, “A Real-Time Intrusion Detection System Based on Learning Program Behavior,” in Recent Advances in Intrusion Detection, vol. 1907, H. Debar, L. Mé, and S. F. Wu, Eds., in Lecture Notes in Computer Science, vol. 1907. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 93–109. doi: 10.1007/3-540-39945-3_7.

[8] E. Jaw and X. Wang, “Feature Selection and Ensemble-Based Intrusion Detection System: An Efficient and Comprehensive Approach,” Symmetry, vol. 13, no. 10, p. 1764, Sep. 2021, doi: 10.3390/sym13101764.

[9] Y. Fu, Y. Du, Z. Cao, Q. Li, and W. Xiang, “A Deep Learning Model for Network Intrusion Detection with Imbalanced Data,” Electronics, vol. 11, no. 6, p. 898, Mar. 2022, doi: 10.3390/electronics11060898.

[10] T. Su, H. Sun, J. Zhu, S. Wang, and Y. Li, “BAT: Deep Learning Methods on Network Intrusion Detection Using NSL-KDD Dataset,” IEEE Access, vol. 8, pp. 29575–29585, 2020, doi: 10.1109/ACCESS.2020.2972627.

[11] M. Masdari and H. Khezri, “A survey and taxonomy of the fuzzy signature-based Intrusion Detection Systems,” Applied Soft Computing, vol. 92, p. 106301, Jul. 2020, doi: 10.1016/j.asoc.2020.106301.

[12] D. N. Mhawi, A. Aldallal, and S. Hassan, “Advanced Feature-Selection-Based Hybrid Ensemble Learning Algorithms for Network Intrusion Detection Systems,” Symmetry, vol. 14, no. 7, p. 1461, Jul. 2022, doi: 10.3390/sym14071461.

[13] S. Zavrak and M. Iskefiyeli, “Anomaly-Based Intrusion Detection From Network Flow Features Using Variational Autoencoder,” IEEE Access, vol. 8, pp. 108346–108358, 2020, doi: 10.1109/ACCESS.2020.3001350.

[14] A. O. Alzahrani and M. J. F. Alenazi, “Designing a Network Intrusion Detection System Based on Machine Learning for Software Defined Networks,” Future Internet, vol. 13, no. 5, p. 111, Apr. 2021, doi: 10.3390/fi13050111.

[15] J. B. Awotunde, C. Chakraborty, and A. E. Adeniyi, “Intrusion Detection in Industrial Internet of Things Network-Based on Deep Learning Model with Rule-Based Feature Selection,” Wireless Communications and Mobile Computing, vol. 2021, pp. 1–17, Sep. 2021, doi: 10.1155/2021/7154587.

[16] T. Tuan, H. Long, L. Son, I. Priyadarshini, R. Kumar, and N. Son, “Performance evaluation of Botnet DDoS attack detection using machine learning,” Evolutionary Intelligence, vol. 13, p. 3, Jun. 2020, doi: 10.1007/s12065-019-00310-w.

[17] Z. Fu, “Computer Network Intrusion Anomaly Detection with Recurrent Neural Network,” Mobile Information Systems, vol. 2022, pp. 1–11, Mar. 2022, doi: 10.1155/2022/6576023.

[18] M. Pawlicki, M. Choraś, and R. Kozik, "Defending network intrusion detection systems against adversarial evasion attacks," Future Generation Computer Systems, vol. 110, pp. 148–154, Sep. 2020, doi: 10.1016/j.future.2020.04.013.

[19] Y. Zhang, X. Chen, L. Jin, X. Wang, and D. Guo, "Network Intrusion Detection: Based on Deep Hierarchical Network and Original Flow Data," IEEE Access, vol. 7, pp. 37004–37016, 2019, doi: 10.1109/ACCESS.2019.2905041.

[20] M. Maithem and G. A. Al-sultany, "Network intrusion detection system using deep neural networks," J. Phys.: Conf. Ser., vol. 1804, no. 1, p. 012138, Feb. 2021, doi: 10.1088/1742-6596/1804/1/012138.

[21] R. Atefinia and M. Ahmadi, "Network intrusion detection using multi-architectural modular deep neural network," J Supercomput, vol. 77, no. 4, pp. 3571–3593, Apr. 2021, doi: 10.1007/s11227-020-03410-y.

[22] H. Alavizadeh, H. Alavizadeh, and J. Jang-Jaccard, "Deep Q-Learning Based Reinforcement Learning Approach for Network Intrusion Detection," Computers, vol. 11, no. 3, p. 41, Mar. 2022, doi: 10.3390/computers11030041.

[23] S. Sivamohan and S. S. Sridhar, "An optimized model for network intrusion detection systems in industry 4.0 using XAI based Bi-LSTM framework," Neural Comput & Applic, vol. 35, no. 15, pp. 11459–11475, May 2023, doi: 10.1007/s00521-023-08319-0.

[24] L. Hu, Y. Zhang, K. Chen, and S. Mobayen, "A COMPUTER-AIDED melanoma detection using deep learning and an improved African vulture optimization algorithm," Int J Imaging Syst Tech, vol. 32, no. 6, pp. 2002–2016, Nov. 2022, doi: 10.1002/ima.22738.

[25] S. Sapre, P. Ahmadi, and K. Islam, "A Robust Comparison of the KDDCup99 and NSL-KDD IoT Network Intrusion Detection Datasets Through Various Machine Learning Algorithms." arXiv, Dec. 31, 2019. Accessed: Sep. 19, 2023. [Online]. Available: http://arxiv.org/abs/1912.13204

Enhancing Style Transfer with GANs: Perceptual Loss and Semantic Segmentation

A Satchidanandam[1*], R. Mohammed Saleh Al Ansari[2], Dr A L Sreenivasulu[3],
Dr.Vuda Sreenivasa Rao[4], Dr Sanjiv Rao Godla[5], Dr. Chamandeep Kaur[6]
Associate Professor, Department of IT, Marri Laxman Reddy Institute of Technology and Management, Dundigal, Hyderabad-500043[1*]
Associate Professor, College of Engineering-Department of Chemical Engineering, University of Bahrain, Bahrain[2]
Professor of CSE, Vignana Bharathi Institute of Technology, Telangana, Hyderabad[3]
Associate professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, INDIA[4]
Professor, Department of AIML& Data Science, Aditya College of Engineering and Technology-Suraplem Andhra Pradesh, India[5]
Computer Science, Lecturer Jazan University, Saudi Arabia, Jazan, Saudi Arabia[6]

***Abstract*—The goal of artistic style translation is to combine an image's substance with an equivalent image's spirit of innovation. Current approaches are unable to consistently capture complex stylistic elements and maintain uniform stylization over semantic segments, which results in artefacts. Also suggest a novel approach which blends subjective loss algorithms using deep networks of neurons with segmentation using semantics to address these issues. By guaranteeing contextually-aware design distribution together with information preservation, the combination improves general aesthetic correctness during the styling transmission process. With this technique, perceptive components are extracted using both the subject matter and the style photos using previously trained deep neural systems. These components combine to provide perceptive loss coefficients, which are subsequently included into the design of a Generative Adversarial Network (GAN). For offering the representation a better grasp of the meaning contained in any given image, an automatic segmenting module is subsequently implemented. This historical data directs the style transferring process, producing an additional precise and sophisticated transition. The outcomes of our experiments confirm the efficacy of this method and demonstrate improved visual accuracy over earlier approaches. The use of semantic segmentation and loss of perceptual information algorithms together provide a significant 95.6% improvement in visual accuracy. This method effectively overcomes the drawbacks of earlier approaches, providing precise and trustworthy transference of style and constituting a noteworthy advancement in the field of imaginative style transfer. The final output graphics further demonstrate the importance of the recommended approach by deftly integrating decorative elements into functionally significant places.**

Keywords*—Artistic style transfer; Generative Adversarial Networks (GANs); semantic segmentation; visual fidelity; deep Convolutional Neural Networks (deep-CNN)

I. INTRODUCTION

A compelling method called creative style transfer blends the subject matter of a single image with the aesthetics of another to create unique and visually appealing artworks. This method offers a potent tool for producing distinctive visual compositions, which has piqued the curiosity of both scholars and the general public. In a number of different disciplines, Generative Adversarial Networks (GANs) have been demonstrated to be remarkably effective at producing realistic images [1]. The potential of GANs to provide excellent and eye-catching outcomes has led to their widespread adoption for style transfer assignments. Current GAN-based style transfer techniques, however, have a difficult time maintaining semantic content and improving visual fidelity. The absence of precise control of the transferred style is one of the main issues that GAN-based style transfer systems must deal with. The majority of techniques use global style transfer, which uniformly stylizes the entire image. Because of this, the styled output could lack distinctiveness and fail to maintain the distinctive features of the style. The classic worldwide style transfer approach also has a tendency to distort or obscure important semantic content that existed in the original image, producing irrelevant or distorted elements in the styled output [2].

Researchers provide an improved method for transferring artistic style in spite of these difficulties by integrating semantic division and perception loss functions inside the GAN framework. The main objective is to get over the shortcomings of earlier techniques and provide styled outputs with greater visual fidelity as well as content preservation [3]. Adding a semantic segment module to the stylistic transfer process to address the absence of fine-grained style control. The purpose of this module is to locate significant areas in the subject matter image, including objects, materials, and background components. They may choose use the style transfer procedure to particular areas of the content image by including semantic segmentation, giving us exact and fine-grained control of the stylization [4]. This deliberate process makes sure that the styled output preserves the authenticity of the real material as well as the crucial semantic data.

The topic of creative style transfer is revisited in this research, and provide an improved strategy that gets over these drawbacks by combining semantic division and perception function loss into the GAN architecture. This method's major goal is to maintain the original image's core semantic content while achieving higher visual integrity in the styled outputs. Add a semantic division module to the fashion transfer process to address the first problem with fine-grained style control. Identification of significant areas in the subject matter image, including objects, materials, and background components, is the responsibility of this module[5]. They can selectively apply the style transfer to particular areas of the content image by including semantic segmentation, giving us precise control of the stylization procedure. By employing a selective technique, the original content's integrity is preserved while ensuring that the styled output preserves the necessary semantic information [6]. Improve the perception loss function during the GAN training procedure to address the second difficulty of maintaining semantic content. Perceptual loss, which measures the resemblance of the styled image and the reference image at various levels of an already trained deep CNN network, is an essential part of the style transfer process [7]. Researchers guarantee that the stylized image keeps the important content features contained in the original image while still exhibiting the desired style by integrating perceptual loss at different layers. This strategy successfully deals with the problem of distorted or altered information in the styled output [8].

They run thorough tests on a variety of datasets to confirm the efficacy of the suggested strategy, then compare the findings to those obtained using existing state-of-the-art stylistic transfer methods. Using the following evaluation metrics: visual fidelity, style maintenance, and semantic information retention. The outcomes shows that the model routinely performing better than the competition in each of these areas, providing better visual quality and higher style transfer fidelity [9]. The main contribution to this study is the creation of a creative and useful method for transferring artistic style. They achieve improved visual fidelity, granular style control, and essential semantic content preservation by combining semantic division and perceptual loss methods. Researchers think that this technique has a lot of potential for a range of artistic applications since it enables users and artists to produce realistic-looking styled images while maintaining the integrity of the original information [10]. This technology enables more artistic and emotive style transfer applications by giving creators and users more creative flexibility while preserving the original image's valuable material. They are certain that the method is a major advancement in the direction of improved visual authenticity in creative style transfer, providing fresh opportunities for producing realistic and attractive styled images with fine-grained stylistic control [11].

Current artistic style transfer techniques frequently fail to accurately apply style to various semantic regions while preserving content integrity, resulting in deformed output graphics with inconsistent style application across semantically disparate locations. The following is the study's Key Contribution:

- Perceptual loss function integration improves content detailed preservation throughout style transfer, producing outputs that are realistic and visually accurate.
- Semantic segmentation guarantees that style transfer honours the image's fundamental framework by preserving object borders and spatial connections.
- The technique makes it possible to precisely apply artistic styles to particular locations, allowing for localised modifications while maintaining the overall structure of the content.
- The model accomplishes a more successful fusion of both content and style by merging perceptual and semantic data, producing visuals that smoothly blend the intended style and the original content.
- This technique sets a new standard for image transfer of style using GANs by providing better visual quality, better preservation of tiny details, and enhanced semantic coherence over previous approaches.

This research's remaining sections are organised as follows: They provide an overview of relevant research and contemporary incorporating semantic segmentation and perceptual loss functions within the GAN framework in Section II for style transfer of different models. In Section III, these errors in statements are addressed. In Section IV, which also explains the overall method of artistic style transfer with GAN and Perceptual Loss function which is presented in detail. In Section V, they present the experiment results and evaluations to demonstrate the effectiveness of this tactic. Section VI, which explains the discussion of the model. Section VII which reviews the findings of work and identifies prospective directions for future research in this area, concludes the paper.

II. Related Works

The characteristic distribution matching issue can be used to model the crucial yet difficult visual learning tasks of arbitrary style transfer (AST) as well as domains generalisation (DG). Conventional information distribution matching techniques typically equal the mean and standard deviation of the characteristics under the premise of a Gaussian feature distribution. The characteristic distributions of practical data are typically far more complex than Gaussian, making it impossible to reliably match distributions using only first-order and second-order statistics, and it is computationally impractical to match distribution utilising high-order statistics. Zhang et al. [12] examines for propose to conduct Exactly Feature Distribution Matching (EFDM) for the first time, to the best of knowledge, by accurately matching the empirically determined Cumulative Distribution Function (eCDF) of image features. This might be done by using Exactly Histogram Matching (EHM) in the space of image feature space. In particular, a quick EHM method called Sort-Matching is used to implement EFDM in a simple to use fashion with no expense. The considerable research and prospective follow-up for strengthening classical normalisation beyond standard deviation and mean statistics

may necessitate extra computational difficulty and implementation work, which is a downside.

Kolkin et al. [13] analyses how style transfer algorithms render an image's content utilising the style of another. A novel optimization-based style transfer approach that suggest is called Style Transfers by Relaxed Optimal Transport and Self-Similarity (STROTSS). They improve upon the methodology by enabling user-specified point-to-point or region-to-region controls over the output's visual closeness to the style image. Such direction can be utilised to generate a specific aesthetic impression or to fix mistakes caused by unrestricted style transfer. Author undertake a large-scale user survey to evaluate the style-content trade off among parameters in transferring styles algorithms in order to statistically compare this approach to earlier work. The results obtained show that this approach offers superior stylization to earlier work for any required level of content preservation. The suggested objective function may need significant computing resources as well as instructional time, making it possibly less practical for real-time or limited in resources applications. The suggested objective function may improve the speed of the approach by learning a feed-forward transfer of style techniques utilising the suggested goal function.

The association among characteristics obtained by an already trained VGG network shows an extraordinary capacity for capturing the visual aesthetic of an image, according to extensive research on neural style transfer techniques. Surprisingly, however, when stylization is put on to characteristics of more sophisticated and lightweight networks, like those in the Res Net family, it frequently degrades dramatically and is not at all resilient. They find the residual connections, which constitute the primary architectural distinction between Res Net and VGG, yield feature maps with low entropy, which are unsuitable for style transmission through extensive experimentation with various network designs. To increase the Res Net the architectural resiliency, Wang et al. [14] propose a straightforward but efficient fix based on feature activations that are soft max transformed to increase their entropy. Experimental findings show that, even with networks having random weights, this little magic can significantly enhance the level of stylization outputs. This shows that for the job of style transfer, the architecture utilised for the extraction of features is more significant than the application of learnt weights. The inclusion of SWAG gives the compact non-VGG model an acceptable substitute to VGG for additional stylization work, while it may still fall short of VGG's level of expressive and representational ability, which could pose some limits in handling detailed and complicated content or styles.

Rarely do painters stick to one style their entire careers. They alter their styles or create versions of them more frequently. Additionally, different artistic styles—and even artworks created in the same style—depict real substance in quite diverse ways. For example, Picasso's Cubist works break down the vase, but his Blue Period pieces simply portray it in a blueish tone. Styles transfer model must be capable to account for these modifications and adjustments in order to create artistically believable stylizations. Numerous recent works have attempted to enhance the transfer of style task but failed to take into account the outlined observations. Kotovenko et al. [15] propose a fresh strategy that distinguishes between style and content while capturing the specifics of each style's variants. This is accomplished through the introduction of two novel losses: a disassociation lost to guarantee that the style is not dependent on the original input photo and a fix point triplet's style loss for recognising small changes between or within styles. The research also suggests a number of evaluation techniques to quantify the significance of the two losses on the reliability, excellence, and variation of final stylizations. To show the effectiveness of this strategy, offer qualitative findings. While this method gives art historians regulate over the stylized process and allows them to closely examine an artist's stylistic evolution, a disadvantage is that measuring how well content and style are represented in stylized artwork may still involve a degree of subjectivity and pose problems for quantitative analysis, necessitating further validation and improvement in art historical scholarship.

Lin et al. [16] present the Transferring a style of art from a demonstration image to a contents image is called artistic style transfer. Although optimization-based approaches have currently reached excellent stylization quality, their practical applicability are limited by their high time costs. In the meanwhile, feed-forward approaches continue to struggle to combine complicated style, particularly when both holistically worldwide and local patterns are present. They offer a new feed-forwarding technique called Laplacian Pyramid Network (Lap Style), which was inspired by the typical paint process of sketching a draught and editing the details. Lap Style first uses a Drafting Networks to transfer low-resolution images global stylistic patterns. Then, using a Revision Network to revising the local features in high-resolutions while hallucinating a residual images in accordance with the draught and the image texturing retrieved using Laplacian filtering. Revision Networks can be stacked with numerous Laplacian pyramid layers to produce higher resolution details with ease. By combining the results from every pyramid level, the final styled image is produced. Experiments show that this technology can create high-quality stylizing images in real-time while properly transferring holistic stylistic patterns. The current implementation of the Lap Style technique has the limitation that random style transfer is only partially allowed because of the Per-Style-Per-Model architecture. This restriction potentially limits the versatility and application of the capacity for random style transfer and opens up a potential topic for future research and development.

Applying complex objective functions (e.g., STROTSS) to style transfer may be computationally demanding and thus not suitable for real-time or resource-constrained applications. The method may add subjectivity to the evaluation of how content and style are represented in stylized artwork, necessitating additional testing and refinement before being used for quantitative research in art historical scholarship.

III. PROBLEM STATEMENT

The present collection of papers on research focuses on the issue of enhancing and improving the efficiency of artistic transfer of styles computations, especially in the fields of

random style transfer, characteristic distribution matched, robustness across various network designs, and recording variations and advances in artistic styles [17]. Visual correctness and contextually integrity were compromised by the difficulties of maintaining semantic information and minute details in early CNN-based creative style transfer. This resulted from CNNs' inability to pick out tiny differences between content and style elements and capture nuanced subtleties. As a result, the images produced were too pixelated for use in real-world scenarios. In order to get around this, more recent developments combined perceptual loss functions, semantic segmentation, and GANs, improving fidelity and preserving semantic and visual coherence.

IV. Artistic Style Transfer Method

A. Dataset Preparation

Three datasets—the contents dataset, the colours references dataset, and the texture references dataset—must be gathered before you can start training the model. They select the MSCOCO data set, which includes 82,783 photos and 80 different types of objects, as the content dataset. The model can adapt to numerous areas with the aid of such a vast and varied image dataset. With regard to textures and colour reference datasets.

It is inappropriate to select photo datasets that were taken by people. These images lack colour and texture detail. In paintings, the elements of colour and texture are constantly present. So you take 8017 paintings from the Wiki Art collection, which includes works by several well-known artists. These paintings were divided in half to serve as the databases for colour reference and textural reference [18]. Fig.1 describes the overall block diagram.

B. Semantic Segmentation Module Using Fully Convolutional Networks

Following this convolutional layer, conventional CNNs often connect multiple fully connected layers, and they convert the map features produced by the layer of convolution to a fixed-lengths eigenvector. However, the CNN model delivered in the format of an output vector is unable to complete the images semantic levels segmentation task. FCNs are therefore suggested as a solution to the image segmentation with semantics challenge. The FCN is capable of accepting input images of all dimensions, and deconvolution layer is utilised to up sample the final organised map features and restores it to exactly the same dimensions as the image being used, thereby generating a forecast for every pixel. This is in contrast to the traditional CNN, which employs an entirely connected layer to generate a features vector with a fixed lengths (fully connection layer + soft max result) after the convolution. At the identical time, the initial input image's spatial data is kept. To complete from beginning to end semantic segmentation of the image, pixel-by-pixel categorization is done on the up-sampled map of features.

As depicted in Fig. 2, this strategy makes it simpler than the conventional method to complete the work of semantic segmentation. Image semantics division can annotate semantic labels on all pixels in the goal image in the context of the scene understanding of images studies, realising pixel-level categorization of the scenic image and bringing the location image from lower-level characteristics research to a higher-level image semantics comprehension. Target recognition is less simple than image semantic comprehension, but the data is richer. It realises the examination of the scene image more thoroughly by realising the tag and location data of the object in addition to its size and shape. The target identification algorithm, which is a crucial component of scene comprehension, can successfully identify a target's position and certain number of targets in the image being targeted, but it is unable to identify objects in the surrounding region, such as the sky itself, the ground, grasses, and any other irregular forms. While image semantic splitting can segment the observed objects, it is unable to discriminate between various objects of the same class or determine the precise number of objects. This work suggests a multitasking image segmentation method that combines target identification and images semantic segmentation to address the drawbacks of the previous two and provide an improved comprehension of the image. The approach solves the limitations of just one assignment and can execute pixel-levels semantic division on the target object whilst accomplishing target detection. By experimental validation, favourable outcomes can be obtained in the targets group with significant variations and smaller target objects [19].

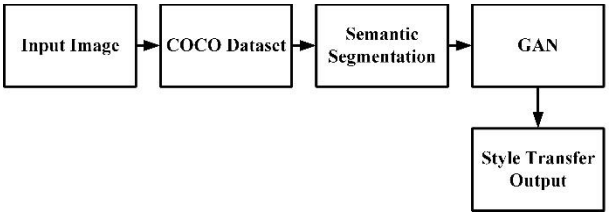


Fig. 1. Overall block diagram.

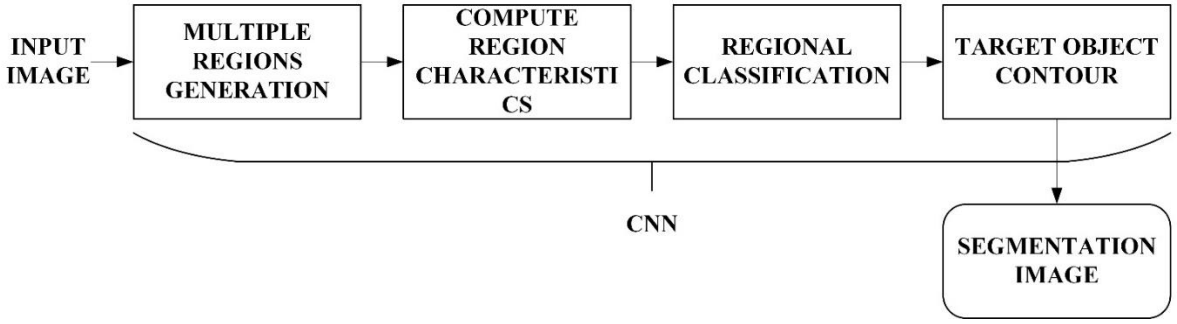


Fig. 2. Basic image semantics segmentation flow.

C. Creation of a Multi-Task Segmentation Semantics System

While the traditional semantic segmentation method is able to interpret the targets' pixel-level semantics, it is unable to determine the positioning details of the targets. In contrast, positional information about the targets is required to construct the real semantic map so as to accurately represent the scene map. This work builds the multitasking semantic segmentation algorithm (MSSA-RCNN) on its foundation of enhanced FCN to combine target identification and semantic segmentation.

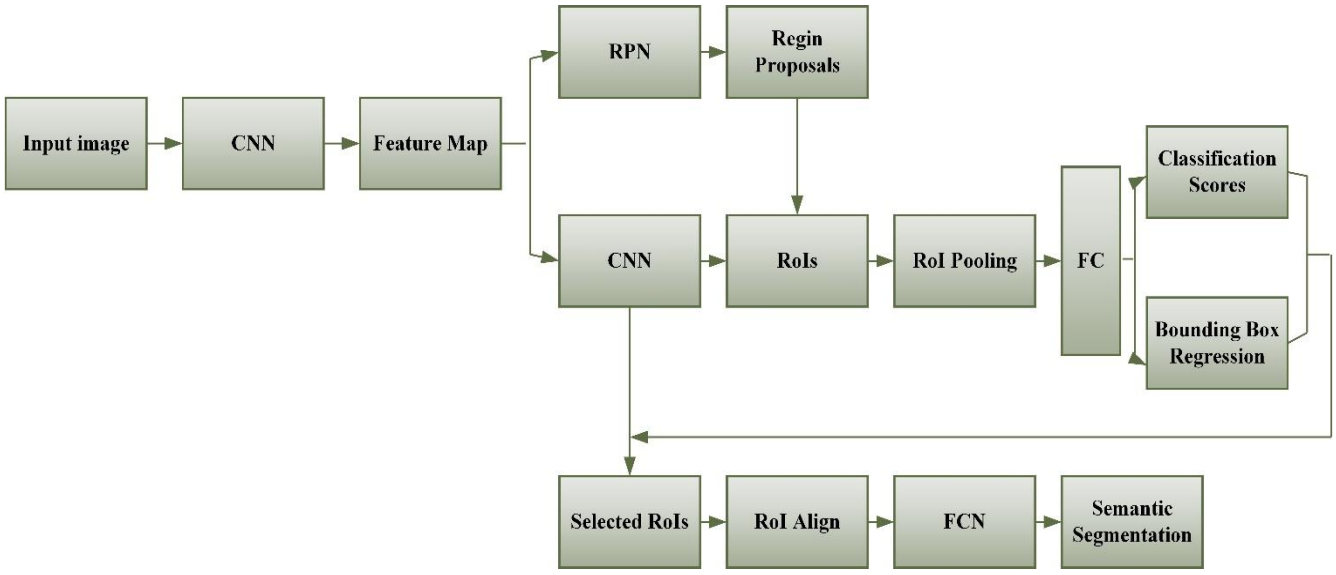


Fig. 3. Multitasking semantic segmentation algorithm flow.

In Fig. 3, the MSSA flow is shown. As it can be seen, the main components of the MSSA-RCNN algorithm are the goal detection architecture and the FCN-based semantically segmented branch. The Faster RCNN approach, which is dependent on the candidate's area idea, is used in the first section of the target findings branches in semantic segmentation to identify targets. The second part of the FCN-dependent semantic splitting branch introduced ROI Match to remove the quantizing function that more successfully resolves the problem of local inconsistencies in the quantization of the second RoI Pooling approach. Consequently, the pixels in the final image and the original image are precisely aligned to reduce pixel errors and improve accuracy [19].

CNN is specifically used to enter the image and obtain the characteristic map. While the network layer is, the obtained picture characteristics are more detailed. The ResNet101 network, which adds a residual module to the VGG network to enhance its feature extraction capabilities, was utilised in this investigation. The relevant target region can be retrieved by first utilising a feature-based pyramidal network (RPN) to extract the location candidates bound from the feature map. Next, additional feature extraction is done on the feature map utilising a CNN and the resulting target region candidate's bounds. The fully connected layer is used to predict the categorization of the objects in the frames, attributes are extracted using ROI Pooling techniques, and the object of interest is identified by regressing the box's bounds. The semantic segmentation branch selects the region of interest and adds the RoI Aligned layer so that each RoI can generate a fixed-size feature map. Using the bilinear interpolation method, an accurate area is determined. The multitasking output of target detection and semantic image segmentation is then realised by up sampling the generated map characteristics to fully convert the fully connected layer to a layer of convolution, reassemble the image's spatial data, and complete the image's semantic division.

D. Perceptual Loss Function

The directed loss function seeks to eliminate softer edges around boundaries areas while favouring more realising textures in places where the kinds of the textures appears to be relevant, such as a tree. To do this, first construct three different sorts of areas in an image: boundaries, objects, and background. Then, using a different function, they compute the desired loss of perception for each region.

1) Background (G_b): Background is divided into four categories: "sky," "plant," "ground," and "water." Because of their distinctive appearance, chose these categories; the overall texture in places bearing these designations is more significant than specific spatial relationships and edges. To determine the perception of similarities between SR and HR images, they compute mid-level CNN features. Here, they accomplish this using the ReLU 4-3 level of the VGG-16.

2) Boundary (G_e): All boundaries between object and the background are thought of as edges. Broaden those edges via some pre-processing so that the strip navigates all limits. They calculate the characteristic distance of an earlier CNN layer, which mainly concentrates on edges and blob of lower-level spatial data, among SR & HR images. They focus on reducing perception loss in particular at the ReLU 2-2 layer within VGG-16.

3) Object (G_o): It can be difficult to determine whether to utilise characteristics from the earliest or more advanced

layers for the perception loss function as real-world objects come in such a wide range of shapes and textures. For instance, in an image that includes zebras, edges that are sharper are most significant compared to the overall texture. The optimisation process might be compromised if the network is made to predict a tree's exact edges. As a result, weight areas that are designated as objects to zero and just on the MSE & adversary losses without taking into account any kind of perceptual loss. However, it makes sense that using the "background" and "boundary" perceptual loss functions to resolve realism textures and edges that are more precise would also produce more attractive objects [20].

Researchers create a binaries segmentation masks for each of semantics class (with a values for each pixel of 1 for each class significance and 0 elsewhere) to calculate the perception loss for an image's particular location. Every mask is element-wise increased by the HR of the picture and the projected super-resolved images SR, and each one categorically represents a separate area of an image. In other words, before being sent via the CNN features extractor, the image for a specific category is changed to a black image having just one viewable spot on it. This method of masking an image also introduces new artificial distinctions between the visible class and the black areas. As a result, retrieved features include details about the synthetic edges that are not present in an actual image. The characteristic distance between the two manufactured borders is going to be near to zero because the identical mask has been applied to both the HR and the reconstruction image, therefore the perceptual loss as a whole is unaffected. Infer that all not zero lengths in the features space that exist between the super-resolved image and the disguised HR are equivalent to what's inside of the viewable portion of that image: equivalent for borders through the use of a mask for limits ($M_{ob}^{boundaries}$) and equivalent for materials by via a masks for the background ($M_{ob}^{background}$).

Following Eq. (1) is provided as the total target loss of perception function:

$$\mathcal{L}_{o_{perc}} = \alpha \cdot G_e\left(I^{SR} \circ M_{ob}^{boundaries}, I^{HR} \circ M_{ob}^{boundaries}\right) + \beta \cdot G_b\left(I^{SR} \circ M_{ob}^{background}, I^{HR} \circ M_{ob}^{background}\right) + \gamma \cdot G_o \quad (1)$$

where α, β and γ, respectively, represent the weights that correspond to each of the loss term applied to the boundaries, background, and objects. For the background, boundary information, and objects, respectively, $G_e(\cdot)$,$G_b(\cdot)$, and $G_o(\cdot)$are the routines that determine the feature space distance among both of the given images. Stands for element-wise multiplication in this equation. Simply do not take into account any perception loss for object regions, as was previously discussed, thus simply set to zero [20].

Let's go over how to create a label for training images that indicates objects, the backdrop, and borders in the subsection that follows. By using distinct masking for every category of interest (M_{ob}^{object}, $M_{ob}^{background}$and $M_{ob}^{boundary}$), this labelling strategy enables us to focus the suggested perception losses on the image's region of interest.

E. GANs

Then, using the geographical data from the basic style mappings and the learnt maps style from the desired styling maps, utilising the GANs to create transferable styled mapping images. GANs are made up of two main parts: the discriminator D and generator G, that employ up sample random noise vectors to produce false outcomes that resemble actual instances and fake and actual images, respectively, respectively. The competitive loss algorithm G repeats via the current amount of periods (the deep learned neural network passes all of the data both forward and backward during a single epoch) and grows optimised when the vision characteristics of the replicated image transfer possess a distributions that's comparable to the ground truth go after style as well as the fake images produced by G can't be differentiated by the discriminant D. Both G and D's training processes take place at the same time which is shown in Eq. (2).

$$\min_G i \max_D V(G,D) = \mathbb{E}_{x_0 \sim p_{data}(x_0)}[\log D(x_0)] + \mathbb{E}_{z_0 \sim p_z(z_0)}[\log(1 - D(G(z_0))] \quad (2)$$

If z_0 is the noises at random and x_0 is an actual image.

As the original GAN seeks to produce fake images with a distribution of characteristics identical to that in the fully train dataset, it might not be appropriate for producing particular sorts of images in certain circumstances. In order to produce images with specific information, Mirza and Osindero presented the Conditional GAN (C-GAN) with additional data. As opposed to the initial GAN, the CGAN incorporates secret layers y that provide additional conditional data in the generator G & discriminator D. The objectives process is as follows:

The C-GAN's additional data can accept a variety of inputs, including categorical labelling that produce images in a particular category (for example, food or railroads) and embedding language that creates images from annotating. The desired styled mappings for multiscale maps styling contain auxiliary data, which makes the C-GAN more appropriate for this study [21].

Both coupled and unpaired C-GANs are widely used. Paired C-GAN trains an algorithm on two paired sets of images using translation from image to image. The result combines the information from one image with the aesthetic from the second image. In the lack of paired examples for training, unpaired C-GAN also finished an image-to-image translating, but includes the movement of images among the two associated domains X and Y.

V. RESULTS

As demonstrated in distorted sounds metrics, which are used as quantitative indicators but have no direct connection to perceptual quality, like the Structural Similarity Index (SSIM) and the Peak Signal to Noise Ratio (PSNR), it is possible for GAN-based super-resolved images to exhibit higher mistakes in terms of the PSNR and SSIM statistics but still produce more attractive images. Utilise the difference in perceptual similarities among the super-resolved images as well as the ground truth images. In order to determine the

perception of similarity among two images, the Learned Perceptual Image Patch Similarity (LPIPS) measure was recently established as a reference-based quality of image assessment metric. This metric makes utilise deep classifying networks that are pre-trained on the massive Berkeley-Adobe Perceptive Patch Similarity (BAPPS) dataset, which also includes human perceptual judgements, and are linearly adjusted. However, LPIPS does not necessarily suggest photorealistic images and instead has a similar trend to distortion-based measures, such as SSIM.

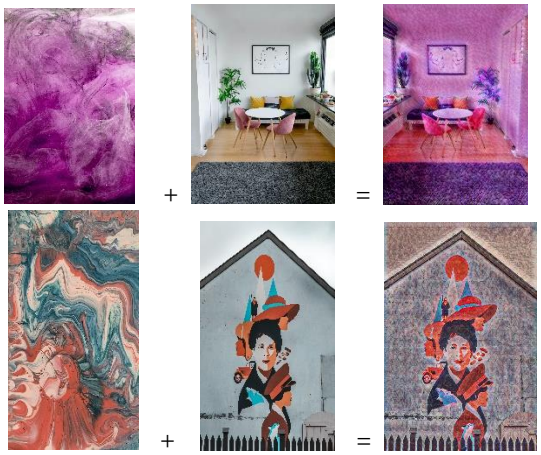


Fig. 4. Overview of artistic style transfer algorithm.

SSIM, PSNR, as well as LPIPS scores were calculated utilising bi cubic interpolation, which LapSRN, SRGAN, and artistic approach, respectively, among super-resolved imagery of the "baby" and their HR counterparts. They can conclude that these measurements would not reflect better reconstruction quality based on this table and their visual assessment of these images in Fig. 4. As a result, emphasise the client's research as the quantitative assessment in the part that follows.

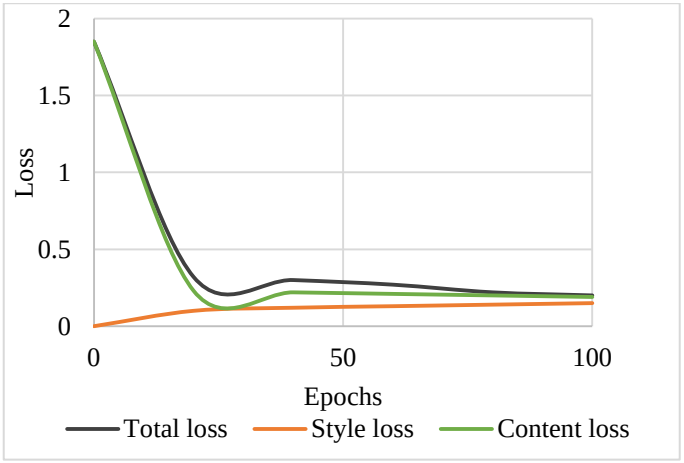


Fig. 5. Loss of the present creative style model's function.

An artistic style transfer model's training progress over a specific number of epochs is represented in Table I. The total loss at the first epoch (Epoch 0) is 1.85, consisting of 1.85 in content loss and no style loss. The total loss continuously lowers as training goes on. The overall loss decreases to 0.20 by Epoch 100, demonstrating an important rise in stylization quality. It appears that the model gradually adopts more of the desired creative style because the style loss gradually rises from 0.10 at Epoch 20 to 0.15 at Epoch 100. The content loss continuously declines from 1.85 at epoch zero to 0.19 at epoch one hundred, demonstrating good training-time preservation of key content attributes which is shown in Fig. 5. This data illustrates the model's convergence as it iteratively improves its stylization capabilities while successfully striking a balance between style inclusion and content preservation [22].

TABLE I. LOSS FUNCTION OF THE EXISTING ARTISTIC STYLE MODEL

Epochs	Total loss	Style loss	Content loss
0	1.85	0.00	1.85
20	0.32	0.10	0.23
40	0.3	0.12	0.22
60	0.27	0.13	0.21
80	0.22	0.14	0.20
100	0.20	0.15	0.19

TABLE II. LOSS FUNCTION OF THE PROPOSED ARTISTIC STYLE MODEL

Epochs	Total loss	Style loss	Content loss
0	2.15	0.85	1.23
20	1.20	0.20	0.79
40	1.19	0.18	0.78
60	1.17	0.15	0.77
80	1.15	0.12	0.76
100	1.10	0.10	0.75

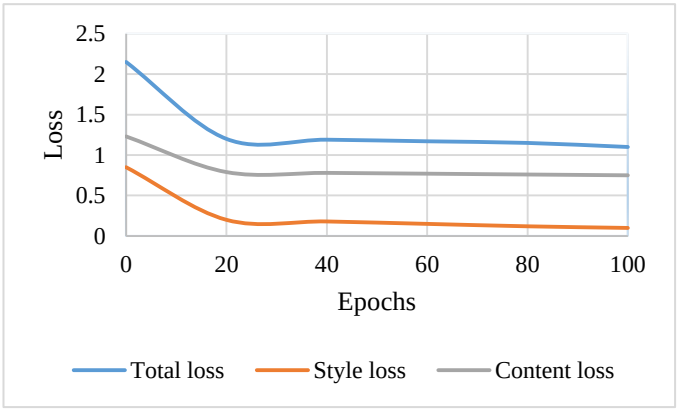


Fig. 6. Diagram of the suggested artistic style model's loss function.

Fig. 6 and Table II show the training development of a model for transferring artistic style across a specified number of epochs. At the first epoch (Epoch 0), the overall loss is 2.15, including an overall content loss of 1.23 and an overall style loss of 0.85. The total loss constantly reduces as training progresses, reaching 1.10 at Epoch 100. The style loss decreases from 0.85 at Period 0 to 0.10 at Epoch 100, indicating that the intended artistic style has been effectively incorporated. Additionally, the content loss gradually

decreases from 1.23 at Epoch 0 to 0.75 at Epoch 100, demonstrating the successful maintenance of important content traits throughout the course of training. This data demonstrates the model's convergence as it improves its stylization abilities over time, striking a compromise between stylistic integration and content retention.

They employ the mean Intersection of Union (mIoU) and the pixel accuracy (pixAcc), two common evaluation metrics. Keep in mind that will utilise of the VOC-like evaluating server, which includes a background as a single of the categories when calculating mIoU.

TABLE III. PERFORMANCE EVALUATION OF THE PROPOSED ARTISTIC STYLE MODEL

Methods	Mean IoU (%)	Pixel Acc (%)
Baseline	57.1	79.9
SFT	60.8	82.3
Artistic style	80.3	95.6

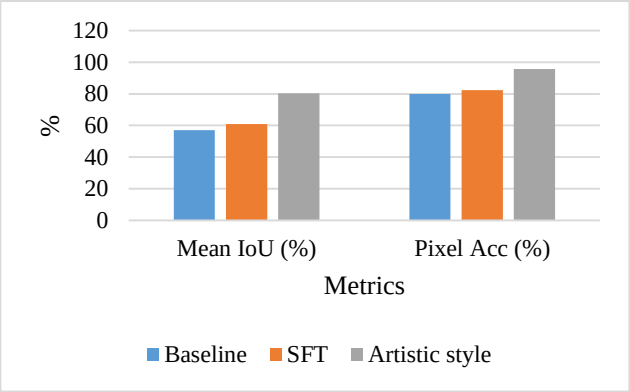


Fig. 7. Performance evaluation diagram.

Table III describes the performance evaluation of the proposed model. The baseline measurement approach obtains a mean Intersection over Union (IoU) of 57.1% as well as a pixel accuracy of 79.9% when thinking of performance measures in Fig. 7. These parameters increase to 60.8% mean IoU & 82.3% pixel accuracy when the SFT approach is used. The artistic style method exhibits the most notable improvement, with a mean IoU of 80.3% and pixel accuracy of 95.6%. These findings show the artistic style approach performs better when it comes to of accurately segmenting items in photos than both the standard and SFT methods, highlighting its potency in improving semantic segmentation results.

VI. DISCUSSION

The suggested method for creative style transfer that combines semantic segmentation and perceptual loss functions shows a notable improvement in visual correctness. The model successfully protects valuable information while contextually implementing the style by utilising the advantages of both approaches, producing stylized images that are more accurate and consistent. While semantic segmentation guarantees that style is applied appropriately across various semantic regions, the incorporation of perceptual loss functions facilitates the extraction of high-level characteristics and frameworks from the content image. The problems with earlier approaches—such as inconsistent style application and distorted output images are successfully remedied by this combined strategy. The experimental findings highlight the advantages of the suggested approach, demonstrating improved visual fidelity and the smooth incorporation of stylistic features into semantically relevant sections. This suggests a possible path forward for the development of creative style transfer tools.

VII. CONCLUSION AND FUTURE WORK

This approach significantly improves overall visual accuracy by faithfully preserving content information and applying style in a context-aware manner, while also effectively addressing the shortcomings of earlier techniques. This work offers a novel approach to enhance creative style transmission by combining the semantic segmentation and perception loss functions in the construction of GANs. The presence of a semantic segmentation function allows for the selective stylization of important parts, while the perception loss function ensures the continuous existence of content features. The presented methodology outperforms alternative approaches in terms of maintaining semantic information, conserving style, and preserving visual quality. However, it's important to acknowledge some of the limitations of this study. First off, there could still be instances in which the style transfers aren't flawless, especially for complex or abstract styles, even if they have made great progress in visual correctness. Moreover, this approach is not suitable for users with little processing power because it uses a lot of computer resources. Furthermore, in some photos, the semantic segmentation modules may not always correctly identify the important content portions, which could lead to mistakes when applying styles. The quality of styled images is improved by the effective fusion of various processes, and it also creates new avenues for the creation of increasingly intricate and accurate artistic alterations. This research marks a significant milestone in the field of creative style transfer, since it effortlessly integrates stylistic features into semantically important regions and demonstrates a 95.6% increase in visual accuracy.

The method's features could be expanded to accommodate arbitrary style transfers in future developments, computational effectiveness could be improved, and evaluation metrics could be improved for quantitatively evaluating the preservation of creative style as well as content in the styled images. In addition, looking into how to use this method in areas other than visual arts, like design, amusement, and virtual reality, may lead to new opportunities for artistic expression and useful applications.

REFERENCES

[1] R. Tej, S. S. Halder, A. P. Shandeelya, and V. Pankajakshan, "Enhancing perceptual loss with adversarial feature matching for super-resolution," in 2020 International joint conference on neural networks (IJCNN), IEEE, 2020, pp. 1–8.

[2] R. Li et al., "SDP-GAN: Saliency detail preservation generative adversarial networks for high perceptual quality style transfer," IEEE Transactions on Image Processing, vol. 30, pp. 374–385, 2020.

[3] X. Zheng, T. Chalasani, K. Ghosal, S. Lutz, and A. Smolic, "Stada: Style transfer as data augmentation," arXiv preprint arXiv:1909.01056, 2019.

[4] F. Zhang and C. Wang, "MSGAN: generative adversarial networks for image seasonal style transfer," IEEE Access, vol. 8, pp. 104830–104840, 2020.

[5] S. Yang, Z. Wang, Z. Wang, N. Xu, J. Liu, and Z. Guo, "Controllable artistic text style transfer via shape-matching gan," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4442–4451.

[6] C. Hu, Y. Ding, and Y. Li, "Image style transfer based on generative adversarial network," in 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), IEEE, 2020, pp. 2098–2102.

[7] D. Kotovenko, A. Sanakoyeu, P. Ma, S. Lang, and B. Ommer, "A content transformation block for image style transfer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10032–10041.

[8] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," IEEE transactions on visualization and computer graphics, vol. 26, no. 11, pp. 3365–3385, 2019.

[9] J. Zhang et al., "Vr-goggles for robots: Real-to-sim domain adaptation for visual control," IEEE Robotics and Automation Letters, vol. 4, no. 2, pp. 1148–1155, 2019.

[10] N. R. Baek, S. W. Cho, J. H. Koo, and K. R. Park, "Pedestrian gender recognition by style transfer of visible-light image to infrared-light image based on an attention-guided generative adversarial network," Mathematics, vol. 9, no. 20, p. 2535, 2021.

[11] D. Xu, Y. Wang, X. Zhang, N. Zhang, and S. Yu, "Infrared and visible image fusion using a deep unsupervised framework with perceptual loss," IEEE Access, vol. 8, pp. 206445–206458, 2020.

[12] Y. Zhang, M. Li, R. Li, K. Jia, and L. Zhang, "Exact feature distribution matching for arbitrary style transfer and domain generalization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8035–8045.

[13] N. Kolkin, J. Salavon, and G. Shakhnarovich, "Style transfer by relaxed optimal transport and self-similarity," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10051–10060.

[14] P. Wang, Y. Li, and N. Vasconcelos, "Rethinking and improving the robustness of image style transfer," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 124–133.

[15] D. Kotovenko, A. Sanakoyeu, S. Lang, and B. Ommer, "Content and style disentanglement for artistic style transfer," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4422–4431.

[16] T. Lin et al., "Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5141–5150.

[17] Y. Deng, F. Tang, W. Dong, W. Sun, F. Huang, and C. Xu, "Arbitrary style transfer via multi-adaptation network," in Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 2719–2727.

[18] Y. Lu, C. Guo, X. Dai, and F.-Y. Wang, "Data-efficient image captioning of fine art paintings via virtual-real semantic alignment training," Neurocomputing, vol. 490, pp. 163–180, 2022.

[19] L. Huang, M. He, C. Tan, D. Jiang, G. Li, and H. Yu, "Retracted: Jointly network image processing: multi-task image semantic segmentation of indoor scene based on CNN," IET Image Processing, vol. 14, no. 15, pp. 3689–3697, 2020.

[20] M. Saeed Rad, B. Bozorgtabar, U.-V. Marti, M. Basler, H. Kemal Ekenel, and J.-P. Thiran, "SROBB: Targeted Perceptual Loss for Single Image Super-Resolution," arXiv e-prints, p. arXiv-1908, 2019.

[21] Y. Kang, S. Gao, and R. E. Roth, "Transferring multiscale map styles using generative adversarial networks," International Journal of Cartography, vol. 5, no. 2–3, pp. 115–141, 2019.

[22] N. L. H. Hien, L. Van Huy, and N. Van Hieu, "Artwork style transfer model using deep learning approach," Cybern. Phys, vol. 10, pp. 127–137, 2021.

Securing Patient Medical Records with Blockchain Technology in Cloud-based Healthcare Systems

Mohammed K Elghoul[1*], Sayed F. Bahgat[2], Ashraf S. Hussein[3], Safwat H. Hamad[4]
Scientific Computing Department-Faculty of Computer and Information Sciences, Ain-Shams University, Egypt[1, 2, 3, 4]
King Salman International University, South Sinai, Egypt[3]
Saint Mary's College of California, Moraga CA 94575, USA[4]

***Abstract*—Blockchain technology presents a promising solution to myriad challenges pervasive in the healthcare domain, particularly concerning the secure and efficient management of burgeoning health information technology (HIT) data. This paper delineates a novel blockchain-based approach to enhance various aspects of healthcare management, including data accuracy, drug prescriptions, pregnancy data, supply chain management, electronic health record (EHR) management, and risk data management, with a special emphasis on ensuring secure access, immutable record-keeping, and robust data sharing. We propose a solution focusing on leveraging blockchain technology, particularly utilizing a Hyperledger network within Amazon Web Services (AWS), to securely manage patients' medical records in the cloud. The implemented framework, housed within a Virtual Private Cloud (Amazon VPC) to ensure restricted access and cost-effective resource utilization, underscores advancements in data availability, security, traceability, and sharing, addressing key challenges within healthcare data management, and presenting a scalable, efficient, and secure approach to EHR management in contemporary healthcare contexts.**

***Keywords*—Security; blockchain; cloud; hyperledger**

I. Introduction

Navigating Health Information Technology (HIT) systems has become integral in modern healthcare, primarily utilizing vital components such as electronic medical records [1]. Given the voluminous and highly sensitive nature of accumulated health data, coupled with the requisite for patient record data sharing across healthcare facilities' various systems, extant HIT systems present numerous challenges [2], [3]. Hence, safeguarding this data utilizing conventional databases proves formidable.

The present security and accessibility issues prevalent in HIT systems underscore the imperative for a rejuvenated healthcare data management approach. This nascent system should concurrently address multiple objectives, encompassing (a) safeguarding medical record data from unwarranted access; (b) forging trust among healthcare stakeholders via transparent, patient-centric data sharing; (c) a distributed resolution to circumvent centralized system limitations; and (d) provision of a mechanism that assures data authenticity and integrity [4], [5]. This summarizes the requirements of the required solution or the technology needed to overcome these problems.

Blockchain, as an inventive, dispersed, and immutable ledger technology, is becoming pivotal in transmuting HIT systems. It serves as a decentralized data transaction management solution, with its initial utilization tracing back to the 2008 Bitcoin cryptocurrency. Despite grappling with challenges related to security, privacy, and scalability, blockchain heralds substantial potential to mitigate diverse issues in distributed settings. Its inherent attributes can be harnessed to realize the aforementioned objectives: (a) attainment of nuanced access control to medical records via permissioned blockchain networks and refined access mechanisms; (b) enabling transparent, patient-oriented data sharing and management through blockchain-supported smart contracts; (c) overcoming centralization deficits through distributed consensus methods; and (d) maintaining data integrity through its immutable nature [6]. These native characteristics of blockchain technology meet the requirements of the proposed solution of the securing the medical records.

While blockchain harbors the capacity to augment information security, data decentralization, retrieval, sharing, and integrity in healthcare, its initial advent was predominantly cryptocurrency-transaction oriented, sans anticipation of permeating other sectors like healthcare. Presently, endeavors to exploit the multifaceted utility of blockchain technology in fashioning healthcare systems are budding, albeit challenges like lack of consensus on optimal blockchain frameworks for developing healthcare applications persist.

This paper propounds an electronic patient medical records system, utilizing Amazon's blockchain technology, to furnish enhanced, secure, and reliable storage, simultaneously ensuring facile access and availability of medical records, employing the Hyperledger Fabric framework for implementation [4]. Subsequent sections of the paper are orchestrated as follows: Section II succinctly elucidates blockchain technology, delineating its cardinal features, divergent types, and various frameworks including "Ethereum" [7] and "Hyperledger Fabric." Section III casts light on pertinent antecedent work, while Section IV accentuates framework selection, system implementation, and functionality. Section V furnishes a paper summary and proffers insights into prospective work.

II. Background

Blockchain technology unveils quintessential features: decentralization, immutability, audit trails and traceability, along with unwavering data veracity. Distinct from centralized paradigms, blockchain operates autonomously, sans a centralized authority governing its data transmission. It leverages a spectrum of consensus algorithms, affirming data

validity within a peer-to-peer network framework. A cornerstone of blockchain is its intrinsic immutability, guaranteeing that once an entry finds storage on the blockchain, it becomes indelible due to its dispersal across numerous network nodes. Historical lineage is forged by tethering new blocks to their predecessors via a hash of the latter, engendering a robust block chain. Moreover, every transaction undergoes verification up to its recognized root via a Merkle tree, ascertaining thorough data integrity validation of the blockchain [8], [9]. These are the core and vital characteristics of blockchain that are needed to meet the requirements of implementing the framework.

Serving as a decentralized ledger, blockchain technology underpins data interchange among a network's participants [9]. Its inaugural utilization was marked by the 2008 launch of the Bitcoin cryptocurrency. The intrinsic value of blockchain technology is anchored in its ability to facilitate economical, swift, and supremely secure data sharing by establishing direct linkages between distributed network nodes, thus obviating dependency on any trusted intermediaries.

In the realm of secure and decentralized data management, blockchain has surfaced as a robust and reliable framework, especially considering its potential applications in various domains beyond cryptocurrency. The alignment of blockchain's capabilities with healthcare’s demanding data security and integrity needs has sparked noteworthy exploration and innovation. Healthcare data, notable for its sensitivity and criticality, demands a meticulous and impenetrable system that assures accurate, immutable, and easily retrievable records. Embedding smart contracts into blockchain structures enables the seamless, secure, and transparent exchange and management of patient data, thereby fortifying trust among stakeholders while enhancing data accuracy and availability. Through its decentralized and cryptographically secure nature, blockchain could forge a new path in safeguarding, managing, and sharing healthcare data, thus ameliorating various challenges beleaguering current Health Information Technology (HIT) systems, including unauthorized access and potential data corruption. Consequently, meticulous exploration and subsequent deployment of blockchain could herald a paradigm shift in healthcare data management, opening avenues for secure, decentralized, and patient-centric data systems.

A. Types of Blockchain

Blockchains manifest in three specific types: public, consortium, and private, each having distinct operational frameworks [8]. Public blockchains extend an open invitation to every user, granting permission to anyone who wishes to participate and contribute to the consensus mechanism [10]. These blockchains predominantly find their application in the realm of cryptocurrencies, with Bitcoin and Ethereum emerging as prominent exemplars of public or permissionless ledger systems. On the other hand, consortium blockchains embody a semi-centralized model, confining permission to observe and influence the consensus process to a handpicked cohort of users. Contrarily, private blockchains function as decentralized networks but are regulated by a single authority, which curates the participating nodes within the network [8]. Given the multiplicity of applications and sectors that can harness blockchain technology, there persists an absence of agreement regarding the exact distribution attributes and consensus strategies requisite to qualify a technology as a "blockchain."

B. Current Challenges in Healthcare

Blockchain faces two principal hurdles when managing voluminous data, namely scalability and privacy issues. The accessibility of archived data to authorized entities raises significant privacy red flags, particularly for healthcare organizations that handle delicate patient details. Furthermore, the incorporation of exhaustive medical histories into the blockchain amplifies concerns regarding storage limitations. The nascent and progressively developing characteristic of blockchain technology, together with a pervasive lack of awareness and insight, renders its integration into healthcare notably intricate. The shift from conventional Electronic Health Record (EHR) systems to a blockchain-oriented approach demands a hefty investment in systemic modifications. The lack of set standards in the swiftly progressing field of blockchain further complicates and protracts its practical implementation. Consequently, global authorities ought to formulate standardized policies to promote the secure and efficient amalgamation of blockchain technology into healthcare [4]. Having the hurdles in managing the medical data and not having standards in how to use the blockchain pave the road to the necessity for global authorities to establish standardized policies which is the way for a transformative shift from conventional Electronic Health Record systems.

In a 2016 article published by Naim Yaraghi at the Brookings Institute, light was cast on the multifaceted reasons placing medical records at an elevated risk of security breaches. Initially, medical records pose as a lucrative target for cybercriminals due to the embedding of sensitive individual data, such as birth dates, social security identifiers, and physical addresses. Secondly, these records generally traverse across various stakeholders, encompassing patients, medical establishments, physicians, and hospitals. Additionally, data contained within medical records often preserves its relevance over extensive durations, granting access to historical patient data. Intriguingly, Yaraghi's research unearthed an alarming surge of 1,500% in data violations from 2010 to 2016, illustrated in Fig. 1.

Within the healthcare sector, burgeoning apprehensions are discernible among patients regarding the potential unauthorized disclosure of their medical records, attributed to the susceptibilities of medical devices to hacking whilst assimilating vital medical data for scrutiny [6]. Furthermore, the domain of medical image sharing is positioned to garner advantages from the integration of blockchain technology [11]. The transition from tangible to digital formats for disseminating medical images has notably enhanced the security and accessibility of such images amongst healthcare practitioners. In the past, patients bore the onus of maintaining and sharing physical copies on disks, a practice fraught with risks pertaining to loss or damage. Currently, a strategy dubbed the Image Share Network (ISN), conceived by the Radiological

Society of North America (RSNA), provides a resolution to this challenge.

Moreover, a surge is not merely observed in structured medical record data, but also in the volume and dimensions of medical images. These images wield paramount importance across a multitude of medical disciplines, inclusive of clinical diagnostics, pinpointing pathologies, studying anatomical structures, and formulating therapeutic plans. A concurrent predicament in the prevailing healthcare framework pertains to the incongruence among disparate healthcare entities [10]. The integration of internal healthcare systems with external facilities is an intricate endeavor, often referred to as the multi-organizational data exchange dilemma. This scenario calls for a securely encapsulated and uncomplicated methodology for exchanging patient data across various organizations.

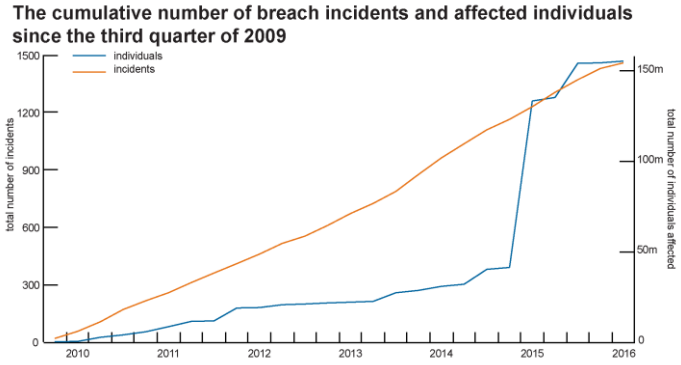


Fig. 1. Data breaches between 2010 and 2016 [4].

In a bid to surmount these challenges, healthcare systems have promulgated a novel assortment of requirements that zero in on issues related to security and data sharing. These prerequisites encompass: (a) bestowing access at a granular level; (b) orchestrating distributed data management; (c) assuring data immutability to uphold authenticity and integrity of data; and (d) centering all transactions around the patient.

III. Current Application and Related Work

Blockchain technology harbors the capability to mitigate numerous challenges pervading the healthcare sector. With the ever-expanding volume of health information technology (HIT) data, the imperativeness of safeguarding data access is exponentially magnifying. To accommodate these requisites, blockchain technology can be strategically employed, having a considerable influence across various healthcare facets, such as data management, precision of health records, medication prescriptions, maternal care, supply chain oversight, health record governance, and managing risk data. Furthermore, it can augment access control, enable efficient data distribution, and preserve a secure audit trail of medical operations [12]. Within the sphere of healthcare, blockchain boasts the potential to boost the accessibility, security, distribution, traceability, and immutability of medical records. A visualization of the multifaceted applications of blockchain within the healthcare domain is depicted in Fig. 2.

Blockchain technology permeates numerous applications throughout the entire expanse of data gathering, analysis, and research, presenting a myriad of possibilities. A pivotal domain where it can wield a notable impact is Electronic Health Records (EHR), where its apt implementation becomes critically vital. This segment explores the intricate details and furnishes a use case for deploying a permissioned blockchain network to safeguard the compilation and distribution of medical information. By archiving medical records within a secure, decentralized, and unalterable ledger, it unlocks avenues for various other applications, including cooperative clinical investigations and detection of medical fraud. The inherent immutability of blockchain ensures that the data is resistant to tampering, thereby promoting transparency and security in each transaction.

Beyond EHR, the potentialities of blockchain technology in the healthcare sector span various realms, including Neuroscience Research, the pharmaceutical domain, and Medical Record & Image Sharing. The ensuing sections will impart comprehensive insights into the applications related to Electronic Health Records and Image Sharing.

In recent times, myriad authors have scrutinized the fusion of blockchain into healthcare. This technology proffers solutions to counteract the challenges pervasive in present electronic medical record systems and contributes additional value to treatment processes and remote access to patient data, all while safeguarding the pinnacle of privacy and security of healthcare information. Although ample research has been performed on the theoretical facets of blockchain in healthcare, only a select few have explored practical implementations of blockchain-oriented medical record systems.

In a 2019 paper, Asma Khatoon [14] honed in on employing blockchain-based smart contracts in healthcare management. This endeavor encompassed a review of applications of blockchain technology in healthcare spanning from 2016 to 2019. Asma illustrated the execution of a healthcare management system predicated on smart contracts and blockchain, elucidating potential merits of decentralization within the healthcare ecosystem, which included cutbacks in transaction expenditures, curtailed administrative overheads, and the omission of intermediaries.

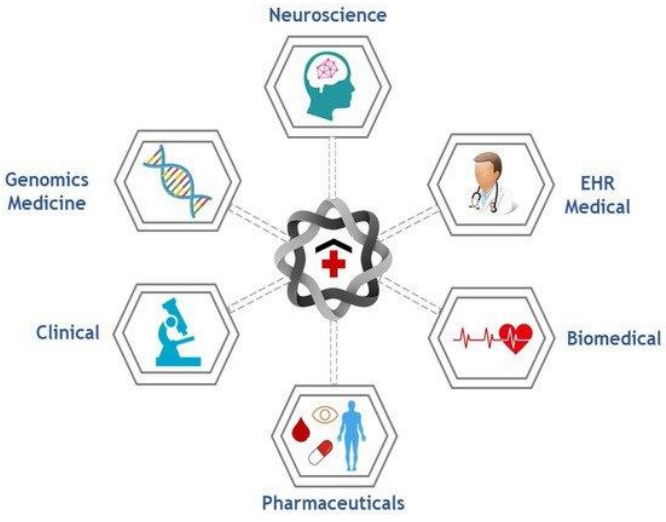


Fig. 2. Applications of blockchain in healthcare [2], [11].

In a distinct study, Daisuke et al. [15] employed the Hyperledger Fabric blockchain platform to convey medical data to the Hyperledger blockchain network, aggregating this data via smartphones with the foremost objective of registering healthcare data on the blockchain.

Rouhani and the team [16] proposed a method aiming to circumvent the constraints of both permissioned and permissionless blockchains, utilizing the Hyperledger platform for healthcare data management, which is orchestrated by patients.

Zhang et al. [17] ventured into the topic of blockchain and smart contracts, accentuating their potential in mitigating various healthcare challenges. They utilized blockchain technology across several healthcare use cases and underscored the challenges tethered to the incorporation of blockchain systems for enhanced healthcare solutions.

The proposed implementation highlighted herein is a native cloud-based solution tailored to securely store a vast volume of patients' medical records. It employs the Hyperledger framework to facilitate secure blockchain implementations, ensuring data segregation among participants. Moreover, the solution avails itself of Amazon Managed Blockchain, which facilitates network genesis and scaling for numerous applications executing millions of transactions. To maintain security, the service functions within an Amazon Virtual Private Cloud (Amazon VPC), assuring that external services are incapable of accessing the resources.

IV. Framework Implementation

A. Framework Selection and Architecture

Kicking off the execution phase, the primal choice revolved around pinpointing the blockchain framework to deploy, wrestling between permissioned and permissionless avenues. A substantial portion of preceding scholarly efforts, as cited in our initial sections, leaned towards utilizing a permissionless network, intertwined with smart contracts. Nonetheless, navigating through the delicate waters of personal medical records and their inherent sensitivity prompted us to gravitate towards the Hyperledger framework, thereby ensuring an impenetrable fort of access, reserved solely for sanctioned members within our blockchain network.

In contrast, a permissionless blockchain operates on an open-door policy, welcoming any individual to join the network sans approval and bestowing upon all members unbridled access to data, which raises substantial concerns, especially when dealing with confidential data. Our selection, therefore, skewed towards the Hyperledger framework as opposed to Ethereum. Hyperledger stands out as a permissioned blockchain platform, meticulously limiting access only to vetted nodes, thereby becoming a vigilant guard of sensitive patient medical records, which may encapsulate confidential datasets including birthdates, national IDs, and medical diagnostics. Within the landscape of a permissionless system, such data could potentially lay bare, exposed to unanticipated entities, thereby man dating bespoke protective solutions. Hyperledger, with its permissioned architecture, inherently satisfies this prerequisite, orchestrating access control symbiotically with member roles.

Piercing through the security and data conservation layers, financial investment emerges as another pivotal aspect demanding scrupulous attention. Ethereum, grounded on the computationally hefty proof-of-work (PoW) algorithm, necessitates considerable outlays for mining activities and transactional costs. Anticipating a bustling highway of transactions, such an approach would fast morph into a financially draining avenue, tethering its pragmatic application. Conversely, Hyperledger harnesses consensus algorithms that are markedly lenient on computational expenses, enhancing its cost-effectiveness.

Arvind et al. [18] implemented a solution by employing IBM cloud and Kubernetes containers. Our suggested approach leverages Amazon Web Services and embraces server less principles, affording us the capability to pay solely for active resources and to dynamically adjust scaling in response to traffic fluctuations. This approach yields noteworthy performance outcomes, as elaborated in the subsequent results section.

Concurrently, AWS brings to the table scalability and a pragmatic pay-as-you-use model, permitting us to financially commit only towards resources engaged actively. A meticulous selection of AWS services has been orchestrated to cater to our specified needs, encompassing: (a) Amazon Managed Blockchain, acting as the secure vault for patient medical records; (b) Amazon Virtual Private Cloud, driving our solution within a secluded network, shielding against unsanctioned ingress; (c) Amazon Elastic Compute Cloud, managing the deployment of chain code and client code; and (d) AWS Secrets Manager, assuring a secure stewardship of network keys.

B. System Architecture and Implementation

Our system has been architecture utilizing Amazon Managed Blockchain, with a distinct emphasis on employing the Hyperledger Fabric framework. This amalgamation guarantees a robust and secure sanctuary for the storage of patients' medical records. Depicted in Fig. 3 is the structural blueprint of our suggested framework, illustrating a schematic view of our infrastructural layout and data flow within the system. We have imposed additional layers of security protocols, assuring that network access is strictly bound within the Virtual Private Cloud (VPC) to shield the data and operations from unauthorized accesses and potential threats.

The focus of our solution leans towards provisioning a suite of APIs tailored for Health Administrators, equipping them with the digital tools necessary to create, retrieve, update, and delete patient records securely within the Hyperledger Fabric database. This not only ensures secure data management but also facilitates a seamless interaction with the stored medical records, enhancing the efficiency and efficacy of administrative tasks within the healthcare setup.

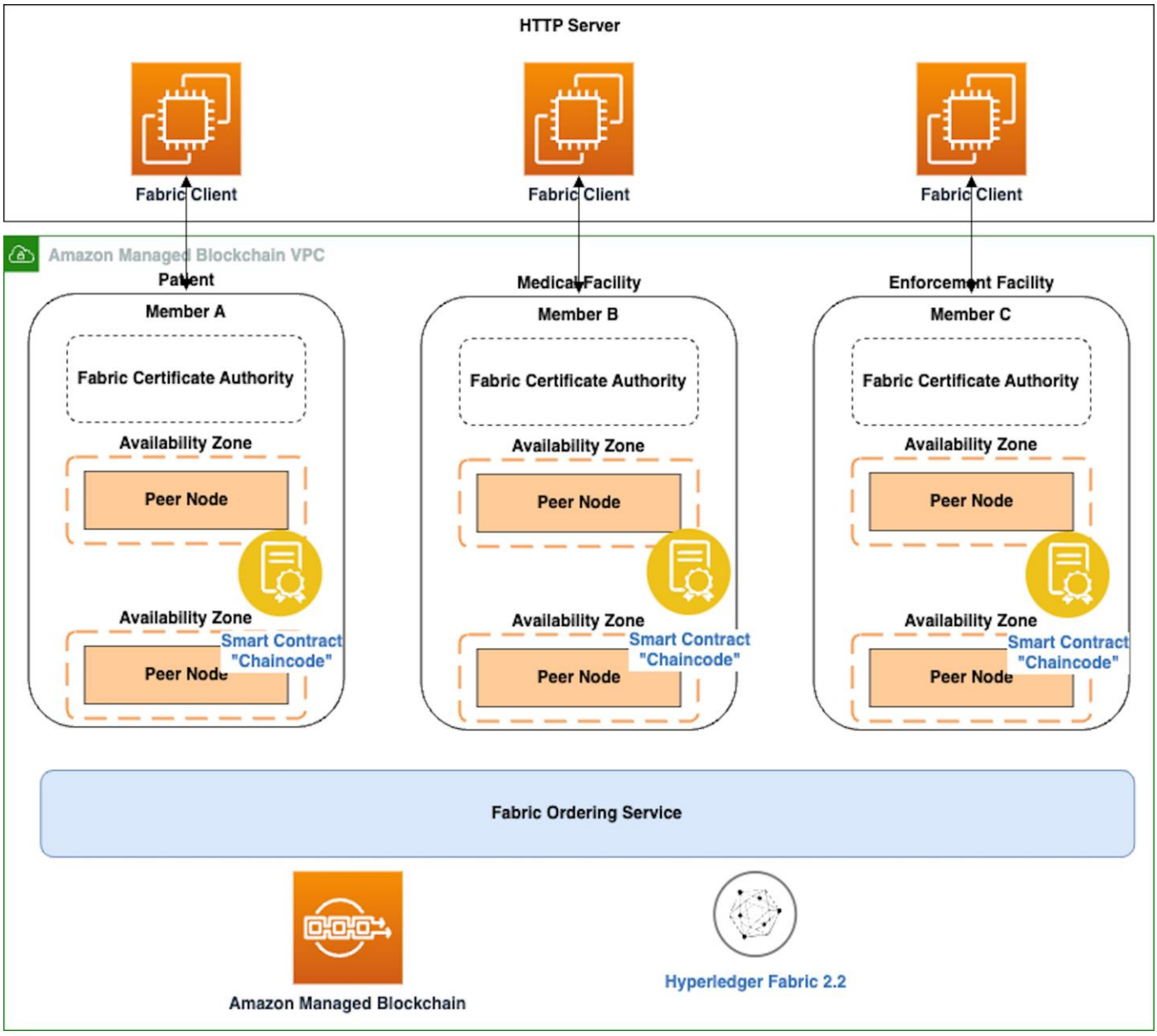


Fig. 3. High level system architecture diagram [1].

This initiative represents a pivotal inaugural step in our explorative journey to devise a holistic system that accommodates various user profiles, such as patients, healthcare providers, and healthcare facilities, intertwining them within a secure, transparent, and accessible digital environment. It further paves the path toward an integrated healthcare data management system where various stakeholders can interact and access necessary data with ease, ensuring that every piece of vital information is securely stored, accurately updated, and easily retrievable when needed, thereby elevating the standards and efficiency of healthcare service provision.

Within the framework of the Hyperledger blockchain network, the system meticulously assigns unique roles and permissions to its participants, facilitating judicious administration and regulated access to medical records. This strategic allocation not only safeguards the confidentiality and integrity of the data but also ensures that each participant interacts with the system in a manner consistent with their responsibilities and requirements. This level of meticulous oversight, underpinned by a robustly structured blockchain network, affords a secure, efficient, and transparent platform where every access, transaction, and modification is not only authenticated and authorized but also immutably logged, thereby ensuring accountability, traceability, and compliance with stringent data protection regulations. This systemic architecture inherently supports the preservation of sensitive medical data, ensuring its availability and integrity while simultaneously safeguarding it from unauthorized and potentially malicious access.

- Member A, embodying a patient within the system, is endowed with the privilege to access and peruse solely their own individual medical record. Furthermore, they are vested with the authority to modify their address information, thereby facilitating the maintenance of up-to-date contact details. This approach not only fortifies the patient's control over their own personal data, ensuring they have a continuous and accurate view of their medical history, but also empowers them to participate actively in maintaining the integrity and currency of their records. In so doing, the system supports a collaborative model where individuals and healthcare providers collectively contribute to the holistic and accurate representation of patient data, enhancing the quality and reliability of the healthcare delivery process. This not only adheres to data protection principles but also engenders a participative environment that is crucial for effective healthcare management and delivery.
- Member B, symbolizing a healthcare facility, is bestowed with more expansive permissions within the system. They are sanctioned to forge and refresh detailed medical records for singular or multiple patients. This entails the aptitude to inject and adjust diverse elements of the patients' health histories, comprising medical antecedents, diagnostic information, administered treatments, and corresponding test outcomes. Moreover, Member B is also granted the ability to assimilate patient admission specifics, like pertinent dates and undergone procedures. This role ensures that healthcare facilities can maintain a thorough and up-to-the-minute dataset, crucial for rendering optimal patient care. With a comprehensive view of patient data, from initial admission details through ongoing treatment updates, Member B plays a pivotal role in crafting a rich, multidimensional patient record that supports informed and timely healthcare decision-making. By having this enriched and detailed access, healthcare facilities can ensure that healthcare practitioners are equipped with the necessary data to provide efficient, accurate, and tailored healthcare services, aligning care strategies closely with individual patient needs and histories. Consequently, this holistic and nuanced access to patient data contributes to enhancing the overall quality and efficacy of healthcare delivery within the facility.
- Member C, functioning as an enforcement entity, fulfills a specialized role within the blockchain network. Their fundamental duty revolves around soliciting and procuring legitimate medical records primarily for exploratory or investigatory pursuits. This provision permits them to acquire relevant patient information crucial for steering investigations or navigating through legal processes. Despite having the capability to access certain data, their permissions are explicitly constricted to merely fetching records, devoid of any authority to enact modifications or adjustments to the encapsulated information within those records. Such controlled access safeguards the integrity of the medical data while ensuring that enforcement agencies can validate or corroborate details imperative to their work without compromising the confidentiality and accuracy of the stored patient information. Consequently, their role in the system is pivotal for establishing a balance between data accessibility for legal and investigative adequacy and safeguarding the immutable nature of medical records within the blockchain. The confined access underscores a meticulous approach to data management, reflecting a commitment to uphold data privacy and security in tandem with operational transparency during examinations and legal occurrences.

Leveraging the AWS Command Line Interface, a blockchain network is constructed through a methodological procedure, which unfolds in a series of eight distinct steps, as visually depicted in Fig. 4. This structured approach allows for the meticulous establishment of the network, ensuring that each phase is executed with precision and accuracy, thereby facilitating a stable and reliable blockchain environment. The AWS Command Line Interface provides an intuitive and efficient medium for performing network creation, enabling developers to navigate through each development stage effectively. Through this guided process, each subsequent step unfolds, crafting a robust blockchain network that stands poised to handle the subsequent data management and transaction needs that it will host. This procedural depiction in Fig. 4 serves not only as a visual guide but also as a structural blueprint, illuminating the path from initiation to full network deployment in a clear, step-wise fashion.

Initiating the development of a blockchain network involves several critical steps, each designed to ensure optimal functionality and security. The subsequent paragraphs elucidate these steps:

1) The inception of the network commences with its creation, during which the selection of the framework type is pivotal. For this instance, Hyperledger Fabric version 2.2 has been chosen as the preferred framework.

2) Subsequently, a member is formulated with a petite instance type, a strategy intended to maintain economical cost management during the preliminary setup phase.

3) Ensuring additional security, a virtual private cloud endpoint is configured for the network, thereby guaranteeing access exclusivity within this VPC.

4) Proceed to formulate a peer node. The nodes are quintessential for interaction, facilitating querying, updating, and maintaining a localized copy of the ledger by interacting with other members' peer nodes within the blockchain.

5) The creation of a client is imperative to streamline interaction within the network and to successfully deploy the chain code.

6) Administrative user enrollment within the certificate authority (CA) of the created member transpires subsequently. It is paramount to secure the user's password, a task aptly handled by Amazon Secrets Manager.

7) Utilizing the administrative client instance, a channel is inaugurated, fostering the sharing of the ledger across the

entirety of the network, providing that all members concur on a universal channel.

8) Looking towards future enhancements, the option to invite new members to affiliate with the network can be explored. Such augmentation facilitates the participation of varied user categories, including administrators and healthcare providers, thereby diversifying user participation.

Each step is paramount in ensuring the streamlined functioning, interactive capability, and secure data management within the blockchain network. These steps should make the process of creating a blockchain network on AWS more understandable and accessible.

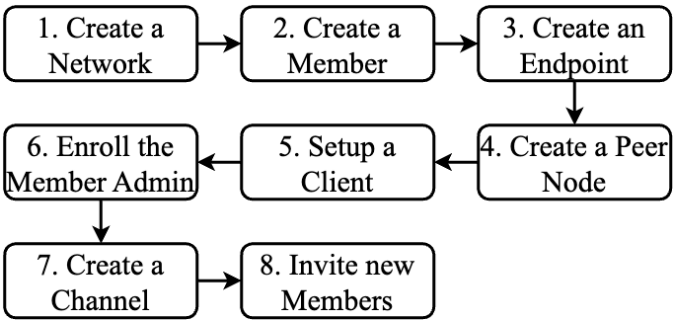


Fig. 4. Create blockchain network steps [4].

C. System Functionality

In the ensuing section, we demystify the findings and performance indicators derived from our experimental exertions. Our investigative ventures were coordinated employing a test dataset that is germane to patient healthcare records. Test automation and quality assurance play an important role in monitoring test results because they reduce human effort and cost and improve the accuracy of results [19]. For the enactment of these experiments, we utilized the open-source Gatling library, involving ourselves in executing concurrent operations which include, but are not limited to, the creation, retrieval, and updating of patient records, alongside obtaining historical data. The script was set into motion with an inaugural group of ten concurrent users and was systematically scaled to integrate up to 2,000 users over a span of 100 seconds. This method of scaling afforded us the ability to assess the system's performance while simultaneously accommodating up to 20 concurrent users for each of the four distinct operations.

The results indicated that the proposed solutions are effective and well-developed for addressing real-world needs. What distinguishes this platform is its effective utilization of Amazon web services. In brief, although the experimental outcomes emphasize the system's potential efficacy in situations with high demand, additional investigation is necessary to confirm its suitability in practical healthcare environments. This involves considering various technical, regulatory, and user experience factors that may impact its performance and acceptance.

The ensuing illustrations elucidate the results procured from the Gatling tool:

- Fig. 5 delineates the spectrum of Response times and the count of requests.
- Fig. 6 unveils the count of active users over the duration of the experiment, reaching an apogee of 1,706 active users at a particular juncture during the experimental phase.
- Fig. 7 sheds light on the dispersion of response times, illustrating the duration requisite to procure a response from the server.

Each diagram provides a visual representation that aids in comprehending the performance and response capabilities of the system during various levels of user load and interaction, offering valuable insights into its operational viability.

An essential consideration is the expense associated with infrastructure. Numerous current healthcare data management systems depend on expansive server farms and physical infrastructure, leading to substantial maintenance and upgrade costs. In comparison, our system utilizes the cost-effective 'Starter' edition of Amazon Managed Blockchain, lessening the financial strain on healthcare providers and enabling scalable expansion without requiring a substantial initial investment.

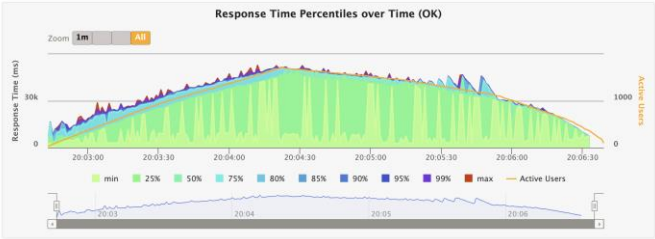


Fig. 5. Response time percentile over time for successful request.

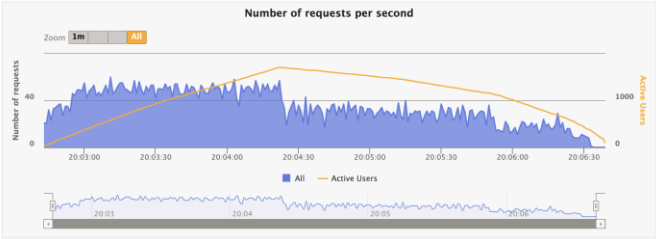


Fig. 6. Number of requests per second.

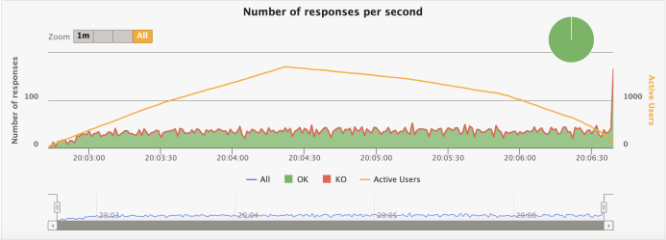


Fig. 7. Number of responses per second.

V. Conclusion and Future Work

This paper unveils a pioneering methodology toward devising a high-performance, cloud-centric solution for safeguarding and accessing medical records within a Hyperledger blockchain network. The implementation is proficient in adroitly managing the migration of historical data, including a magnitude of five million records, and can effectively administrate daily data influxes. The system's capability to accommodate concurrent users was tested utilizing the starter edition of Amazon Managed Blockchain machine type, thereby illustrating its scalability and cohesive integration with AWS services.

Moreover, the paper accentuates the crucial role of testing automation and quality assurance in authenticating experimental outcomes, diminishing human labor and financial expenditure while enhancing precision.

The suggested system is predominantly centered around data storage and analysis, laying a foundation for future enhancements and the integration of novel features to amplify functionality and security. This encompasses potential amalgamation with nascent technologies like artificial intelligence and machine learning to expedite diagnostic processes. Furthermore, enhancements such as multifaceted authentication to secure personal, confidential information and the option to scale to more substantial machine sizes for bolstered performance are contemplated. These prospective enhancements underscore the potential for ongoing system development to cater to the evolving demands of healthcare data management.

References

[1] M. K. Elghoul, S. F. Bahgat, A. S. Hussein, and S. H. Hamad, “Management of medical record data with multi-level security on Amazon Web Services,” SN Appl Sci, vol. 5, no. 11, p. 282, Nov. 2023, doi: 10.1007/s42452-023-05502-9.

[2] S. Elgayar, S. Hamad, and E.-S. El-Horbaty, “Revolutionizing Medical Imaging through Deep Learning Techniques:An Overview.,” International Journal of Intelligent Computing and Information Sciences, vol. 23, no. 3, pp. 59–72, Sep. 2023, doi: 10.21608/ijicis.2023.211266.1274.

[3] salma mostafa ahmed helmy, A. Mahmoud Amar, and E.-S. El-Horabty, “Internet of Things (IoT) based smart device for cardiac patients monitoring using Blynk App,” International Journal of Intelligent Computing and Information Sciences, vol. 0, no. 0, pp. 0–0, Dec. 2022, doi: 10.21608/ijicis.2022.139226.1182.

[4] M. K. Elghoul, S. F. Bahgat, A. S. Hussein, and S. H. Hamad, “Secured Cloud-based Framework for Electronic Medical Records using Hyperledger Blockchain Network,” Egyptian Computer Science Journal, vol. 46, no. 2, Sep. 2022.

[5] C. C. Agbo and Q. H. Mahmoud, “Comparison of blockchain frameworks for healthcare applications,” Internet Technology Letters, vol. 2, no. 5, Sep. 2019, doi: 10.1002/itl2.122.

[6] M. Elghoul, S. Bahgat, A. Hussein, and S. Hamad, “A Review of Leveraging Blockchain based Framework Landscape in Healthcare Systems,” International Journal of Intelligent Computing and Information Sciences, vol. 0, no. 0, pp. 1–13, Oct. 2021, doi: 10.21608/ijicis.2021.75531.1095.

[7] P. J. Taylor, T. Dargahi, A. Dehghantanha, R. M. Parizi, and K.-K. R. Choo, “A systematic literature review of blockchain cyber security,” Digital Communications and Networks, vol. 6, no. 2, pp. 147–156, May 2020, doi: 10.1016/j.dcan.2019.01.005.

[8] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, “An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends,” in 2017 IEEE International Congress on Big Data (BigData Congress), IEEE, Jun. 2017, pp. 557–564. doi: 10.1109/BigDataCongress.2017.85.

[9] M. Hölbl, M. Kompara, A. Kamišalić, and L. Nemec Zlatolas, “A Systematic Review of the Use of Blockchain in Healthcare,” Symmetry (Basel), vol. 10, no. 10, p. 470, Oct. 2018, doi: 10.3390/sym10100470.

[10] A. Hasselgren, K. Kralevska, D. Gligoroski, S. A. Pedersen, and A. Faxvaag, “Blockchain in healthcare and health sciences—A scoping review,” Int J Med Inform, vol. 134, p. 104040, Feb. 2020, doi: 10.1016/j.ijmedinf.2019.104040.

[11] V. Patel, “A framework for secure and decentralized sharing of medical imaging data via blockchain consensus,” Health Informatics J, vol. 25, no. 4, pp. 1398–1411, Dec. 2019, doi: 10.1177/1460458218769699.

[12] V. L. Lemieux, “Trusting records: is Blockchain technology the answer?,” Records Management Journal, vol. 26, no. 2, pp. 110–139, Jul. 2016, doi: 10.1108/RMJ-12-2015-0042.

[13] A. A. Siyal, A. Z. Junejo, M. Zawish, K. Ahmed, A. Khalil, and G. Soursou, “Applications of Blockchain Technology in Medicine and Healthcare: Challenges and Future Perspectives,” Cryptography, vol. 3, no. 1, p. 3, Jan. 2019, doi: 10.3390/cryptography3010003.

[14] A. Khatoon, “A Blockchain-Based Smart Contract System for Healthcare Management,” Electronics (Basel), vol. 9, no. 1, p. 94, Jan. 2020, doi: 10.3390/electronics9010094.

[15] D. Ichikawa, M. Kashiyama, and T. Ueno, “Tamper-Resistant Mobile Health Using Blockchain Technology,” JMIR Mhealth Uhealth, vol. 5, no. 7, p. e111, Jul. 2017, doi: 10.2196/mhealth.7938.

[16] S. Rouhani, L. Butterworth, A. D. Simmons, D. G. Humphery, and R. Deters, “MediChainTM: A Secure Decentralized Medical Data Asset Management System,” Jan. 2019, doi: 10.1109/Cybermatics_2018.2018.00258.

[17] P. Zhang, J. White, D. C. Schmidt, and G. Lenz, “Design of Blockchain-Based Apps Using Familiar Software Patterns with a Healthcare Focus,” in Proceedings of the 24th Conference on Pattern Languages of Programs, in PLoP ’17. USA: The Hillside Group, 2017.

[18] A. Panwar, V. Bhatnagar, M. Khari, A. W. Salehi, and G. Gupta, “A Blockchain Framework to Secure Personal Health Record (PHR) in IBM Cloud-Based Data Lake,” Comput Intell Neurosci, vol. 2022, pp. 1–19, Apr. 2022, doi: 10.1155/2022/3045107.

[19] A. Ali, H. Maghawry, and N. Badr, “Automation of Performance Testing: A Review,” International Journal of Intelligent Computing and Information Sciences, vol. 0, no. 0, pp. 1–16, Dec. 2022, doi: 10.21608/ijicis.2022.161846.1219.

Hyperchaotic Image Encryption System Based on Deep Learning LSTM

Shuangyuan Li[1], Mengfan Li[2], Qichang Li[3], Yanchang Lv[4]
Information Construction Office, Jilin Institute of Chemical Technology, Jilin, China[1]
School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin, China[2, 3, 4]

***Abstract*—This paper introduces an advanced method for enhancing the security of image transmission. It presents a novel color image encryption algorithm that combines hyperchaotic dynamics and deep learning medium and long short-term memory (LSTM) networks. Firstly, the chaotic sequence is generated using the Lorenz hyperchaotic system, then the Lorenz chaotic system is discretized and iteratively processed using the fourth-order Runge-Kutta (RK4) method, and then the deep learning LSTM model is used to transform the chaotic sequence processed by the Lorenz hyperchaotic system into a new sequence for training. Finally, according to the new chaotic signal, the Arnold disruption and Deoxyribo Nucleic Acid (DNA) encoding double disruption diffusion are performed to derive the ultimate encrypted image. Through the analysis of multiple color image simulation experiments, the algorithm presented in this paper can well realize the encryption on color images and can achieve lossless encryption, with strong resistance to differential attack, statistical attack and violent attack. Compared with the literature analysis, the correlation coefficient, information entropy and pixel change rate of this paper are closer to the ideal value, and it has higher security and better encryption effect.**

Keywords*—Image encryption; Lorenz Chaotic System; LSTM model; deep learning; DNA encoding

I. Introduction

With the advancement of information technology, big data, 5G and cloud computing technologies are inseparable from many areas of daily life, such as education, military, medicine and scientific conferences. However, while the Internet has brought great convenience, large amounts of data face numerous challenges, such as espionage, theft, usurpation, and modification, leading to numerous information security incidents. One such area is digital images, where the transmission of image data is an integral part of Artificial Intelligence (AI) systems. Through the use of image encryption technology, images can be encrypted so that they are not easily stolen or tampered with during transmission, thus ensuring the security of transmission. In the field of AI, personal image data, such as face recognition and human posture recognition, plays a significant role. Image encryption technology can be used to encrypt and protect these sensitive image data to ensure the security of personal privacy.

As an important carrier of information, digital image is a two-dimensional image composed of discrete pixels (picture element), each pixel represents a point in the image with specific position, brightness and color information. They can be generated by digital cameras, scanners or computers, and stored and processed in digital form. They contain important information in areas such as defense medicine, and education [1]. To guarantee the security of the transmission process, the most effective way is encryption. Traditional encryption methods include symmetric encryption, which is one of the earliest and simplest encryption methods that performs encryption and decryption operations on plaintext by using the same key, and some of the most commonly used symmetric encryption algorithms include the Advanced Encryption Standard (AES), Data Encryption Standard (DES), Triple Data Encryption Standard (3DES), and others [2]. Symmetric encryption offers the benefits of excellent efficiency and high speed, but it needs to secure the key transmission process. Asymmetric encryption, also known as public key encryption, involves a public key that can be used by anyone to encrypt data, while the private key is retained exclusively by the key holder for decrypting the data. Prominent asymmetric encryption methods encompass Rivest-Shamir-Adleman (RSA) and Elliptic Curve Cryptography (ECC), Digital Signature Algorithm (DSA) [3]. These algorithms are primarily developed for text-based application information and are less efficient in encrypting images with high image pixel correlation and extensive redundancy. It is difficult for traditional image encryption methods to overcome the difficulties in key distribution, resulting in illegal theft of ciphertext and low security. Deep learning algorithms have high complexity, good chaos characteristics and strong parameter sensitivity, and can be applied in the field of image encryption. Therefore, this paper, we investigate image encryption algorithms for chaotic systems with more complex performance and higher security.

II. Current Status of Research

With the dissemination of digital images in the network, the security of digital images has emerged as a significant concern [4]. To enhance data security, numerous image encryption algorithms have been introduced, including those based on chaotic systems [5], compression perception [6], DNA coding [7], S-box transform [8] and transform domain [9]. In the 1960s, Lorenz, a scientist from Massachusetts Institute of Technology (MIT) in the United States, proposed chaotic system for the first time, and attracted widespread attention and research, and scholars from various countries gradually utilized chaotic systems in picture encryption. In 1997, Fridrich pioneered the application of chaotic systems in image encryption for the inaugural time [10]. Usually, chaos-based image encryption algorithm mainly consists of disarray and diffusion, disarray is to diminish the correlation between adjacent pixels in a plaintext picture by changing the spatial

position between pixels, and diffusion refers to changing the pixel value of an image.

Chaotic systems are a class of dynamic nonlinear systems that exhibit non-stationarity, unpredictability and strong sensitivity. Due to their unpredictability and complexity they are widely used in information security and random number generation. Chaotic signal as the core of chaotic image encryption consists of the following two kinds of chaotic systems: one is a one-dimensional chaotic system, a nonlinear dynamical system containing only one independent variable, such as Logistic mapping, Henon mapping, Sine mapping, etc., which is simple in structure, high in operational efficiency, and has different dynamical characteristics, but the system key space is small and vulnerable to attacks, and can be determine the chaotic characteristics by analyzing the bifurcation map, periodicity and so on. Another kind of chaotic system is multi-dimensional chaotic system, which contains multiple variables and has more complex and diversified dynamics characteristics, such as Lorenz system, Chen system, etc. These systems introduce more state variables, have higher complexity and heightened sensitivity to starting conditions, making them prevalent in the realm of image encryption.

In order to enhance the security of image encryption, Alghamdi Yousef and Munir Arslan proposed an encryption algorithm with a nonlinear feedback shift register in their literature [11]. This algorithm improves security through multiple rounds of encryption, row substitution, column substitution, and bit-level substitution, resulting in higher quality and efficiency. The paper provides detailed insights into how the algorithm achieves these improvements. Chen Xin et al. proposed the algorithm of quantum chaos and DNA coding in literature [12], by studying the properties of IEA-QCDC, plaintext attack was used to obtain the key, and DNA coding operation was used to obtain the encrypted image, making full use of the security defects of IEA-QCDC and achieving better encryption effect. Ding Dawei, Wang Jin et al. proposed a fractional-order amnesia-coupled chaotic mapping (MCCM), using internal parameters, to make the system more stable and more conducive to the application in the field of chaotic engineering, through experimental analysis, the system exhibits a heightened level of dynamic complexity, leading to the development of a secure medical picture encryption scheme, which provides technical support for the security of the medical field, with good security and robustness [13]. Liang Qin et al. tackled the problem of insufficient security in one-dimensional chaotic systems. They introduced a novel one-dimensional chaotic system with the aim of bolstering security. They introduced a new one-dimensional chaotic system designed to enhance security. This new system, known as one-dimensional sine-cosine chaotic mapping (SCCM), leverages index mismatch of chaotic sequences in image rows and columns, and incorporates random DNA code selection to enhance the encryption process's resistance against plaintext attacks, resulting in improved encryption effectiveness [14]. Francesco Castro et al. introduced a secure fingerprint authentication image encryption scheme. The primary objective of this scheme is to bolster the security and resilience of safeguarding medical images. To address the issue of medical image insecurity during transmission and safeguard patient privacy, the scheme employs chaotic encryption with a replacement key for encrypting private images. Additionally, it implements a hybrid encryption approach based on the ECC and AES, resulting in improved security [15]. Meanwhile, as deep learning technology continuously advances, more algorithms that utilize the fusion of deep learning and chaotic systems are widely used. Ulises Manuel Ramirez Alcocer et.al introduced a deep learning approach that utilizes the LSTM network to monitor medical processes. This approach has been found to yield improved results in the medical field [16].

To summarize the state of multi-dimensional chaotic systems and the extensive application of deep learning, this paper introduces a novel approach that leverages LSTM networks to generate new chaotic signals. These signals are subsequently applied in chaotic systems and iterated using Lorenz mapping to create pseudo-random sequences. The algorithm further employs the Arnold algorithm and DNA coding for double disruption to enhance security. This leads to an expanded key space and increased resistance to attacks. Compared with the reference [17], the original equalization method using 3D histogram is discarded, and the LSTM model, which has better performance in long sequences, is used to enhance the sequence complexity. The traditional encryption method is only related to the image pixel arrangement, and the encryption is too single and simple, this paper increases the number of parameters through the addition of deep learning LSTM, so that the chaotic system reaches a super chaotic state, and at the same time, constructs the DNA encoding rules, which ultimately affects the factors to increase, and improves the security of the algorithm, as well as the encryption efficiency, and has a higher value of using it in the confidentiality.

III. BASIC THEORY

A. Lorenz Chaotic System

Lorenz chaotic system is a three-dimensional nonlinear system of ordinary differential equations, which is composed of three coupled nonlinear differential equations with good chaotic properties and initial value sensitivity, and is often widely used as a classical pseudorandom generator in the field of image encryption [18]. The Lorenz system is defined as depicted in Eq. (1).

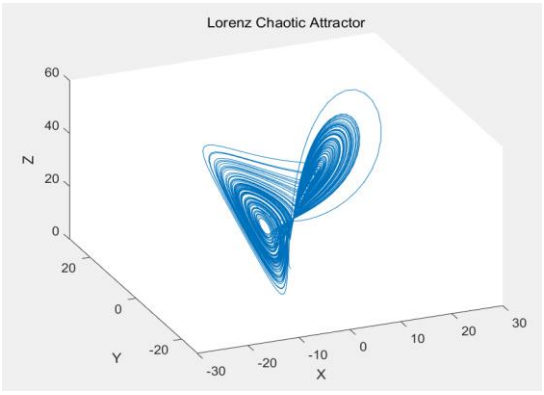


Fig. 1. Phase diagram of hyperchaotic lorenz attractor.

$$\begin{aligned} \dot{x} &= a(y-x)+w, \\ \dot{y} &= cx-y-xz, \\ \dot{z} &= xy-bz, \\ \dot{w} &= -yz+rw, \end{aligned} \tag{1}$$

where, $\dot{x}$, $\dot{y}$, $\dot{z}$, $\dot{w}$ are the chaotic states of the system, for the control parameters a, b, c and r of the Lorenz system, when these parameters satisfy a=10, b=8/3, c=28, and $-1.52 \leq r \leq -0.06$, the system enters a state of hyperchaotic. The attractor phase diagram of the hyperchaotic Lorenz system can be observed in Fig. 1. When the system control parameter r=-1, the Lyapunov exponents in Eq. (1) are as follows in order: 0.3381, 0.1586, 0, and -15.1752. Notably, this sequence contains two Lyapunov exponents greater than zero. The chart of the Lyapunov exponent diagram is shown in Fig. 2.

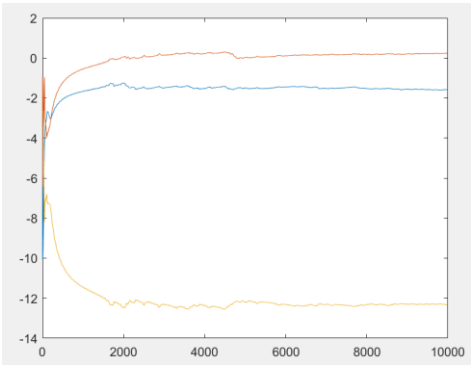


Fig. 2. Lyapunov exponent plot.

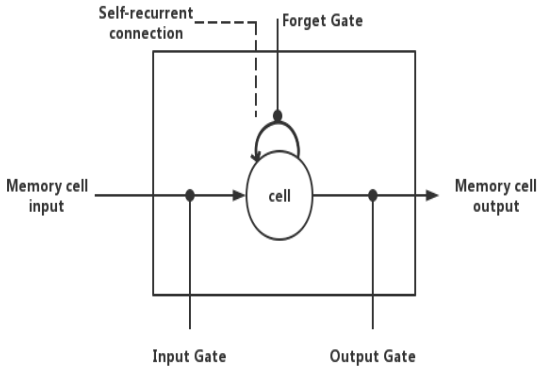


Fig. 3. LSTM model diagram.

B. LSTM Model

LSTM can be viewed as a specialized variant of Recurrent Neural Network (RNN), primarily developed to address the challenges related to gradient vanishing and gradient explosion when training with lengthy sequences [19]. LSTM's unique architecture and memory retention mechanisms make it particularly well-suited for a wide range of applications. RNN is a kind of neural network dealing with sequential data. Based on the RNN model, the LSTM model introduces a gating mechanism (Gates) to solve the short-term memory problem of RNN, which is able to capture long-term dependencies, thus enabling the recurrent neural network to better leverage long-range temporal information and have better performance in longer sequences. The LSTM model includes three key logical control units: the Input Gate, Output Gate, and Forgetting Gate. These units are connected to a multiplication element, forming the core structure of the LSTM model, as illustrated in Fig. 3. These components play a crucial role in controlling information flow and memory retention within the network, allowing it to excel in tasks involving longer sequences.

The cell is a memory unit that represents the memory of the neuron's state, providing the LSTM unit with the capability to store, retrieve, reset, and update long-term information. When the moment t, the LSTM neural network defines the formula as:

$$\begin{aligned} f_t &= \text{sigmid}\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \\ i_t &= \text{sigmid}(W_i \cdot [h_{t-1}, x_t] + b_i) \\ o_t &= \text{sigmid}(W_o \cdot [h_{t-1}, x_t] + b_o) \\ \tilde{c}_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\ c_t &= f_t \times c_{t-1} + i_t * \tilde{c}_t \\ h_t &= o_t \times \tanh(c_t) \end{aligned} \tag{2}$$

where, f_t、i_t、o_t and c_t denote Forget Gate, Input Gate, Output Gate, and cell, respectively, as mentioned in Fig. 3. W_*, on the other hand, denotes the recursive connection weights of the corresponding gates, respectively, and sigmid and tanh denote the two activation functions.

C. Arnold Mapping

Arnold mapping is a chaotic mapping technique that repeatedly applies folding and stretching transformations within a bounded region, serving as a primary method for permutation. The transformation principle involves performing a shearing transformation along the x-axis, followed by a similar transformation along the y-axis. Finally, a modulo operation is used to implement cut-and-fill operations, altering the layout of image grayscale values by changing the pixel coordinates. The specific transformation formula is:

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_n \\ y_n \end{pmatrix} mod \begin{pmatrix} M \\ N \end{pmatrix} \tag{3}$$

D. DNA Coding

In the field of biology, medicine and other applications, the information of many organisms in nature exists in DNA, and DNA can be arranged according to different rules and sequences to achieve the effect of disarray through the disarrangement of rows and columns. DNA consists of four kinds of DNA, which are adenine (A), thymine (T), guanine (G) and cytosine (C). In accordance with the rules of complementary base pairing in biology, A is paired with T, while G is paired with C. In computer systems, the complementarity of the two bases is determined by the fact that they are complementary to each other. In a computer system, the complementary pairing rules are similar to the pairing rules for binary coding, 0 and 1. Therefore, the four varieties of deoxyribonucleotides can be denoted using two binary digits, yielding a sum of 24 feasible codes. However, only eight coding rules, in accordance with DNA coding standards, are retained, as shown in Table I [20]. Every pixel in a grayscale image can be denoted using 8-bit binary numbers consisting of 0 and 1. Similarly, each pixel value can be encoded using a DNA sequence comprising four nucleotides. As an illustration,

the decimal number 232 can be expressed in binary as 11101000, and following the DNA coding rules, this binary sequence can be converted to the DNA sequence TGGA.

In the process of image encryption, DNA arithmetic rules are employed, encompassing DNA addition, DNA subtraction, and DNA dissimilarity operations. Adding and subtracting DNA sequences follow similar principles to traditional algebraic calculations [21], and the rules for these arithmetic operations are provided in Table II and Table III.

TABLE I. DNA CODING RULES

Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Rule 7	Rule 8
A-00	A-00	C-00	G-00	C-00	G-00	T-00	T-00
T-11	T-11	G-11	C-11	G-11	C-11	A-11	A-11
C-01	G-01	A-01	A-01	T-01	T-01	C-01	G-01
G-10	C-10	T-10	T-10	A-10	A-10	G-10	C-10

TABLE II. DNA ADDITION RULES

+	A	T	C	G
A	A	T	C	G
G	G	C	T	A
T	T	A	G	C
C	C	G	A	T

TABLE III. DNA SUBTRACTION RULES

-	A	T	C	G
A	A	T	C	G
G	G	G	C	A
T	T	T	A	C
C	C	C	G	T

IV. IMAGE ENCRYPTION

A. Encryption Algorithm

The image encryption algorithm presented in the paper follows the traditional rules of disarray and diffusion, and proposes a double disarray by combining Arnold mapping and DNA encoding. During the encryption process, the initial key of the original image P is first obtained using a hash function. Following this, a chaotic sequence is produced using the hyperchaotic Lorenz system, and the system is subjected to discretization and iterative processing using the Runge-Kutta method. The deep learning LSTM model is utilized to leverage temporal features for both training and analyzing the chaotic sequence, ultimately generating a novel chaotic sequence. Finally, double permutation is applied to shuffle and diffuse the pixels, resulting in the ultimate encrypted image. The primary encryption procedure of the algorithm are illustrated in Fig. 4.

B. Encryption Process

1) Key generator: The key for the plaintext image P is determined through the SHA-256 function, which serves as the fundamental component of the key generator. First take the plaintext image P as input, use the function to generate a 64-bit digest, convert it to 256-bit binary as output, and divide it into groups of every 8 bits to get a 32-bit segmented keystream $K = \{k_1, k_2, k_3, \cdots k_{32}\}$, where k_i, $i = 1,2,\cdots 32$ are all 8-bit hash values.

According to the generation rule of key flow, K is processed as follows, according to Eq. (4) to Eq. (7), the initial state parameters of the system $\{x_0, y_0, z_0, w_0\}$ can be computed, and through Eq. (8), we get the key s_0 which is related to the plaintext, and then we generate pseudo-random sequences S_1 , S_2 used for the subsequent Arnold diffusion.

$$x_0 = \left((k_{17} \oplus k_{18}) \oplus (k_{18} \oplus k_{19}) \oplus (k_{19} \oplus k_{20})\right) \times \sum_{j=1}^{j=32} \frac{k_j \times 2^{j-1}}{2^{47}} \quad (4)$$

$$y_0 = \left((k_{21} \oplus k_{22}) \oplus (k_{22} \oplus k_{23}) \oplus (k_{23} \oplus k_{24})\right) \times \sum_{j=1}^{j=32} \frac{k_j \times 2^{j-1}}{2^{47}} \quad (5)$$

$$z_0 = \left((k_{25} \oplus k_{26}) \oplus (k_{26} \oplus k_{27}) \oplus (k_{27} \oplus k_{28})\right) \times \sum_{j=1}^{j=32} \frac{k_j \times 2^{j-1}}{2^{47}} \quad (6)$$

$$w_0 = ((k_{29} \oplus k_{30}) \oplus (k_{30} \oplus k_{31}) \oplus (k_{31} \oplus k_{32})) \times \sum_{j=1}^{j=32} \frac{k_j \times 2^{j-1}}{2^{47}} \quad (7)$$

$$s_0 = \text{zeros}(1, n), n = 2MN \quad (8)$$

2) Preprocessing: The RK4 is a suitable method for solving differential equations. It achieves higher computational accuracy by approximating the differential equations in four discrete steps over a specific time interval of the solution. The discretized equations are as follows:

$$\begin{aligned} x_{n+1} &= x_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ k_1 &= f(t_n, x_n) \\ k_2 &= f\left(t_n + \frac{h}{2}, x_n + \frac{h}{2}k_1\right) \\ k_3 &= f\left(t_n + \frac{h}{2}, x_n + \frac{h}{2}k_2\right) \\ k_4 &= f(t_n + h, x_n + hk_3) \end{aligned} \quad (9)$$

In this method, the subsequent value x_{n+1} is calculated by multiplying the current value x_n with the period h and reckon slope. The slopes are represented as follows: k_1 is the slope at the start of the period, k_2 and k_3 correspond to the slopes in the middle of time, and k_4 indicates the slope at the conclusion of the time period. The Lorenz chaotic system is discretized using RK4 and iterated T+MN times, discarding the first T times to eliminate transient effects and increase safety, four new sequences x_1, y_1, z_1, w_1 are obtained.

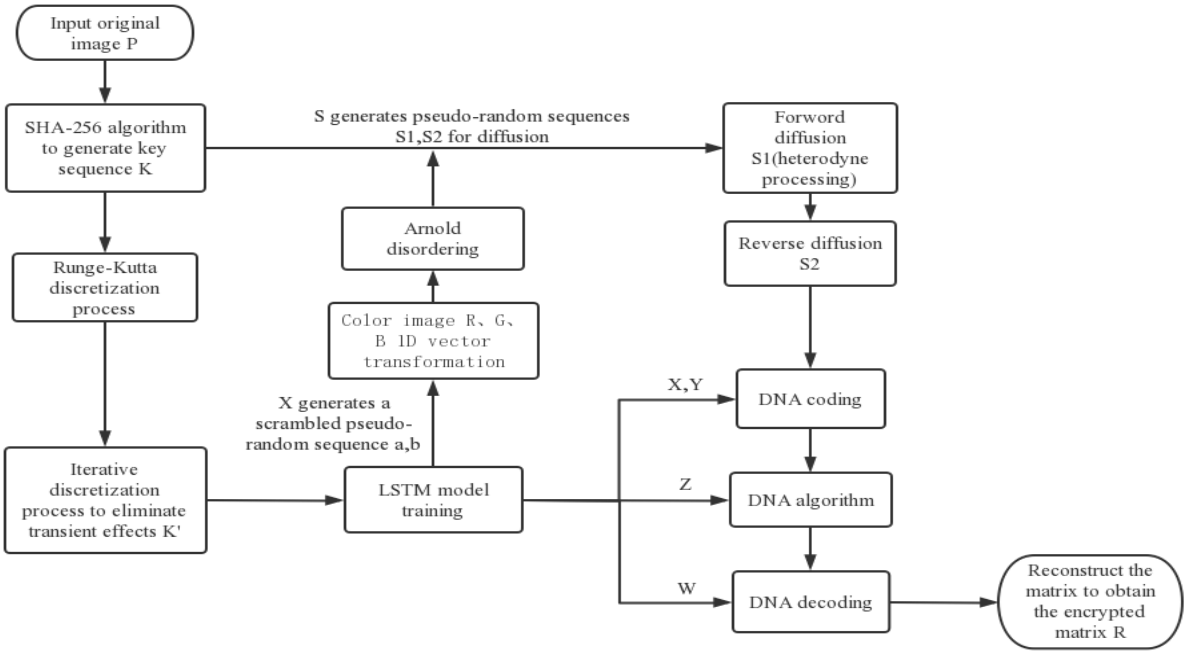


Fig. 4. Encryption flow chart.

3) LSTM training: By configuring the parameters of LSTM, the preprocessed sequences are trained, and a portion of the length l of the pseudo-random sequences x_1, y_1, z_1, w_1 are respectively selected for deep learning. The initial learning rate is represented as 'r', the learning rate decreasing factor is 'p', and the length of the selected training sequences is 'l'. Following the completion of the training process, four novel pseudo-random sequences, namely X, Y, Z, and W, are generated. During the training process, the root-mean-square error (RMSE) serves as a more effective metric for quantifying the disparity between predicted and actual values. It offers heightened sensitivity to anomalous data and is defined as the square root of the average of the squared differences between the predicted and actual values, divided by the number of observations, denoted as n. In Fig. 5, the RMSE is depicted as it evolves with the number of iterations during the training process. As the number of iterations increases, the RMSE becomes smaller and gradually approaches 0, signifying that the model can make more accurate data predictions.

4) Double disorder modeling: The color original image P, with dimensions M×N, is separated into three individual color channels: R, G, and B. The pseudo-random sequences a, b are generated using LSTM generated sequences X. The decomposed channels are initially converted into one-dimensional vectors, and then transformed into the coordinates using a, b to get the transformed coordinates q, as shown in Eq. (10), and then Arnold disambiguation is carried out.

$$\begin{aligned} a &= X(1:M*N) \\ b &= X(M*N+1:2*M*N) \\ q &= mod(b(i)+a(i)*i, M*N)+1 \end{aligned} \tag{10}$$

After disorganization, the key S_1, S_2 obtained in the key generation stage is used for forward and reverse diffusion according to Eq. (11) and Eq. (12) to obtain the new single-channel image. The key S_1 is used for forward diffusion, which is a kind of diffusion processing method based on different-or operation, and S_2 is the key for reverse diffusion, where B_i represents the three channels R, G, B, and B_i' represents the new R, G, B channels.

$$\begin{aligned} B_i &= (B_{i-1}+S_1+P_i) mod 256 \\ C_i &= (C_{i+1}+S_2+B_i) mod 256 \end{aligned} \tag{11}$$

$$B_i' = reshape(C_i, M, N) \tag{12}$$

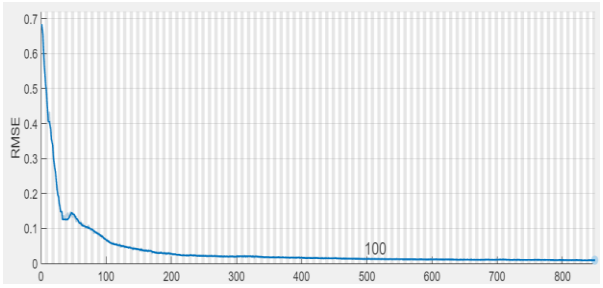


Fig. 5. RMSE variation curve.

Through DNA encoding, the bases A, T, G, C and the binary pixel values are disambiguated according to certain rules, which consumes less time and has a strong disambiguation effect and is highly applicable in image data encryption [22]. Each chunk of the graphic representation P is encoded according to the rules in Table I, and the encoding rules are determined by the pseudo-random sequence X to get the matrix P_i'. The obtained new three single channels are encoded according to Y to get the disarrayed matrix Q_i'. The encoded chunks are subjected to the DNA operation of P_i' and

Q_i' in accordance with the rules established by the sequence Z, and the encrypted matrix Q_i is obtained. Finally, DNA decoding of Q_i is performed using W. Each chunk are combined to create the ultimate encrypted image Q .

C. Decryption Process

The decryption process is the inverse application of the encryption. The individual channels of the encrypted picture are isolated, and the separated image is subjected to DNA coding inverse disambiguation, plus mode inversion, forward inverse diffusion. The Arnold inverse transformation is employed to extract the individual channels from the plaintext image, and subsequently, these channels are reconstructed and combined to obtain the original plaintext P.

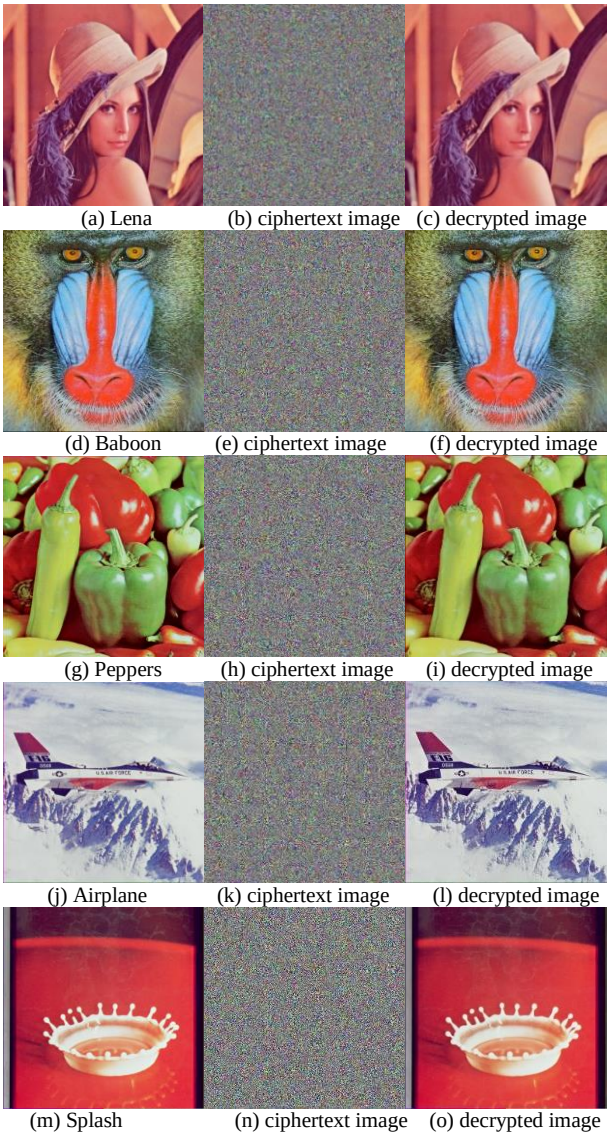
V. Encryption Effect Analysis

In this paper, we employ the MATLAB R2022b simulation platform to carry out encryption and decryption operations using the proposed algorithm, and we analyze its performance. The hardware specifications during the experiments include an Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz, 1.80 GH. The experiments were conducted using 512×512 pixel color images of Lena, Baboon, Peppers, Airplane, and Splash color images for simulation testing. And the tests are performed in terms of histogram analysis, adjacent pixel correlation analysis, information entropy, differential attack analysis, robustness analysis and key space analysis. Fig. 6 shows the encryption and decryption effect of the color image, the encrypted image is entirely devoid of any visible information from the original image, making it impossible for an attacker to extract any meaningful data from the image, enough to resist the common means of attack such as differential attack, violence attack, and the algorithm yields a strong encryption effect.

A. Histogram Analysis

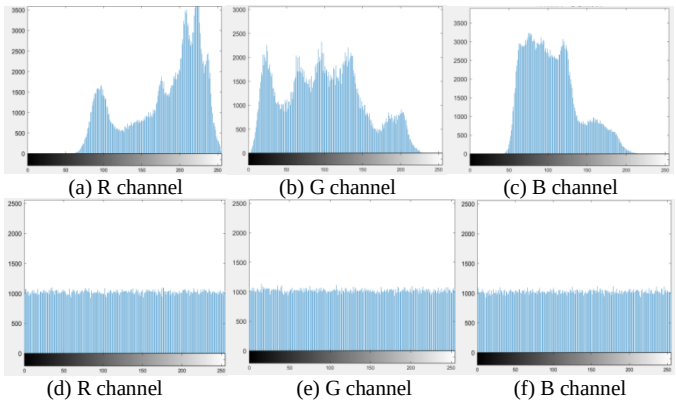
Histograms can display image information and offering a visual representation of the arrangement of individual grayscale values in picture. The frequency of occurrence is counted according to the size of the gray values. A better encryption algorithm should ideally render the histogram of the encrypted image so indistinct that a clear distinction cannot be discerned, the histogram can be evenly distributed. Histograms of plaintext images often exhibit clear statistical patterns, and attacks that exploit these patterns are known as statistical attacks [23]. To enhance the ability to resist statistical attacks, the histogram of the encrypted image should tend to a straight line, and there is a big difference between it and the plaintext histogram. For example, Fig. 7 shows the comparison of histograms before and after Lena's encryption.

From the above figure, it is evident that the pixels between the original plaintext images have strong statistical regularity and are more susceptible to statistical attacks, the distribution of elements in the encrypted Lena image exhibits greater uniformity, the encrypted histogram is smoother, and making it difficult to extract information from the encrypted image, thus enhancing security and increasing resistance to statistical attacks by potential attackers.



(a) Lena (b) ciphertext image (c) decrypted image
(d) Baboon (e) ciphertext image (f) decrypted image
(g) Peppers (h) ciphertext image (i) decrypted image
(j) Airplane (k) ciphertext image (l) decrypted image
(m) Splash (n) ciphertext image (o) decrypted image

Fig. 6. Encryption and decryption effect diagram.



(a) R channel (b) G channel (c) B channel
(d) R channel (e) G channel (f) B channel

Fig. 7. Histograms of lena images before and after R, G and B channel encryption: (a)-(c) Plaintext; (d)-(f) Ciphertext.

B. Correlation Analysis

During the image encryption process, the characteristics of image pixels make the correlation between neighboring pixels strong, and at the same time, it can disrupt the correlation between neighboring elements by disambiguating the pixels of the original picture, so as to achieve the purpose of resisting statistical attacks [24]. In a high-quality encryption algorithm, the lower the correlation between adjacent elements post-encryption, the more favorable the outcome. To evaluate an encryption algorithm's ability to withstand statistical attacks more effectively, we randomly selected 5000 pairs of pixel values from the Lena image in various orientations. And then examined the correlation coefficients in the three directions both before and after encryption, separately for the three single channels of the Lena image. The outcomes of this analysis are presented in Fig. 8, 9, and 10. This can be clearly seen from these pictures the strong correlations present in the plaintext image across all directions before encryption. However, post-encryption, the image exhibits a more uniform pixel distribution and scattered throughout the rectangular square matrix, there is no difference in the whole plane leading to a reduction in pixel-to-pixel correlations. The formula for calculating the correlation between two adjacent pixels can be expressed as follows:

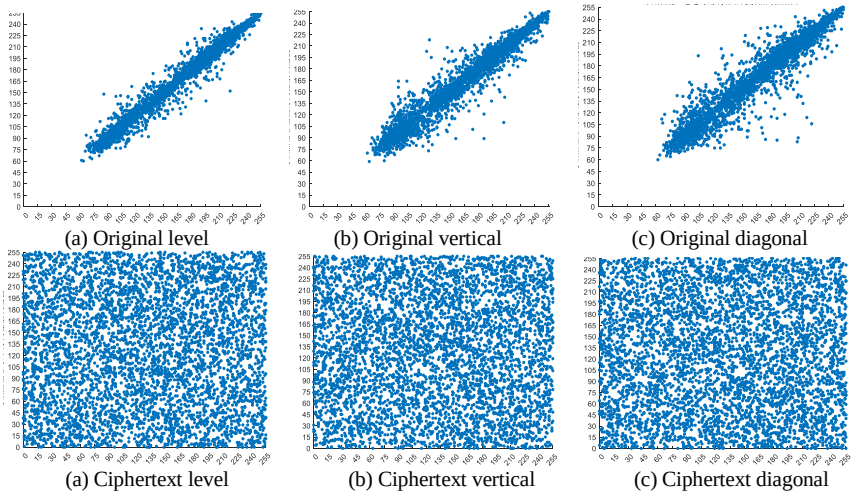
$$E(x) = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{13}$$

$$D(x) = \frac{1}{N}\sum_{i=1}^{N} \left(x_i - E(x)\right)^2 \tag{14}$$

$$\mathrm{cov}(x,y) = \frac{1}{N}\left(x_i - E(x)\right)\left(y_i - E(y)\right) \tag{15}$$

$$r_{xy} = \frac{\mathrm{cov}(x,y)}{\sqrt{D(x)}\sqrt{D(y)}} \tag{16}$$

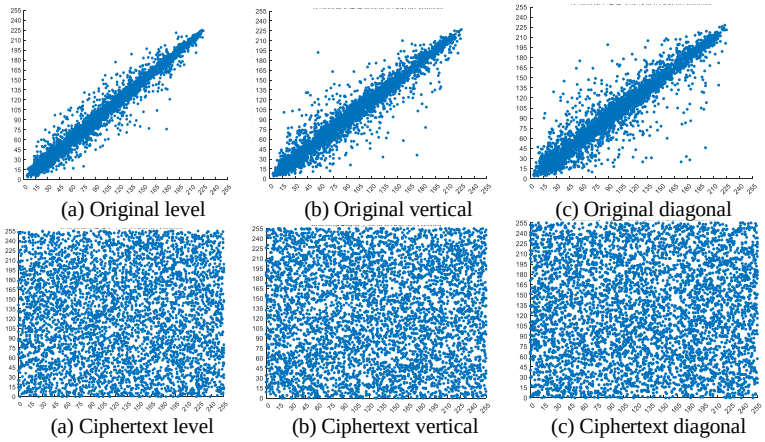
where $E(x)$ and $E(y)$ denote the mathematical expectation of x, y respectively, and $\mathrm{cov}(x,y)$ denotes the covariance of x, y . Through simulation test, compare and analyze the correlation coefficients before and after encryption of Lena and Baboon. Furthermore, a comparison was made with the correlations mentioned in the literature [25] [26], as shown in Table IV. The analysis reveals that the correlation of different channels (R, G, and B) in different directions before encryption is close to 1, indicating a higher correlation. In contrast, the correlation in the post-encryption image approaches the ideal value of 0. This algorithm offers a high level of security and exhibits enhanced resistance to statistical attacks.



(a) Original level (b) Original vertical (c) Original diagonal

(a) Ciphertext level (b) Ciphertext vertical (c) Ciphertext diagonal

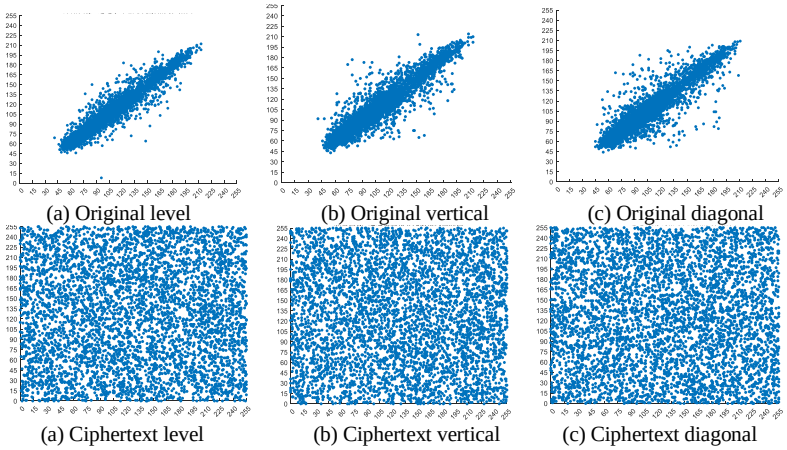
Fig. 8. Neighboring pixel correlation distribution before and after R-channel encryption of lena image.



(a) Original level (b) Original vertical (c) Original diagonal

(a) Ciphertext level (b) Ciphertext vertical (c) Ciphertext diagonal

Fig. 9. Distribution of neighboring pixel correlation before and after G-channel encryption of lena image.



(a) Original level (b) Original vertical (c) Original diagonal

(a) Ciphertext level (b) Ciphertext vertical (c) Ciphertext diagonal

Fig. 10. Distribution of correlation of neighboring pixels before and after encryption of B-channel of lena image.

TABLE IV. CORRELATION COEFFICIENTS OF DIFFERENT IMAGES IN THE THREE DIRECTIONS

Imagery	Directional	Lena			Baboon			Literature [25]	Literature [26]
		R	G	B	R	G	B		
Plaintext image	level	0.9805	0.9713	0.9387	0.9257	0.8673	0.9120	0.9726	0.9563
	vertical	0.9899	0.9836	0.9609	0.8664	0.7666	0.8852	0.9507	0.9242
	diagonal	0.9710	0.9565	0.9199	0.8531	0.7427	0.8510	0.9346	0.9015
Ciphertext image	level	-0.0036	0.0099	-0.0048	0.0022	-0.0488	0.0001	0.0042	0.0053
	vertical	0.0017	0.0182	0.0056	0.0024	-0.0146	-0.0046	0.0027	0.0059
	diagonal	-0.0036	0.0125	-0.0193	-0.0016	-0.0069	-0.0091	0.0070	0.0031

TABLE V. INFORMATION ENTROPY

Information entropy	R-channel		G-channel		B-channel	
Imagery	pre-encryption	post-encryption	pre-encryption	post-encryption	pre-encryption	post-encryption
Lena	7.2531	7.9993	7.5940	7.9993	6.9684	7.9993
Baboon	7.7067	7.9993	7.4744	7.9992	7.7522	7.9992
Peppers	7.3388	7.9993	7.4963	7.9993	7.0583	7.9994
Airplane	6.7178	7.9993	6.7990	7.9993	6.2138	7.9993
Splash	6.9481	7.9993	6.8845	7.9992	6.1265	7.9993

C. Information Entropy

Entropy is employed to characterize the intricacy of phenomena, and information entropy serves as a quantitative gauge of the level of randomness within a source, that is, it describes the complexity of a system. This metric can be applied to assess the randomness of an image by quantifying the dispersion of pixels with distinct grayscale values across different color channels. A distribution with higher uniformity indicates higher resistance to statistical attacks. Greater information entropy signifies increased complexity within the picture. An ideal encryption should possess an information entropy of 8. The calculation formula is:

$$H(x) = -\sum_{i=1}^{2N-1} P_i \log_2 P_i \quad (17)$$

In Eq. (17), P_i represents the occurrence rate of message i and $H(x)$ is the information entropy. Table V reveals that the information entropy of different channels in the encrypted image closely approaches the target value of 8. This suggests that the encrypted image significantly differs from the plaintext image and is less susceptible to leakage, thus indicating the algorithm in this paper has a higher level of security.

D. Differential Attack

A differential attack is one type of chosen plaintext attack, for a common method of cracking the ciphertext image, the attacker is able to the relationship between the picture before and after the encryption, to find out the law of the ciphertext image to be deciphered. The anti-differential performance mainly depends on the degree of sensitivity in the plaintext. In the encryption process, the Normalized Pixel Contrast Ratio (NPCR) and Unified Average Changing Intensity (UACI) are two variables to measure the two variables that quantify the disparity between two images and are quantitative tests to analyze the original and encrypted images [27]. NPCR measures the proportion of pixels that differ in the same positions between two images, relative to the total number of pixels in those images. UACI calculates the average magnitude

of changes between the images before and after transformations, and its formula is as follows:

$$NPCR = \frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} D(i,j)}{M \times N} \times 100\% \quad (18)$$

$$D(i,j) = f(x) = \begin{cases} 1, C_1(i,j) \neq C_2(i,j) \\ 0, otherwise \end{cases} \quad (19)$$

$$UACI = \frac{1}{M \times N}\frac{\Sigma(C_1(i,j) - C_2(i,j))}{255} \times 100\% \quad (20)$$

Among them, the ideal NPCR and UACI values are 99.6094% and 33.4635%, respectively. The algorithm's effectiveness in countering differential attacks improves as the measured values approach these ideal benchmarks. In this paper, a comparative analysis is conducted using the literature [28] [29], as illustrated in Table VI. The analysis demonstrates that the NPCR and UACI values in this algorithm closely approach the true values, indicating strong encryption efficacy and enhanced resistance to differential attacks.

TABLE VI. KEY SENSITIVITY

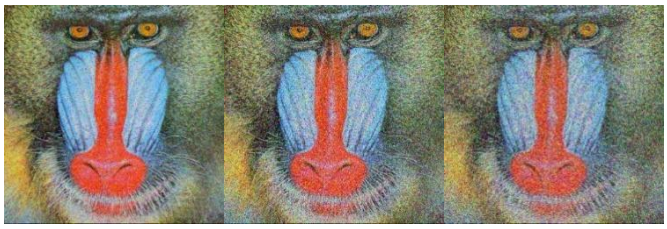
Imagery	NPCR/%	UACI/%
Lena	99.60	33.47
Baboon	99.62	33.45
Peppers	99.62	33.46
Airplane	99.61	33.47
Splash	99.61	33.47
Literature [28]	99.66	33.61
Literature [29]	99.61	33.48

E. Key Space

The key space represents the collection of all potential keys capable of generating an encryption key. The size of the key space is contingent upon the length of the security key, and a sufficiently extensive key space is effective in thwarting brute force attacks. When the key space of the whole algorithm reaches 2^{100}, it can show that the security of the algorithm can be guaranteed [30]. In the chaotic system used in this paper, the initial key is generated by the SHA-256 algorithm. Without considering the influence of other factors, the size of the security key is 256 bits, and the size of the key space is $2^{256} > 2^{100}$, which can withstand brute force attacks and exhibits a high level of security.

F. Robustness Analysis

Images are susceptible to interference from external factors during transmission, and the robustness of an encryption system is to assess how an image performs when subjected to various forms of interference. It is used to judge the encryption and decryption quality of an encryption system, with the aim of determining whether the system can effectively protect the encryption and integrity of the image when confronted with noise interference [31]. Salt and Pepper Noise (SPN) is a familiar type of noise in digital images that can significantly affect image quality, causing it to become unclear or distorted. The image decrypted with various levels of SPN added to the Baboon image is shown in Fig. 11. It is evident from the image that as the intensity of the added SPN increases, from (a) with intensity 0.1 to (c) with intensity 0.3, the quality of the deciphered image deteriorates and becomes increasingly blurry. However, it is still possible to discern the information in the original picture. This suggests that the algorithm introduced in this paper showcases substantial resistance to noise interference.



(a) (b) (c)

Fig. 11. Decrypted baboon images subjected to varying levels of SPN: (a) 0.1 (b) 0.2 (c) 0.3.

VI. CONCLUSION

This paper introduces a hyperchaotic image encryption algorithm based on LSTM. It involves processing the chaotic sequence generated by the Lorenz system in order to create new chaotic signals. This is achieved by leveraging the sequence data processing capabilities of the LSTM. Following this, the data undergoes iterative processing using the hyperchaotic system. Simultaneously, the Arnold algorithm, along with DNA coding and arithmetic rules, is applied to perform double diffusion of the data. This enhances the algorithm's complexity. In this paper, the above problems are described in detail and the experiment and analysis are carried out. The experimental results show that the algorithm can effectively ensure the privacy of the image through the encryption and decryption of the color RGB image. Meanwhile, comparisons with other algorithms, and an analysis of the algorithm's performance in terms of histograms, correlations, information entropy, and more, it has been verified that the encryption system exhibits strong resistance to differential attacks, statistical attacks, and brute-force attacks. It efficiently diminishes the correlation among adjacent pixels, thus bolstering the algorithm's complexity, security, and encryption efficacy. The DNA encoding technology employed in the algorithm falls within the realm of bioengineering, which suggests that the algorithm could be combined with biomedical treatments in the future to provide better protection for the transmission of medical data. In fact, there are many applications of deep learning algorithms to images, which can achieve good encryption effects and improve efficiency, which is the direction of future research and improvement.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for taking time to guide the completion of this paper.

REFERENCES

[1] Hasan A A ,Ali A M K ,Talib A A .Image encryption based on 2DNA encoding and chaotic 2D logistic map[J].Journal of Engineering and Applied Science,2023,70(1).

[2] Yongsheng H ,Han W ,Luoyu Z .Color image encryption base on a 2D hyperchaotic enhanced Henon map and cross diffusion[J].Alexandria Engineering Journal,2023,73.

[3] Wuyan L, Limin Z, Zhongbao Y, et al. Image encryption algorithm based on hyperchaotic system and dynamic DNA encoding[J]. Physica Scripta,2023,98(11).

[4] Wu J, Chen D, Liu H. Computer Network Security in the Era of "Internet +"[J]. Journal of Artificial Intelligence Practice,2022,5(3).

[5] Wang Y, Wu C, Kang S, et al. Multi-channel chaotic encryption algorithm for color image based on DNA coding[J]. Multimedia Tools and Applications,2020,79(prepublish).

[6] Wen W, Hong Y, Fang Y, et al. A visually secure image encryption scheme based on semi-tensor product compressed sensing [J]. Signal Processing, 2020, 173(4): 107580-107597.

[7] Wang X, Wang Y, Zhu X, et al. A novel chaotic algorithm for image encryption utilizing one-time pad based on pixel level and DNA level [J]. Optics and Lasers in Engineering, 2020, 125(8):105851-105863.

[8] Naseer Y, Shah T. Advance image encryption technique utilizing compression, dynamical system and s-boxes [J]. Mathematics and Computers in Simulation (MATCOM), 2020, 178(6): 207-217.

[9] Wu X, Kan H, Kurths J. A new color image encryption scheme based on DNA sequences and multiple improved 1D chaotic maps [J]. Application Software Computation, 2015, 37(6): 24-39.

[10] FRIDRICH J. Symmetric ciphers based on two-dimensional chaotic maps [J].International Journal of Bifurcation and Chaos, 1998, 8:1259-1284.

[11] Yousef A, Arslan M. An Image Encryption Algorithm Based on Trivium Cipher and Random Substitution[J]. SN Computer Science,2023,4(6).

[12] Xin C ,Simin Y ,Qianxue W , et al.On the cryptanalysis of an image encryption algorithm with quantum chaotic map and DNA coding[J].Multimedia Tools and Applications,2023,82(27).

[13] Dawei D, Jin W, Mouyuan W, et al. Controllable multistability of fractional-order memristive coupled chaotic map and its application in medical image encryption[J]. The European Physical Journal Plus,2023,138(10).

[14] Liang Qin,Zhu Congxu. A new one-dimensional chaotic map for image encryption scheme based on random DNA coding[J]. Optics and Laser Technology,2023,160.

[15] Castro F, Impedovo D, Pirlo G. A Medical Image Encryption Scheme for Secure Fingerprint-Based Authenticated Transmission[J]. Applied Sciences,2023,13(10).

[16] Alcocer R M U, Leal T E, Romero G, et al. A Deep Learning Approach for Predictive Healthcare Process Monitoring[J]. Information,2023,14(9).

[17] Malik S, Shah T. Color multiple image encryption scheme based on 3d-chaotic maps [J]. Mathematics and Computers in Simulation (MATCOM), 2020, 178(5): 646-666.

[18] Li T, Yan W, Chi Z. A new image encryption algorithm based on optimized Lorenz chaotic system[J]. Concurrency and Computation Practice and Experience,2020,34(13).

[19] Xinhe W. Water quality prediction based on AR and LSTM model[J]. Journal of Physics: Conference Series,2023,2580(1).

[20] WANG X, ZHAO M.An image encryption algorithm based on hyperchaotic system and DNA coding[J].Optics&Laser Technology,2021,143(14):107316.

[21] Thorat O, Mangrulkar R. Combining DNA sequences and chaotic maps to improve robustness of RGB image encryption[J]. International Journal of Computational Science and Engineering,2023,26(2).

[22] Ur M R, Arslan S, Bello A U. Securing Medical Information Transmission Between IoT Devices: An Innovative Hybrid Encryption Scheme Based on Quantum Walk, DNA Encoding, and Chaos[J]. Internet of Things,2023,24.

[23] Jian Z, Jifeng G, Donglei L. An efficient image encryption algorithm based on S-box and DNA code[J]. Measurement: Sensors,2023,29.

[24] Heping W, Yiting L. Cryptanalyzing an image cipher using multiple chaos and DNA operations[J]. Journal of King Saud University - Computer and Information Sciences,2023,35(7).

[25] Zhiqiang C, Wencheng W, Yuezhang D, et al. Novel One-Dimensional Chaotic System and Its Application in Image Encryption[J]. Complexity,2022,2022.

[26] Qin L, Congxu Z. A new one-dimensional chaotic map for image encryption scheme based on random DNA coding[J]. Optics and Laser Technology,2023,160.

[27] CHEN G R, MAO Y B, CHUI C K. A symmetric image encryption scheme based on 3D chaotic cat maps[J]. Chaos, Solitons and Fractals, 2004, 21(3): 749-761.

[28] Kari P A, Navin H A, Bidgoli M A, et al. A new image encryption scheme based on hybrid chaotic maps[J]. Multimedia Tools and Applications,2020,80(2).

[29] Chai, X., Fu, J., Zhang, J. et al. Exploiting preprocessing-permutation–diffusion strategy for secure image cipher based on 3D Latin cube and memristive hyperchaotic system. Neural Comput & Applic 33, 10371–10402 (2021).

[30] Geng S, Li J, Zhang X, et al. An Image Encryption Algorithm Based on Improved Hilbert Curve Scrambling and Dynamic DNA Coding[J]. Entropy,2023,25(8).

[31] Qiang L, Hanqiang H, Xiao-Wen Z, et al. Image encryption using fission diffusion process and a new hyperchaotic map[J]. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena,2023,175(P1).

Thai Finger-Spelling using Vision Transformer

Kullawat Chaowanawatee, Kittasil Silanon*, Thitinan Kliangsuwan
College of Computing, Prince of Songkla University Phuket Campus, Kathu, Phuket, Thailand

***Abstract*—In this paper, we present a finger-spelling recognition system that is based on Thai Sign Language (TFS) and employs a deep learning model called vision transformer. We extracted the 15 characters of the Thai alphabet from publicly available and our collected datasets to establish the recognition system. To train the learning model, we employed four EVA-02 vision transformer models, each of which showed impressive performance across different model sizes. We conducted four experiments to determine the most effective performance model. In Experiment 1, we directly trained the model to compare its performance. In Experiment 2, we used augmentation techniques to generate additional datasets. Experiment 3 utilized the Test-Time Augmentation (TTA) technique to generate test images with random variations. Lastly, in Experiment 4, we used Pseudo-Labelling (labeling labeled and unlabeled data) in each batch to train the model network. Furthermore, we developed a mobile application that collects user image data and provides helpful information related to finger-spelling, such as meanings, gestures, and usage examples.**

Keywords*—*Thai finger-spelling; vision transformer; deep learning; image recognition

I. INTRODUCTION

Sign language is a method of communication used by individuals who are deaf or hard of hearing. There are two primary types of sign language: vocabulary signs, which use hand shapes, movements, and facial expressions to convey word meanings, and finger-spelling, which uses only hand shapes to spell out technical words, such as names and locations. However, finger-spelling is not commonly used in everyday conversations, so many deaf individuals, especially children, struggle with it. Several finger-spelling recognition research studies have been proposed to help their skills. For instance, A. Deza and D. Hasan [1] used pre-processing techniques such as boosting contrast, edge enhancement, and the Convolution Neural Network (CNN) model to classify the American Sign Language (ASL) finger-spelling dataset. These techniques enabled the authors to achieve a validation accuracy of approximately 77%. R. A. Alawwad et al. [2] introduced Arabic Sign Language (ArSL) recognition systems using a Faster Region-based Convolutional Neural Network (R-CNN). The proposed approach yielded 93% accuracy and confirmed the robustness of background variations. K. R. Prajwal et al. [3] proposed a British Sign Language (BSL) finger-spelling recognition using transformer architecture to train models with noisy labels. Researchers employed a multi-stage training approach to achieve better performance. E. M. Martin and F. M. Espejo in [4] presented a system to interpret the Spanish Sign Language (LSE) finger-spelling. The pre-trained CNN model and Recurrent Neural Network (RNN) have been tested and compared. The CNN obtained a much better accuracy, with 96.42% being the maximum value. V. J. Schmalz [5] applied deep learning and fine-tuning techniques to build an automatic recognition system for Italian Sign Language (LIS) finger-spelling. The proposed CNN and VGG19 models were used for large-scale image and video recognition. The system obtained an accuracy of 99% with VGG19 and 97% with the CNN architecture. E. J. Ong et al. [6] suggested a Sequential Pattern Tree-based (SP-Tree) multi-class classifier for German Sign Language (DGS) and Greek Sign Language (GSL) finger-spelling recognition. Their proposed SP-Tree Boosting algorithm-based recognition model performs better than the Hidden Markov Model (HMM). X. Jiang et al. [7] proposed a novel finger-spelling identification method for Chinese Sign Language (CSL) via AlexNet-based transfer learning and Adam optimizer, which tested four different configurations of transfer learning. Although there has been research on finger-spelling in many languages, Thai Sign Language (TSL) finger-spelling will be the main focus of this work. P. Nakjai et al. [8] investigated a YOLO based on the convolution neural network architecture to localize and classify TSL finger-spelling. The system achieved the mean Average Precision (mAP) of 82.06% under a complex background and 84.99 % under a plain background. S. Lata and O. Surinta [9] proposed an end-to-end TSL finger-spelling recognition based on the YOLOv3 objection detection framework to create the most robust model with high recognition performance. P. Nakjai and T. Katanyukul [10] proposed automatic TSL finger-spelling recognition using CNN-based and Histogram of Oriented Gradients (HOG) based approaches. Experimental results have shown the viability of the proposed method, which achieves mAP at 91.26%. J. Sanalohit and T. Katanyyukul [11] invented MediaPipe Hands (MPH) trained model for hand-keypoint detection. The MPH can satisfactorily address single-hand schemes with accuracy of 84.57%

According to studies, the CNN model is the most commonly employed technology in finger-spelling research. Various model architectures [12], such as Xception, VGG16, InceptionV3, or ConvNeXt, are provided. However, the CNN model has some drawbacks, such as requiring a lot of data and computational resources and being sensitive to spatial distortions and variations. For this reason, a new deep learning technology called vision transformer (ViT) has been introduced for image recognition applications. The ViT offer a different approach by treating images as sequences of patches and applying self-attention to capture global dependencies and contextual information. This allows them to learn from less data and perform very well on image classification tasks. Therefore, we are interested in using the ViT approach due to its advantages in our work for TSL finger-spelling recognition. To find out the most effective performance recognition model. We performed four experiments. In Experiment 1, the model will be trained directly to compare model performance.

Experiment 2 will generate additional data sets using the augmentation technique. Experiment 3 uses the Test-Time Augmentation (TTA) technique to generate test images with random variations. Experiment 4 employed Pseudo-Labelling (labeled and unlabeled data) in each batch to train the model network. We organize the paper as follows. Section II describes the details of the dataset and ViT model. We present the experiment details and results in Section III. Section IV shows the implemented application using the recognition model as its classifier. Finally, we conclude this paper in SectionV.

II. Datasets and Vision Transformer Model for Image Classification

In this section, we first describe the classes of the dataset. Then we describe the details of vision transformer model and it's characteristic.

A. Datasets forTSL Finger-Spelling

This system recognizes 15 letters of TSL finger-spelling (see Fig. 1 and Table I) which use static-single-hand postures. We collected image hand postures from public datasets. The dataset was randomly partitioned into training and test sets, with an 80:20 ratios, resulting in 1,522 and 381 images in the respective subsets.

TABLE I. Static-Single-Hand Posture for TSL Finger-Spelling

Letter	Hand posture
" ก " (Ko kai)	Point the index and middle fingers with a thumb between them.
" ต " (To tao)	Form a fist with your thumb between the middle and index finger.
" ส " (So suea)	Make a fist and put your thumb on top of your fingers.
" พ " (Po phan)	Press your thumb against your middle finger with index pointed.
" ห " (Ho hip)	Stick out your middle and index fingers.
" บ " (Bo bimai)	Hold your fingers together with your thumb across your palm.
" ร " (Ro ruea)	Cross your index finger over your middle finger.
" ว " (Wo waen)	Hold up three fingers and spread them apart
" ด " (Do dek)	Touch your fingertips to your thumb and point.
" ฟ " (Fo fan)	Press your index finger and thumb together with straight fingers.
" ล " (Lo ling)	Form an L-shape with your thumb and index finger.
" ย " (Yo yak)	Stick out your pinkie and thumb.
" ม " (Mo ma)	Hold an invisible ball and poke your thumb through.
" น " (No nue)	Poke your thumb between your middle and ring.
" อ " (O ang)	Make a fist with your thumb pointing up.

B. Vision Transformer Model

The Vision Transformer (ViT) [13] are a type of deep learning architecture that uses the Transformer architecture. This is a significant departure from traditional computer vision architectures that rely on convolutional neural networks (CNNs). The ViT model works by first dividing the image into a sequence of patches. Each patch is then represented as a vector. These vectors for each patch are subsequently fed into a transformer encoder, a stack of self-attention layers. Self-attention is a mechanism that enables the model to learn long-range dependencies between the patches. This is crucial for image classification, as it allows the model to understand how different parts of an image contribute to its overall label. The output of the Transformer encoder is a sequence of vectors representing the image's features. The features are then used to classify the image. The ViTs have emerged as a powerful and promising alternative to CNNs in computer vision. Their ability to capture long-range dependencies, global attention, and data efficiency has made them a preferred choice for various computer vision tasks.

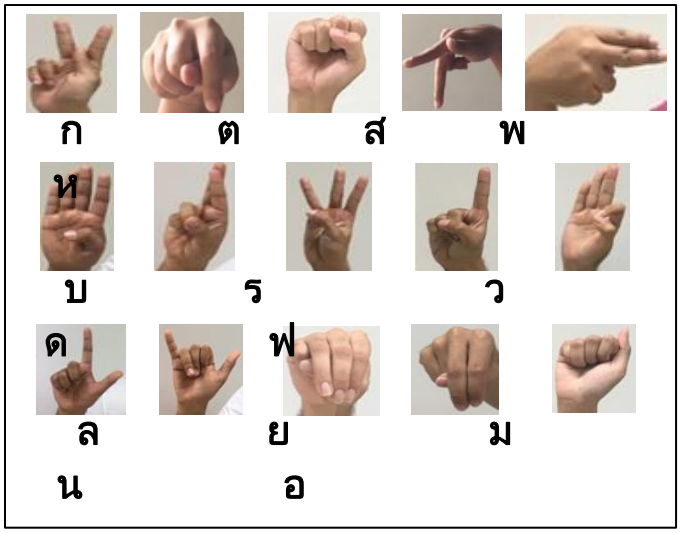


Fig. 1. Static-single-hand posture for TSL finger-spelling.

Currently, the ViT model that is gaining attention is a model named EVA02, presented by Y. Fang et al. [14]. The EVA-02 architecture is designed to be more efficient than previous Transformer-based models, enabling it to process images with greater speed and reduced computational resources. This superior performance makes it a valuable tool for tasks like object detection, semantic segmentation, and visual understanding. There are four variations of the EVA-02 model, with sizes ranging from 6M to 304M parameters. These models perform magnificently in image recognition tasks (see Table II).

TABLE II. Four Variations of EVA-02 Models

EVA-02 Model	Parameters (Million)	FLOPs (Billion)	Top-1 Accuracy on ImageNet (%)
EVA-02 Tiny	10M	0.6	76.2
EVA-02 Small	31M	1.6	77.5
EVA-02 Base	101M	3.3	79
EVA-02 Large	304M	7.6	81.4

III. Experimental Works

Four experiments were conducted to assess the classification performance of TSL finger-spelling recognition as discussed in this section.

A. Vision Transformer Training

In this experiment, we utilized four variations of the EVA-02 model, namely Tiny, Small, Base, and Large. All models

were trained directly on the dataset, with the input image size set to 224x224 pixels. We considered some crucial hyper parameters, including the learning rate of 2e-3, batch size set to 16, cross-entropy loss function, and 30 epochs for training the model. The K-Fold Cross Validation technique in [15] evaluates the predictive model. The dataset is divided into five subsets or folds. The model is trained and evaluated five times, using a different fold as the validation set. This technique aids in model assessment, selection, and hyper parameter tuning, providing a more reliable measure of a model's effectiveness. The model's generalization performance is evaluated using accuracy and weighted F1-score. Accuracy is a simple measure of overall correctness. At the same time, the weighted F1-score is a more sophisticated metric that considers class imbalances. Testing model results are shown in Table III. The large EVA-02 model at 1st fold achieved the best performance at 94.6% accuracy and 94.4% weighted F1-score. Fig. 2 shows the classification results of all 15 classes as a part of class performance analytics. Most errors occur in class "Mo ma" which produces incorrect predictions for classes "So suea" and "To tao" as a consequence of similar hand posture shapes, as displayed in Fig. 3. To improve the model's performance, we will discuss in the next section about the possibility of increasing the number of training examples.

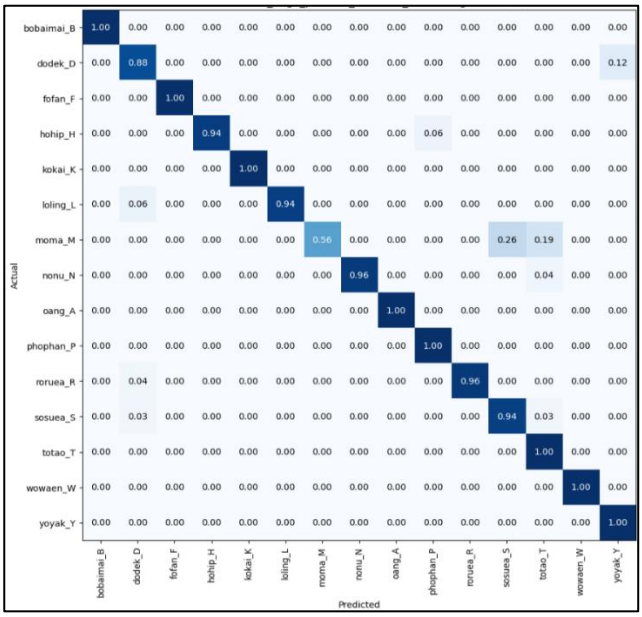


Fig. 2. Confusion matrix of the large EVA-02 at 1st fold.

Fig. 3. Classification error example of the large EVA-02 at 1st fold.

B. Data Augmentation

Data augmentation [16] is typically used during model training that expands the training set with modified copies of samples from the training dataset. This experiment simulates horizontal/vertical flipping, space transformation, random focus, noise addition, and slight rotation to train the model on more generalized data. Fig. 4 shows an example of a dataset after augmentation. Testing model results with the augmentation dataset are shown in Table IV. The large EVA-02 model at 3rd fold achieved the best performance at 94.6% accuracy and 94.5% weighted F1-score. Fig. 5 shows the classification results of all 15 classes as a part of class performance analytics. The model performance was improved when using the data transformation method, especially class "Mo ma"; the accuracy result increased to 70%. Class "So suea" and class "To tao" recognition errors decreased to 15% and 7%, respectively. For other classes, such as class "Po phan" or class "Yo yak", the accuracy dropped to 77% and 94% individually. In addition to augmenting the training dataset to improve the model's performance, we can also increase the number of testing datasets to enhance the accuracy of the model, which we will discuss in the following experiment.

TABLE III. MODEL PERFORMANCE EVALUATION

Model	Fold	Test dataset	
		Accuracy	Weighted F1
Tiny	1	0.6846	0.6693
	2	0.6590	0.6467
	3	0.6692	0.6608
	4	0.6821	0.6770
	5	0.7077	0.6835
Small	1	0.5897	0.5729
	2	0.6051	0.5809
	3	0.5462	0.5322
	4	0.6436	0.6366
	5	0.5590	0.5416
Base	1	0.8846	0.8784
	2	0.8513	0.8496
	3	0.8744	0.8723
	4	0.8897	0.8874
	5	0.9231	0.9210
Large	1	0.9462	0.9441
	2	0.8513	0.8377
	3	0.9128	0.9093
	4	0.8872	0.8832
	5	0.9359	0.9357

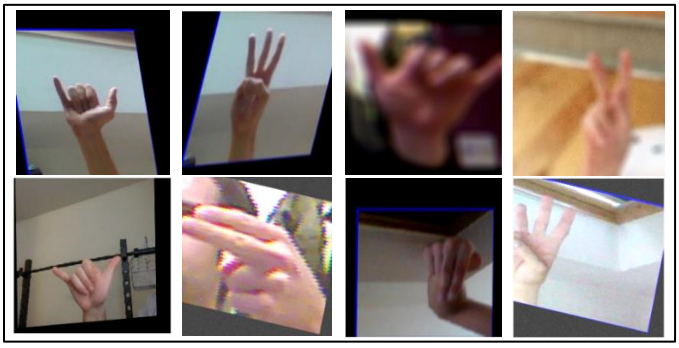


Fig. 4. Example of the data augmentation for training set.

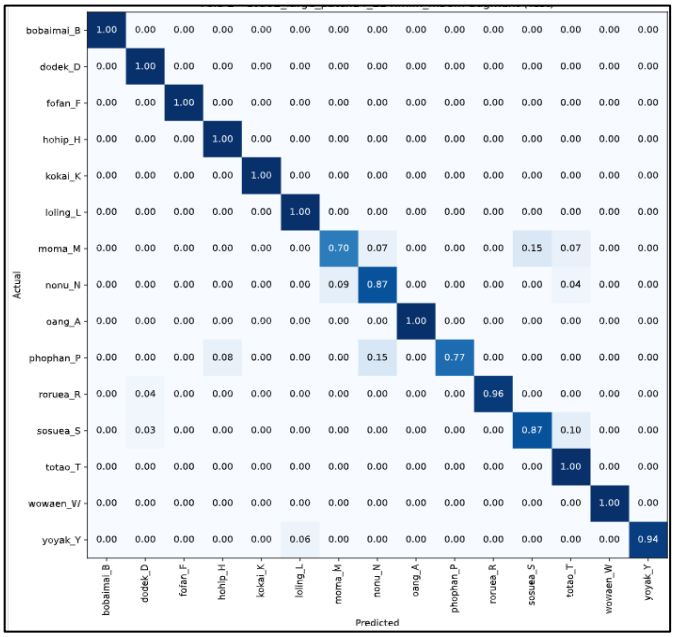


Fig. 5. Confusion matrix of the large EVA-02 at 3rd fold with data augmentation.

TABLE IV. DATA AUGMENTATION MODEL PERFORMANCE EVALUATION

Model	Fold	Test dataset	
		Accuracy	Weighted F1
Tiny	1	0.6590	0.6455
	2	0.7667	0.7521
	3	0.7231	0.7213
	4	0.8385	0.8303
	5	0.7821	0.7755
Small	1	0.8205	0.8082
	2	0.7051	0.6923
	3	0.8051	0.8016
	4	0.8308	0.8243
	5	0.8359	0.8270
Base	1	0.8641	0.8549
	2	0.9077	0.9034
	3	0.9231	0.9217
	4	0.9231	0.9219
	5	0.8769	0.8714
Large	1	0.9462	0.9435
	2	0.9462	0.9428
	3	0.9462	0.9450
	4	0.7513	0.7364
	5	0.9051	0.9041

C. Test-Time Data Augmentation

Test-Time Data Augmentation (TTA) [17] is a technique that can boost a model performance by applying augmentation to the test dataset. TTA is performed on multiple augmented copies of each image in the test dataset, having the model predict each and then returning an ensemble of those predictions to get a higher overall accuracy. For example, the image in Fig. 6 applies two transformations (rotation and color change) together with the original image. All of these images are passed to the same model and the results are averaged.

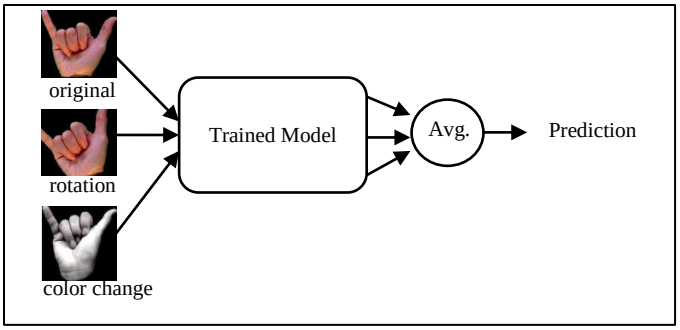


Fig. 6. Test-time data augmentation example.

TTA can be implemented to a model that has already been trained because, unlike training dataset augmentation, no model changes are required. We compared performance results of test dataset without TTA from previous experiment and test dataset using the TTA technique of models. From Table V, the large EVA-02 model at 3rd fold using TTA technique achieved best performance at 95.9% accuracy and 95.8% weighted F1-score. Fig. 7 shows the classification results of all 15 classes as a part of class performance analytics. Class "**Mo ma**" was improved, with accuracy increasing to 77%, while the other classes almost all exceeded 80% accuracy. In the following experiment, we discussed the ways to combine labeled and unlabeled dataset to improve the model's performance.

TABLE V. TEST-TIME DATA AUGMENTATION MODEL PERFORMANCE EVALUATION

Model	Fold	Test dataset without TTA		Test dataset using TTA	
		Accuracy	Weighted F1	Accuracy	Weighted F1
Tiny	1	0.6590	0.6455	0.6744	0.6597
	2	0.7667	0.7521	0.7872	0.7753
	3	0.7231	0.7213	0.7205	0.7152
	4	0.8385	0.8303	0.8538	0.8436
	5	0.7821	0.7755	0.7744	0.7677
Small	1	0.8205	0.8082	0.8282	0.8176
	2	0.7051	0.6923	0.7103	0.7005
	3	0.8051	0.8016	0.8026	0.7965
	4	0.8308	0.8243	0.8333	0.8229
	5	0.8359	0.8270	0.8179	0.8082
Base	1	0.8641	0.8549	0.8769	0.8679
	2	0.9077	0.9034	0.9154	0.9127
	3	0.9231	0.9217	0.9205	0.9196
	4	0.9231	0.9219	0.9359	0.9349
	5	0.8769	0.8714	0.8949	0.8888
Large	1	0.9462	0.9435	0.9513	0.9491
	2	0.9462	0.9428	0.9436	0.9402
	3	0.9462	0.9450	0.9590	0.9586
	4	0.7513	0.7364	0.7718	0.7588
	5	0.9051	0.9041	0.9026	0.9011

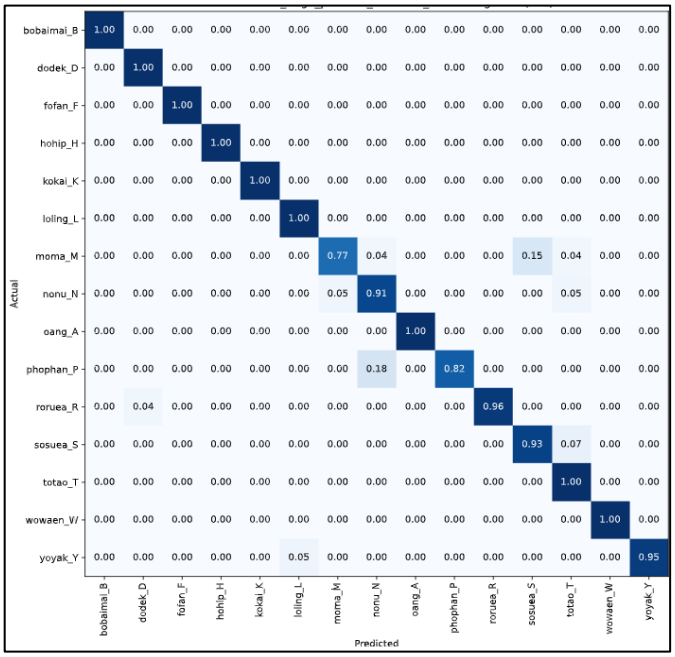


Fig. 7. Confusion matrix of the large EVA-02 at 3rd fold with TTA data augmentation.

D. Pseudo-Labeling

Pseudo labeling [18] is the process of using the trained model to predict labels for unlabeled data. A model has trained with the dataset containing labels and that model is used to generate pseudo labels for the unlabeled dataset. Finally, both the original labels dataset and pseudo labels dataset are combined to retrained the model. In this work, the best performance model from previous experiment is selected to predict labels for unlabeled data. The unlabeled data is collected from our subjects who were asked to stand in front of a white background. The labels dataset and pseudo labels dataset are used to retrain EVA-02 model. Finally, the new retrained model is used to predict the test dataset with TTA technique. Table VI shown performance model using Pseudo labeling. The large EVA-02 model at 1st fold using Pseudo labeling technique achieved best performance at 96.7% accuracy and 96.6% weighted F1-score. Fig. 8 shows the classification results of all 15 classes as a part of class performance analytics. Class "Mo ma" continues to be improved with accuracy increasing to 78%, while for some classes the accuracy has decreased but is still within the acceptable range of at least 80%. This experiment produced good recognition results, as indicated by 11 classes with 100% accuracy, more than any previous experiment. Based on the experiment, the most effective model will be selected as the recognition model which will be used in the program we will develop to recognize TSL finger-spelling recognition. The best EVA-02 model for user will depend on work specific needs. If users are looking for a model that is fast and efficient, then the EVA-02 Tiny or Small models may be a good choice. If user needs the best possible accuracy, then the EVA-02 Base or Large model is the best option.

TABLE VI. PSEUDO-LABELING MODEL PERFORMANCE EVALUATION

Model	Fold	Pseudo-labeling	
		Testing dataset using TTA	
		Accuracy	Weighted F1
Tiny	1	0.7515	0.7332
	2	0.8090	0.7978
	3	0.7885	0.7713
	4	0.8583	0.8509
	5	0.7947	0.7873
Small	1	0.8871	0.8830
	2	0.8604	0.8508
	3	0.8973	0.8942
	4	0.8480	0.8391
	5	0.9014	0.8992
Base	1	0.9425	0.9418
	2	0.9630	0.9628
	3	0.9405	0.9365
	4	0.9405	0.9372
	5	0.9507	0.9492
Large	1	0.9671	0.9663
	2	0.9651	0.9633
	3	0.9446	0.9395
	4	0.9405	0.9382
	5	0.9425	0.9406

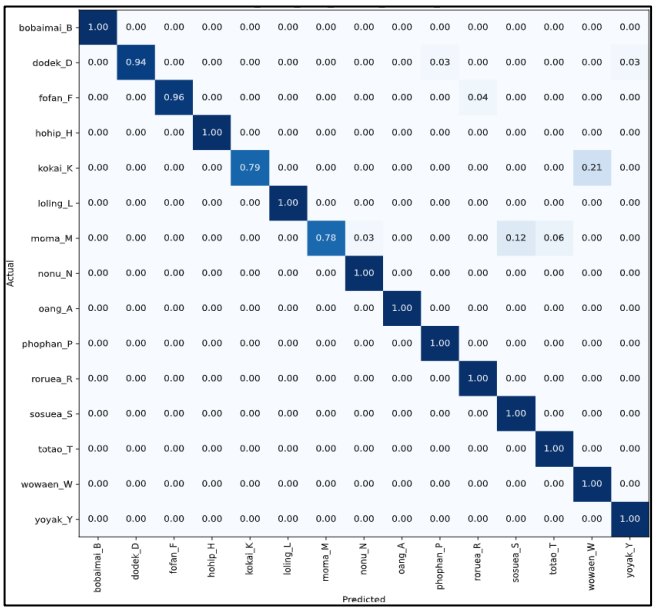
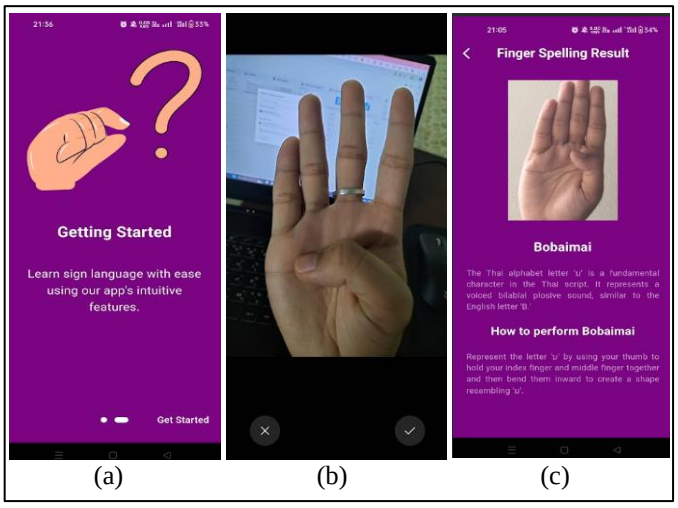


Fig. 8. Confusion matrix of the large EVA-02 at 1st fold with pseudo labeling.

IV. APPLICATION

In this section, we explain our mobile application that users can upload and classify their hand posture images with this application. The application is named "Thai Finger-Spelling Recognition" (see Fig. 9 (a)). We used the Flutter framework to build the front-end part of our mobile application to interact with user, while Python and the Flask framework were employed to develop the backend part to image classification process. The best classification model from experiments is deployed on the server. Users can either capture their own hand posture images or choose from their mobile gallery using the application's main interface (see Fig. 9 (b)). Subsequently, the application uploads the image to the server. The server then prompts the trained model to predict the most likely classes of Thai finger-spelling. The output class retrieves additional details from the database, including example images, class names, alphabet information, and hand posture demonstrations (see Fig. 9 (c)).



(a) (b) (c)

Fig. 9. Thai finger-spelling application: (a) User interface (b) Image capturing (c) Classification information.

V. CONCLUSION

In the study, we proposed training model experiments for Thai sign language finger-spelling recognition. The experiment consisted of vision transformer model training, training with data augmentation, test-time data augmentation, and pseudo-labeling. Our study achieved good performance compared with the experimental phases. In addition, we developed a mobile application. User can upload image data and process from classification to receive useful information related to finger-spelling, such as example image, hand postures, and usage information. For future work, we will collect more data and experiment for several types of Thai finger-spellings such as alphabets, vowels, and tones. Furthermore, we will be testing out other deep learning models like CNN, Capsule Networks, or Generative Adversarial Networks (GANs). The purpose is to compare their recognition performance with the ViT model and reduce model size while maintaining accuracy. This will make it possible to deploy the model in environments with limited resources such as mobile devices or embedded systems.

ACKNOWLEDGMENT

The authors would like to thank College of Computing and Prince of Songkla University, Phuket Campus, Thailand for their support and for providing a working area to test our system.

REFERENCES

[1] Deza and D. Hasan, "MIE 324 final report : sign language recognition," The computer engineering research group (the university of Totonto) [online] March 17 2023 Available: https://www.eecg.utoronto.ca/~jayar/mie324/asl.pdf

[2] R. A. Alawwad, O. Bchir, M. M. B. Ismail, "Arabic sign language recognition using Faster R-CNN," International Journal of Advanced Computer Science and Applications, vol. 12, no. 3, 2021

[3] K. R, H. Bull, L. Momeni, S. Albanie, G. Varol, and A. Zisserman, "Weakly-supervised Fingerspelling Recognition in British Sign Language Videos," 33rd British Machine Vision Conference, 2022.

[4] E. Martinez-Martin and F. Morillas-Espejo, "Deep learning techniques for Spanish sign language interpretation," Computational Intelligence and Neuroscience, vol. 2021, 2021.

[5] V. J. Schmalz, "Real-time Italian Sign Language Recognition with Deep Learning," in CEUR Workshop Proceedings, 2022, vol. 3078, pp. 45–57.

[6] E.-J. Ong, H. Cooper, N. Pugeault, and R. Bowden, "Sign language recognition using sequential pattern trees," in 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 2200–2207.

[7] X. Jiang, B. Hu, S. Chandra Satapathy, S.-H. Wang, and Y.-D. Zhang, "Fingerspelling identification for Chinese sign language via AlexNet-based transfer learning and Adam optimizer," Scientific Programming, vol. 2020, pp. 1–13, 2020.

[8] P. Nakjai, P. Maneerat, and T. Katanyukul, "Thai finger-spelling localization and classification under complex background using a YOLO-based deep learning," in Proceedings of the 11th International Conference on Computer Modeling and Simulation, 2019, pp. 230–233.

[9] S. Lata and O. Surinta, "An end-to-end Thai fingerspelling recognition framework with deep convolutional neural networks," ICIC Express Letters, vol. 16, no. 5, pp. 529–536, 2022.

[10] P. Nakjai and T. Katanyukul, "Hand sign recognition for Thai finger-spelling: An application of convolution neural network," Journal of Signal Processing Systems, vol. 91, pp. 131–146, 2019.

[11] J. Sanalohit and T. Katanyukul, "Thai finger-spelling recognition: Investigating MediaPipe Hands potentials," arXiv preprint arXiv:2201.03170, 2022.

[12] Keras, "Keras Documentation: Keras applications," Keras. [Online]. Available: https://keras.io/api/applications. [Accessed: 25-Sep-2023]

[13] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR, 2021.

[14] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao, "EVA-02: A Visual Representation for Neon Genesis," arXiv preprint arXiv:2303.11331, 2023.

[15] S. Pandian, "K-Fold Cross Validation Technique and its Essential," The Analytics Vidhya [online] September 25 2023 Available: https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/

[16] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," Information, vol. 11, no. 2, 2020, doi: 10.3390/info11020125.

[17] M. Kimura, "Understanding test-time augmentation," in International Conference on Neural Information Processing, 2021, pp. 558–569.

[18] D.-H. Lee and others, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in Workshop on challenges in representation learning, ICML, 2013, vol. 3, no. 2, p. 896.

Investigation of Deep Learning Based Semantic Segmentation Models for Autonomous Vehicles

Xiaoyan Wang*, Huizong Li
School of Computer Science and Technology, Nanyang Normal University, Nanyang, Henan, 473061, China

***Abstract*—Semantic segmentation plays a pivotal role in enhancing the perception capabilities of autonomous vehicles and self-driving cars, enabling them to comprehend and navigate complex real-world environments. Numerous techniques have been developed to achieve semantic segmentation. Still, the paper emphasizes the effectiveness of deep learning approaches because they have demonstrated impressive capabilities in capturing intricate patterns and features from images, resulting in highly accurate segmentation results. Although various studies have been conducted in literature, there is needed for a careful investigation and analysis of the existing methods, especially in terms of two critical aspects: accuracy and inference time. To address this need for analysis and investigation, the research focuses on three widely-used deep learning architectures: ResNet, VGG, and MobileNet. By thoroughly evaluating these models based on accuracy and inference time, the study aims to identify the models that strike the best balance between precision and speed. The findings of this study highlight the most accurate and efficient models for semantic segmentation, aiding the development of reliable self-driving technology.**

Keywords*—Semantic segmentation; autonomous vehicles; deep learning approaches; performance analysis; accuracy; inference time

I. Introduction

Semantic segmentation plays a pivotal role in the realm of computer vision, enabling machines to comprehend visual scenes by assigning each pixel of an image to a specific object category or class [1, 2]. This technique holds immense significance in a plethora of applications, including the navigation of autonomous vehicles [3]. The autonomous driving landscape, characterized by the emergence of self-driving cars, has transformed transportation paradigms [4]. Precise scene understanding through semantic segmentation is paramount in ensuring these vehicles' safe and efficient control in real-time scenarios [5, 6], enabling them to make informed decisions based on the interpretation of their surroundings from video feeds.

Autonomous vehicles, commonly referred to as self-driving cars, are reshaping the future of transportation [7]. Their ability to navigate complex environments autonomously relies on a myriad of technological advancements, and semantic segmentation stands as a linchpin among these. The process of accurately segmenting objects within a scene in real-time video feeds empowers self-driving cars to make split-second decisions [7-9], ensuring pedestrian safety, identifying lane boundaries, and interpreting traffic signals.

Existing methodologies in semantic segmentation for autonomous vehicles have made substantial strides. Deep learning-based approaches, in particular, have garnered significant attention due to their exceptional performance in complex tasks [10, 11]. This preference is attributed to their ability to automatically learn intricate features and patterns from vast datasets, ultimately leading to heightened accuracy [12]. Among the deep learning architectures, ResNet [13], VGG [14], and MobileNet [15] have emerged as frontrunners due to their efficiency in capturing nuanced spatial relationships and features within images [16]. However, despite these advancements, a need persists to identify the most effective and efficient deep learning-based method that strikes a balance between accuracy and inference time, thus optimizing the performance of semantic segmentation for autonomous vehicles.

The statement of the research problem is: How to achieve semantic segmentation for autonomous vehicles and self-driving cars using deep learning models that have high accuracy and low inference time. Correspondingly, the research questions are: What are the strengths and weaknesses of ResNet, VGG, and MobileNet architectures for semantic segmentation? How do these models compare in terms of accuracy and inference time on different datasets and scenarios? Which model(s) can provide the best balance between precision and speed for semantic segmentation?

In this study, we delve into the realm of DL-based models for semantic segmentation in autonomous vehicles, aiming to identify the most effective and efficient solutions. We examine three popular DL architectures: ResNet, VGG, and MobileNet, renowned for their contributions to computer vision tasks. Through a comprehensive analysis, we evaluate these models in terms of foreground accuracy, dice coefficient, and inference time, three crucial performance metrics in the context of autonomous driving systems.

Our findings reveal that certain DL models exhibit notable accuracy and efficiency in semantic segmentation for autonomous vehicles. By conducting an in-depth comparison between ResNet, VGG, and MobileNet architectures, we shed light on their respective strengths and weaknesses. Moreover, we identify the DL models that excel in terms of accuracy and inference time, providing valuable insights for practitioners and researchers in the field. The results of this study serve as a guide to selecting appropriate DL models for real-time semantic segmentation tasks in autonomous vehicles, ultimately contributing to the advancement and reliability of self-driving technologies. By rigorously evaluating these models' performance on video data, this study aims to

contribute insights that advance the state-of-the-art in semantic segmentation for autonomous vehicles. This research endeavors to identify the most effective and efficient deep learning approach through meticulous experimentation and analysis, thereby fostering safer and more reliable autonomous driving systems.

II. Related Works

Ghosh et al. [17] introduced SegFast-V2, an approach to semantic image segmentation tailored for autonomous driving scenarios. Notably, the method prioritizes efficiency by utilizing fewer parameters within deep learning frameworks. With a focus on achieving accurate semantic segmentation, especially in the context of self-driving vehicles, SegFast-V2 presents a solution that balances computational efficiency and performance. The research contributes to advancing the field of autonomous driving by addressing the challenge of efficient and effective semantic segmentation, which is crucial for safe and reliable navigation in complex environments.

Colley et al. [18] investigated the impact of visualizing semantic segmentation in highly automated vehicles on trust, situation awareness, and cognitive load. By examining how providing visual cues of semantic segmentation affects drivers' perceptions and cognitive demands, the research aims to uncover insights into human-vehicle interaction dynamics. By analyzing the implications of semantic segmentation visualization on trust levels, understanding of the driving context, and mental workload, the paper enhances the design and implementation of automated driving systems to optimize driver experience, safety, and overall performance.

Nesti et al. [19] assessed the resilience of semantic segmentation methods employed in autonomous driving scenarios against real-world adversarial patch attacks. Focusing on the critical task of accurately segmenting objects in complex driving environments, the study investigates the vulnerability of these methods to deliberate perturbations introduced by adversarial patches. By subjecting various semantic segmentation models to these real-world attacks, the research endeavors to unravel the potential weaknesses and challenges of such vulnerabilities in ensuring safe and reliable autonomous driving systems. Through meticulous evaluation and analysis, the paper sheds light on the robustness of semantic segmentation techniques under adversarial conditions, offering valuable insights into enhancing the security and performance of self-driving vehicles. The author in Mo et al. [20] conducts a comprehensive review of the latest advancements in semantic segmentation technologies grounded in deep learning methodologies. By critically examining the current state-of-the-art approaches, the study aims to provide an in-depth understanding of the evolution and capabilities of deep learning-based semantic segmentation. Through the analysis of various models, architectures, and techniques, the paper contributes to the field's knowledge by outlining cutting-edge solutions that leverage deep learning for precise object delineation and scene understanding in diverse applications.

Dang et al. [21] presented a lightweight pixel-level semantic segmentation technique based on deep learning for the purpose of detecting and analyzing sewer defects. By leveraging deep learning methods, the approach offers an efficient solution for identifying and classifying sewer system issues through pixel-level segmentation. The study's focus on lightweight architecture signifies a commitment to computational efficiency while maintaining accurate defect identification. This research contributes to the field of sewer infrastructure maintenance by offering a streamlined approach that employs deep learning for detailed and effective defect analysis, enhancing the overall assessment and management of sewer systems.

As results, there are many existing methods for semantic segmentation, but they need to be carefully investigated and analyzed, especially in terms of two critical aspects: accuracy and inference time. Accuracy is the measure of how well the model can correctly segment the image and match the ground truth labels. Inference time is the measure of how fast the model can process the image and produce the segmentation output. These two aspects are important because they affect the performance and safety of the autonomous vehicles and self-driving cars. A model that has high accuracy can provide more reliable and detailed information for the vehicle, while a model that has low inference time can respond more quickly and adapt to changing situations. Therefore, the research paper wants to find the best balance between accuracy and inference time for semantic segmentation.

III. Material and Method

A. Dataset Overview

The Cambridge-driving Labeled Video Database, commonly known as CamVid, is a comprehensive and meticulously annotated dataset designed to advance the field of computer vision, particularly in the context of autonomous driving and scene understanding. The CamVid stands as a vital resource in the realm of computer vision with its diverse and meticulously labeled video sequences. CamVid features a diverse collection of high-resolution video sequences captured from a moving vehicle navigating through urban and suburban environments. These videos encompass a wide range of real-world driving scenarios, presenting challenges such as varying lighting conditions, dynamic traffic, and intricate road layouts. One of the distinguishing aspects of CamVid is its extensive labeling. Each frame of the dataset is meticulously annotated with pixel-level semantic segmentation labels. This means that every pixel in the video frames is categorized, providing a detailed understanding of the objects and structures present in the scenes. Such detailed annotations enable the training and evaluation of advanced machine-learning models for tasks like object detection, semantic segmentation, and instance segmentation. CamVid's applications extend beyond autonomous driving research. The dataset's rich annotations make it highly suitable for projects related to urban scene understanding, environmental monitoring, and general semantic segmentation challenges. The dataset consists of a series of videos, each accompanied by semantic labels that categorize object classes. These labels are accompanied by additional metadata. The database includes accurate reference labels that link every individual pixel to one of 32 predefined semantic categories. Fig. 1 [22] demonstrates the semantic classes of the CamVid dataset.

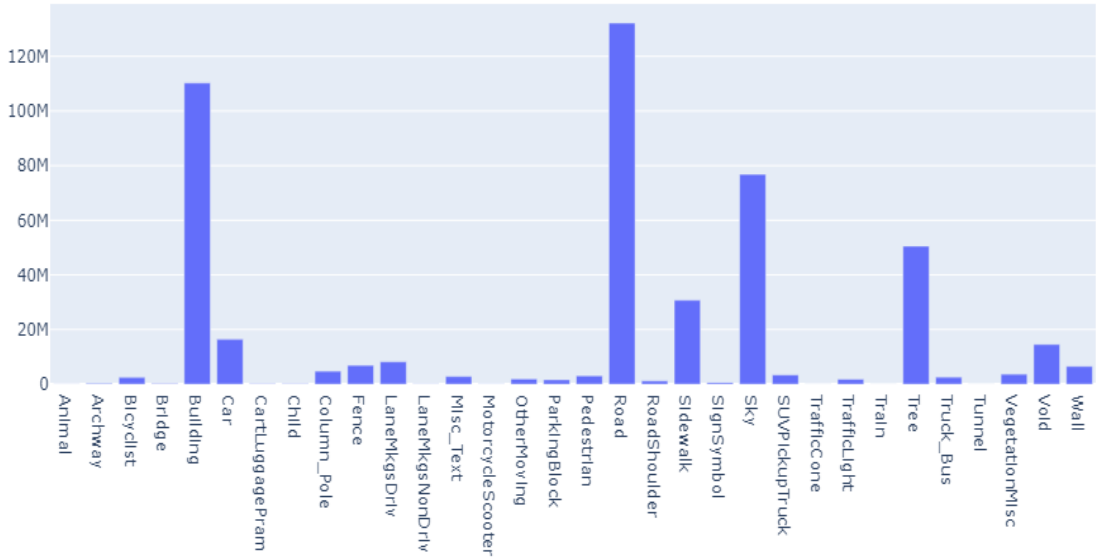


Fig. 1. Semantic classes of the camvid dataset.

B. Model Learning

The proposed model in this study is designed to acquire detailed annotation for each individual pixel within a scene recorded from the perspective of an autonomous agent. The primary task of this model is to classify and isolate every pixel in the given scene into one of 32 specific categories. These categories include items such as roads, pedestrians, sidewalks, and cars, as showcased in the animated image of our product. This enables interaction with any particular image.

The main objective is to understand and interpret the scene with exceptional precision. This understanding is achieved by categorizing each pixel into specific classes, which represent different objects or entities within the scene. For instance, a road, a pedestrian, a sidewalk, a car, and more – all of these are examples of classes that the model identifies. Imagine a picture of a street: the road, the people walking on the sidewalk, the parked cars, and other elements are contained. Our model performs something similar but for each and every pixel in the scene. It determines if a pixel belongs to the road, the sidewalk, a person, a car, or one of the other predefined categories – a total of 32 categories.

C. Backbones

For our initial set of experiments, we opted to employ a straightforward architecture that draws inspiration from the UNet model. This architecture incorporates backbones like ResNet50, VGG19, and MobileNetV2. Despite its simplicity in terms of implementation, this architecture has proven to be remarkably robust in terms of its performance. In other words, it strikes a balance between being relatively easy to create and yielding impressive results in various tasks.

1) ResNet50 Backbone: The UNet architecture is a convolutional neural network (CNN) design that excels in image segmentation tasks. It consists of an encoding path that gradually reduces spatial resolution while capturing features and a decoding path that restores the resolution while refining segmentation maps. In this baseline, we enhance the UNet with a ResNet50 backbone, which is a deep residual network known for its excellent performance in various computer vision tasks. ResNet50 incorporates skip connections to mitigate vanishing gradient issues during training. As shown in Fig. 2 [23], the architecture of ResNet50 is depicted.

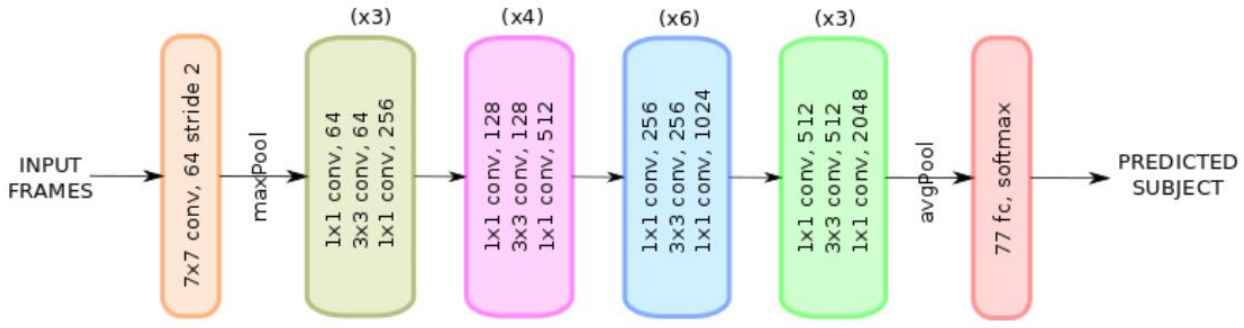


Fig. 2. ReseNet50 architecture [23].

As shown in Fig. 2, the ResNet50 is a specific variant of the ResNet architecture, characterized by its depth of 50 layers. It builds upon the original ResNet's innovation of residual connections, enhancing its capacity to capture complex patterns and features from images. By incorporating a series of residual blocks, ResNet50 enables the efficient training of deeper neural networks while mitigating issues related to vanishing gradients. This architecture has proven highly effective in various computer vision tasks, such as image recognition and segmentation. In the context of enhancing the perception capabilities of autonomous vehicles, ResNet50's depth and feature-extraction prowess contribute to accurate and detailed semantic segmentation, aiding in the vehicles' understanding and navigation of intricate real-world environments.

2) VGG19 backbone: Similar to the previous architecture, this baseline employs a UNet structure but integrates a VGG19 backbone. VGG19 is a deep CNN architecture known for its simplicity and effectiveness. It consists of multiple convolutional layers followed by max-pooling, and it captures progressively complex features through its layers. This backbone enhances the UNet's feature extraction capabilities, contributing to better segmentation performance. Fig. 3 demonstrates the architecture of VGG19.

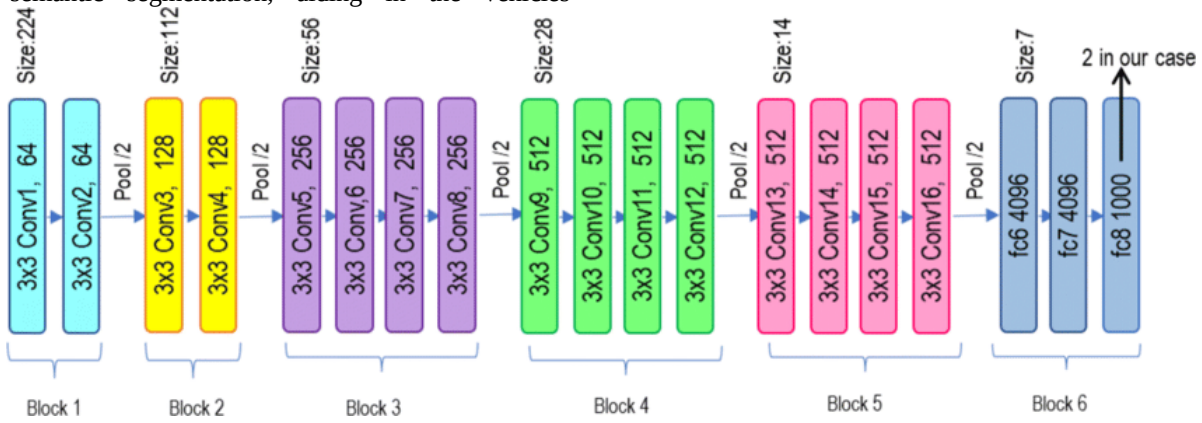


Fig. 3. VGG19 architecture [24].

As shown in Fig. 3, VGG19 is a convolutional neural network architecture renowned for its simplicity and effectiveness in image recognition tasks. With 19 layers, it follows a straightforward design principle of stacking multiple 3x3 convolutional layers, followed by max-pooling layers for down-sampling. This repetitive structure results in a deep network capable of capturing intricate features at different levels of abstraction. VGG19's uniform architecture makes it easy to understand and implement, contributing to its popularity. In the context of enhancing the perception capabilities of autonomous vehicles, VGG19's depth and feature-extraction capabilities play a crucial role in semantic segmentation, enabling the vehicles to accurately perceive and navigate complex real-world scenarios.

3) MobileNetV2 Backbone: The UNet design combined with a MobileNetV2 backbone represents a lightweight yet powerful configuration. MobileNetV2 is optimized for efficiency and speed, making it suitable for real-time applications on resource-constrained devices. It utilizes depthwise separable convolutions to reduce computational complexity while preserving accuracy. Fig. 4 illustrates the architecture of MobileNetV2

The architecture's core component is depth-wise separable convolutions. In these convolutions, the spatial information is decoupled from the channel-wise information, reducing the computational load. Each convolution is divided into a depth-wise convolution, which applies a single convolutional filter to each input channel, followed by a point-wise convolution that merges the outputs into the desired number of output channels. MobileNetV2 also employs skip connections to retain important features, facilitating the flow of gradients during training. These innovative design choices collectively result in a lightweight architecture capable of achieving impressive accuracy on tasks like image classification and semantic segmentation.

D. Hyperparameter Tuning

To enhance the efficacy of our baseline model, a dual focus on both optimal model selection and hyperparameter tuning becomes imperative. The crux lies in identifying not only the most suitable model architecture but also the optimal configuration of hyperparameters for training. To achieve this, we utilize a Bayesian hyperparameter search methodology, a sophisticated technique aimed at systematically exploring the hyperparameter space to unearth the combination that yields the most favorable results.

The essence of this method revolves around minimizing the model's loss function when evaluated against a dedicated validation dataset. By leveraging a Bayesian approach, we dynamically adapt the search process based on previous iterations, progressively honing in on the most promising areas of the hyperparameter space. This method is especially effective in mitigating the challenges posed by high-dimensional and complex search spaces. Ultimately, the outcome of this meticulous hyperparameter search is a refined model configuration that not only aligns with the chosen architecture but also significantly bolsters the model's performance, setting the stage for more accurate and robust predictions.

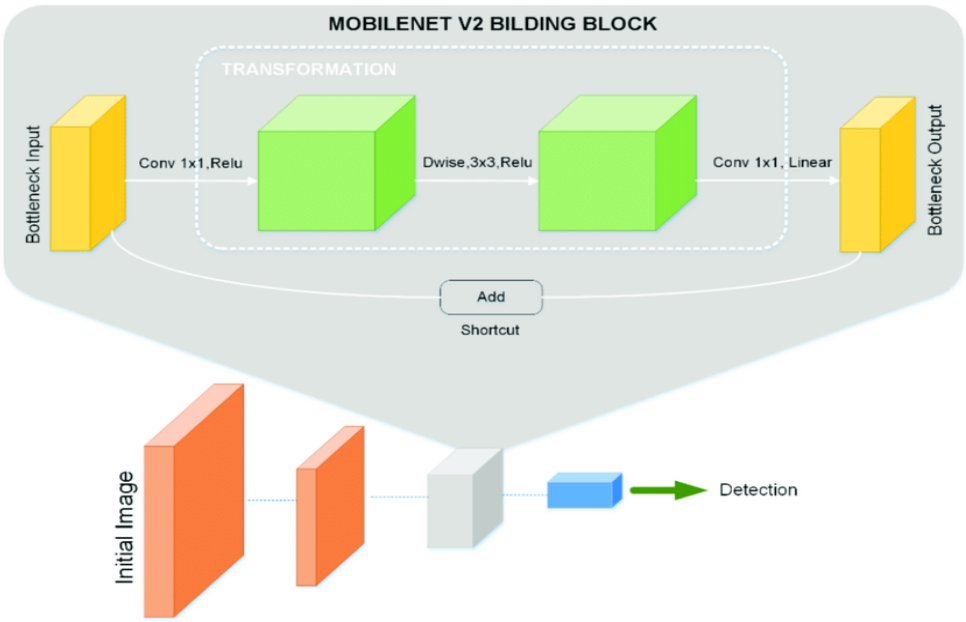


Fig. 4. MobileNetV2 architecture [25].

IV. Experimental Results and Performance Analysis

The forthcoming section delves into the outcomes of our experiments, shedding light on the results obtained through rigorous testing and evaluation. It's important to underscore that for reasons of safety paramountcy, we have prioritized specific classes during the training process. These priority classes encompass entities that play pivotal roles in ensuring safety within various scenarios. The prioritized classes, due to their pronounced safety implications, encompass Pedestrians, Bicyclists, Children, Cars, Heavy Vehicles, and Traffic lights. These categories encapsulate elements that are central to the smooth functioning of urban environments and the safety of both pedestrians and drivers. By focusing on these classes during training, we aim to equip our model with the capability to distinctly recognize and respond to these critical entities.

When evaluating the performance of models in semantic segmentation tasks, two essential metrics that provide valuable insights are Foreground accuracy and the dice coefficient. Foreground accuracy measures the precision with which a model correctly classifies the foreground objects of interest, which is particularly important in scenarios where specific classes carry more significance.

A. Foreground Accuracy

Foreground accuracy, in the context of semantic segmentation, is a metric used to gauge the accuracy of a model's predictions specifically concerning the foreground objects or classes of interest. Unlike overall accuracy, which considers all classes equally, foreground accuracy focuses solely on how well the model correctly identifies and classifies the relevant objects, ignoring the background and other unimportant classes. This metric provides a more insightful evaluation of a model's performance in tasks where certain classes are of greater significance than others, such as object detection or scene segmentation. It is computed by dividing the number of correctly classified foreground pixels by the total number of foreground pixels and can be represented as:

$$Forground\ Accuracy = True\ Positive\ (TP) + False\ Negative(FN) \Big/ True\ Positive\ (TP)$$

B. Dice Metric

The dice coefficient, also known as the F1 score, offers a comprehensive assessment of segmentation accuracy by considering false positives and false negatives. It quantifies the overlap between the predicted segmentation and the ground truth, producing a value between 0 and 1, where 1 indicates perfect alignment. Fig. 5, 6 and 7 show dice metric for the methods. The dice metric is calculated as follows:

$$Dice = 2 * True\ Positive\ (TP) \Big/ 2 * True\ Positive\ (TP) + False\ Positive(FP) + False\ Negative(FN)$$

C. Backbone Experiments

In this section, we delve into the backbone experiments, wherein the focus lies on the fundamental architectural components of our models. These backbone models excel at assimilating contextual information from expansive image regions. This is achieved by adeptly pooling features through a variety of window sizes and seamlessly integrating them using both residual connections and adaptable weights. Our investigation involves subjecting the baseline models to thorough experimentation, including an exploration of different loss functions, to comprehensively understand their performance and capabilities. We present some experiments corresponding to the baselines. Table I presents the experiment name associated with each backbone.

D. Experiments with Hyperparameters

As discussed in Section 5.6, in the process of assessing the effectiveness of models such as ResNet50, VGG19, and MobileNetV2, we harness the capabilities of the Sweep tool. This utility streamlines the execution of a Bayesian hyperparameter search strategy, which is employed to minimize the model's loss on the validation dataset. Sweeps significantly simplify our ability to conduct various experiments while utilizing this search method. The results of foreground accuracy and dice are shown in Fig. 8, 9, 10 and 11.

TABLE I. EXPERIMENT NAME FOR THE BACKBONES

Experiment name	Backbone name
baseline-train-1	ResNet50
baseline-train-2	VGG19
baseline-train-3	MobileNetV2

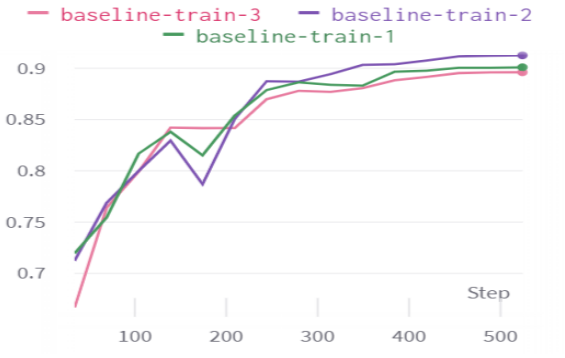


Fig. 5. Foreground accuracy for baseline experiments.

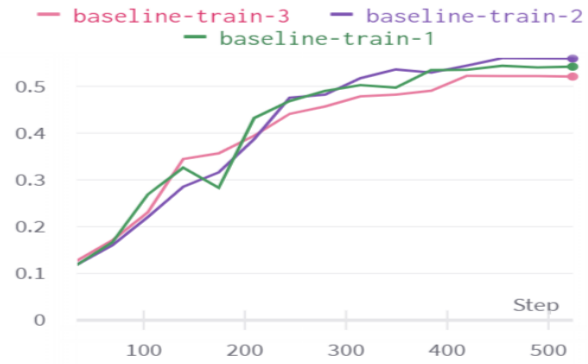


Fig. 6. Dice score for baseline experiments.

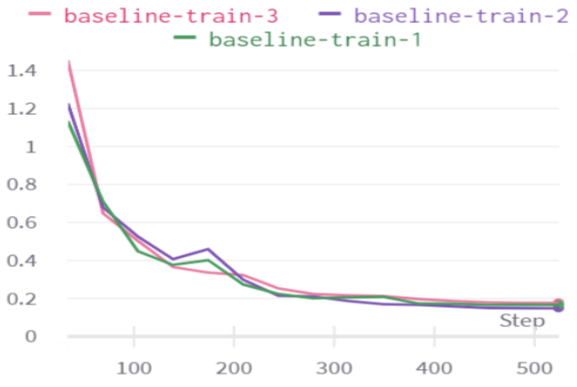


Fig. 7. Validation loss for baseline experiments.

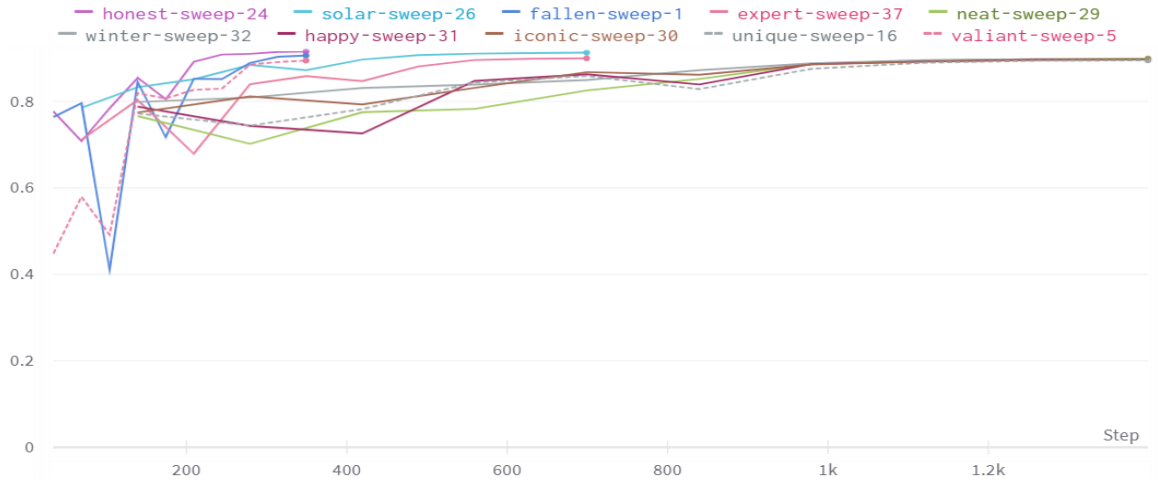


Fig. 8. Foreground accuracy.

Fig. 8 depicts foreground accuracy for various hyper-parameters tuning using Sweep. The provided chart illustrates the Foreground accuracy achieved across ten distinct experiments, offering a comparative analysis of various fine-tuned models. Among these experiments, the recorded accuracy values reveal a hierarchy of performance. Notably, the experiment named "honest-sweep-24" attains the highest accuracy, followed by "solar-sweep-26," "valiant-sweep-5," "solar-sweep-26" once again, and "expert-sweep-37." Conversely, the models "neat-sweep-29," "winter-sweep-32," "happy-sweep-31," "iconic-sweep-30," and "unique-sweep-16" exhibit the lowest accuracy values.

Starting with the highest accuracy achieved in the experiment labeled "honest-sweep-24," it showcases the model's exceptional ability to accurately classify foreground objects. The meticulously tuned parameters of this model contribute to its precision in segmenting relevant classes within the image. This high accuracy score signifies that the model successfully distinguishes and labels the target objects, a crucial feat in tasks like object recognition or scene understanding. The reliability of the "honest-sweep-24" experiment's outcome implies that its fine-tuning process effectively optimized its performance, rendering it a formidable contender in semantic segmentation tasks.

On the other hand, Fig. 9 presents the performances of the models with various backbones in terms of mean foreground accuracy. The backbones involve resnet34, resnet50, Resnet18, vgg19, mobilenetv2_100, mobilenetv3_large_100, mobilenetv3_small_050. The mean foreground accuracy indicates how well the model is performing in correctly classifying instances belonging to the class of interest, often in the context of object recognition or segmentation.

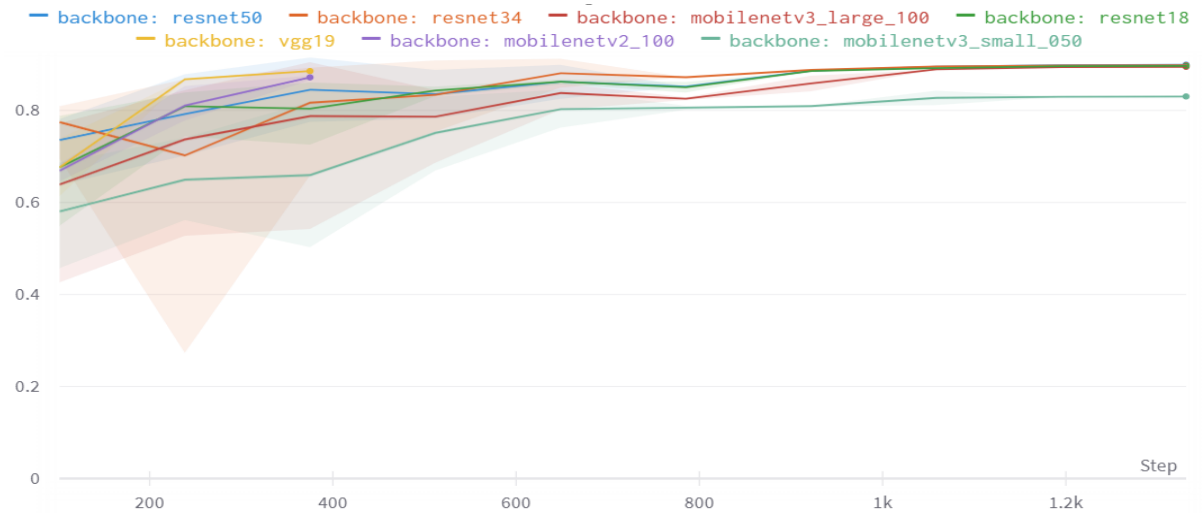


Fig. 9. Mean foreground accuracy for different backbones.

As shown in Fig. 9, among the listed backbone models, "Resnet18" stands out as the best performer, with an accuracy of 89.67%. Resnet18 is a variant of the Residual Network (ResNet) architecture, designed with 18 layers. This architecture utilizes residual blocks, allowing it to efficiently train deep neural networks by mitigating the vanishing gradient problem. Resnet18's success can be attributed to several factors. Firstly, its moderate depth strikes a balance between model complexity and capacity, preventing overfitting while still capturing intricate features in the data. Secondly, Resnet18's residual connections enable efficient gradient flow during training, fostering better convergence and feature representation. Thirdly, Resnet18's design incorporates skip connections, which allow information to bypass certain layers, further enhancing its ability to capture relevant features.

Comparatively, other backbones might struggle due to either excessive complexity leading to overfitting (as with deeper architectures) or limited depth hindering feature extraction (as with shallower architectures). Resnet18 strikes a favorable balance, resulting in its superior mean foreground accuracy. Its intermediate depth, residual connections, and skip connections collectively contribute to achieving a strong balance between capacity and generalization, making Resnet18 a top performer in the comparison.

Finally, the mean dice score presents different backbones, as shown in Fig. 8. It quantifies the similarity between predicted and ground truth segmented regions by measuring the overlap of pixels. This metric's significance lies in its ability to assess the model's capability to accurately delineate object boundaries and capture fine-grained details in complex scenes, providing a comprehensive measure of segmentation quality and performance.

As demonstrated in Fig. 10 presents the mean dice score values collected for different backbone architectures, including "resnet34," "resnet50," "vgg19," "mobilenetv2_100," and "mobilenetv3_large_100." These scores reflect the performance of each backbone on specific tasks or datasets. Notably, "resnet34" emerges as the top-performing architecture with the highest dice scores across multiple columns, followed closely by "resnet50." Both these architectures consistently exhibit superior performance compared to others like "vgg19," "mobilenetv2_100," and "mobilenetv3_large_100." The provided scores reflect the efficacy of these backbones in tackling the given tasks, with "resnet34" standing out as a particularly strong model.

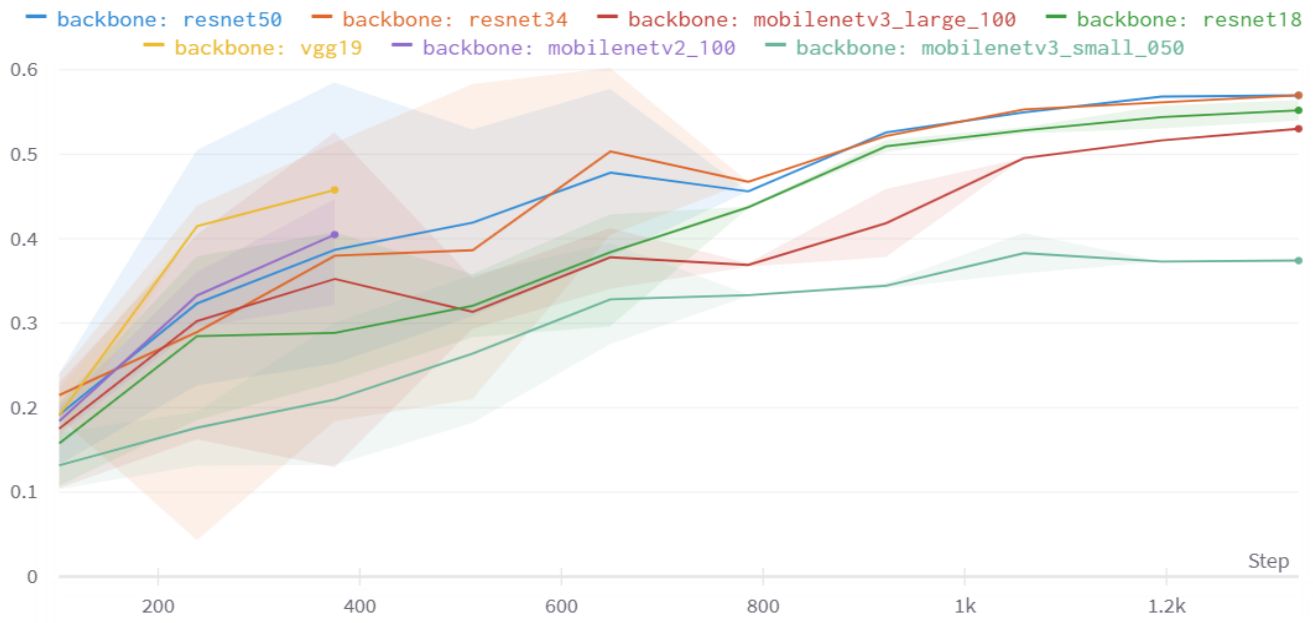


Fig. 10. Mean dice score for the backbones.

E. Inference Time

Inference time refers to the amount of time a model takes to process an input and produce an output prediction. In the context of semantic segmentation models, it measures how quickly a model can analyze an image and generate pixel-wise segmentation results.

The inference times were collected for various backbone models, including "resnet34," "resnet50," "Resnet18," "vgg19," "mobilenetv2_100," "mobilenetv3_large_100," and "mobilenetv3_small_050." Notably, the "Resnet18" model achieved the lowest inference time, while "Vgg19" had the highest.

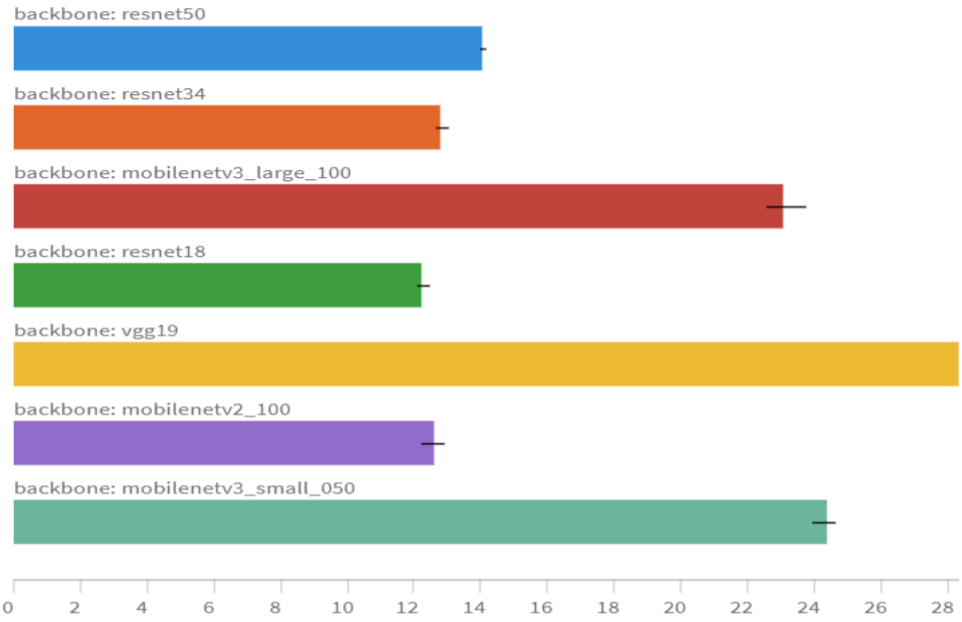


Fig. 11. Inference time of different backbones.

As illustrated in Fig. 11, the superior model in terms of inference time appears to be "Resnet18," which exhibits the lowest processing time among the listed models. This faster inference time can be attributed to Resnet architectural design, which balances depth and complexity, enabling efficient feature extraction while minimizing computational overhead. In contrast, "Vgg19," while offering strong performance, likely incurs higher inference times due to its greater depth and more complex architecture. Therefore, "Resnet18" emerges as the superior choice for applications that prioritize faster semantic segmentation inference times.

V. DISCUSSION

The analysis conducted yields critical insights that significantly inform the optimization of the training process and the selection of backbone architectures for a semantic segmentation model. Firstly, it's evident that employing lower learning rates and weight decay parameters leads to improved foreground accuracy and dice scores. This underscores the necessity of precise parameter tuning to achieve superior segmentation results.

Secondly, the study identifies the batch size and image resize factor as key factors with strong positive correlations to the evaluation metrics. This highlights the pivotal role these factors play in shaping model performance, emphasizing their potential for enhancing accuracy and dice scores.

Additionally, caution is advised against the utilization of VGG-based backbones for the final model. The findings suggest that these architectures are susceptible to vanishing gradients, which can impede gradient propagation during training and hinder optimal model performance.

Ultimately, the analysis underscores the superior performance of ResNet backbones across various metrics. ResNet34 and ResNet50 emerge as optimal choices for the final model due to their impressive performance and quicker inference times compared to other architectures. These insights provide actionable recommendations for refining model training, selecting appropriate backbones, and ultimately improving the efficiency and accuracy of semantic segmentation models.

VI. CONCLUSION

This study extensively explores various deep learning models for semantic segmentation, particularly ResNet, VGG, and MobileNet architectures, aiming to determine the most effective and efficient approach in terms of accuracy and inference time. Through thorough analysis of real-world video data, the research strives to advance semantic segmentation for autonomous vehicles, enhancing their safety and reliability. The investigation's outcomes offer crucial insights into optimizing training processes and selecting backbone architectures, with lower learning rates and weight decay parameters enhancing accuracy, while the batch size and image resize factor positively influence model performance. Caution against using VGG-based backbones due to vanishing gradients is noted, favoring ResNet34 and ResNet50 for their strong metrics and quicker inference times. These findings provide actionable guidelines for refining model training and selecting suitable backbones to enhance the efficiency and precision of semantic segmentation models. For future studies, firstly, researchers could explore hybrid architectures that combine the strengths of ResNet, VGG, and MobileNet for semantic segmentation. This approach seeks to harness the unique features of each architecture to create novel solutions that offer a balance between accuracy, efficiency, and robustness. Secondly, a promising avenue involves enhancing the adversarial robustness of semantic segmentation models. This entails investigating techniques to counter real-world adversarial attacks, particularly in the context of self-driving vehicles. By developing defense mechanisms against such attacks, researchers could contribute to improving the reliability and safety of autonomous driving systems.

ACKNOWLEDGMENT

This work was supported in part by the Natural Science Foundation of Anhui Province of China under Grant 1808085MG221.

REFERENCES

[1] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," International journal of multimedia information retrieval, vol. 7, pp. 87-93, 2018.

[2] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," Neurocomputing, vol. 406, pp. 302-321, 2020.

[3] B. Chen, C. Gong, and J. Yang, "Importance-aware semantic segmentation for autonomous vehicles," IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 1, pp. 137-148, 2018.

[4] Q. Sellat, S. Bisoy, R. Priyadarshini, A. Vidyarthi, S. Kautish, and R. K. Barik, "Intelligent semantic segmentation for self-driving vehicles using deep learning," Computational Intelligence and Neuroscience, vol. 2022, 2022.

[5] M. Ivanovs, K. Ozols, A. Dobrajs, and R. Kadikis, "Improving semantic segmentation of urban scenes for self-driving cars with synthetic images," Sensors, vol. 22, no. 6, p. 2252, 2022.

[6] M. C. ANG, A. AGHAMOHAMMADI, K. W. NG, E. SUNDARARAJAN, M. MOGHARREBI, and T. L. LIM, "MULTI-CORE FRAMEWORKS INVESTIGATION ON A REAL-TIME OBJECT TRACKING APPLICATION," Journal of Theoretical & Applied Information Technology, vol. 70, no. 1, 2014.

[7] A. Moorthy, B. Sivashanmugam, R. Sriram, and M. Swathi, "Real Time Image and Video Semantic Segmentation For Self-Driving Cars," Journal of Survey in Fisheries Sciences, vol. 10, no. 2S, pp. 3208-3216, 2023.

[8] Q. H. Che, D. P. Nguyen, M. Q. Pham, and D. K. Lam, "TwinLiteNet: An Efficient and Lightweight Model for Driveable Area and Lane Segmentation in Self-Driving Cars," arXiv preprint arXiv:2307.10705, 2023.

[9] M. Ang, E. Sundararajan, K. Ng, A. Aghamohammadi, and T. Lim, "Investigation of Threading Building Blocks Framework on Real Time Visual Object Tracking Algorithm," Applied Mechanics and Materials, vol. 666, pp. 240-244, 2014.

[10] Q. Sellat, S. K. Bisoy, and R. Priyadarshini, "Semantic segmentation for self-driving cars using deep learning: a survey," in Cognitive Big Data Intelligence with a Metaheuristic Approach: Elsevier, 2022, pp. 211-238.

[11] V. Bhavadharshini, S. Mridula, B. Sakthipriya, and J. J. Gracewell, "Semantic Segmentation using Convolutional Neural Networks," in 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), 2023: IEEE, pp. 481-488.

[12] C. Chen, C. Wang, B. Liu, C. He, L. Cong, and S. Wan, "Edge intelligence empowered vehicle detection and image segmentation for autonomous vehicles," IEEE Transactions on Intelligent Transportation Systems, 2023.

[13] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," arXiv preprint arXiv:1603.08029, 2016.

[14] A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, "Going deeper in spiking neural networks: VGG and residual architectures," Frontiers in neuroscience, vol. 13, p. 95, 2019.

[15] D. Sinha and M. El-Sharkawy, "Thin mobilenet: An enhanced mobilenet architecture," in 2019 IEEE 10th annual ubiquitous computing, electronics & mobile communication conference (UEMCON), 2019: IEEE, pp. 0280-0285.

[16] M. Abu, A. Amir, Y. Lean, N. Zahri, and S. Azemi, "The performance analysis of transfer learning for steel defect detection by using deep learning," in Journal of Physics: Conference Series, 2021, vol. 1755, no. 1: IOP Publishing, p. 012041.

[17] S. Ghosh, A. Pal, S. Jaiswal, K. Santosh, N. Das, and M. Nasipuri, "SegFast-V2: Semantic image segmentation with less parameters in deep learning for autonomous driving," International Journal of Machine Learning and Cybernetics, vol. 10, pp. 3145-3154, 2019.

[18] M. Colley, B. Eder, J. O. Rixen, and E. Rukzio, "Effects of semantic segmentation visualization on trust, situation awareness, and cognitive load in highly automated vehicles," in Proceedings of the 2021 CHI conference on human factors in computing systems, 2021, pp. 1-11.

[19] F. Nesti, G. Rossolini, S. Nair, A. Biondi, and G. Buttazzo, "Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2280-2289.

[20] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," Neurocomputing, vol. 493, pp. 626-646, 2022.

[21] L. M. Dang et al., "Lightweight pixel-level semantic segmentation and analysis for sewer defects using deep learning," Construction and Building Materials, vol. 371, p. 130792, 2023.

[22] S. Rakshit, "Training Semantic Segmentation Models for Autonomous Vehicles (A Step-by-Step Guide)," 2022. [Online]. Available: https://wandb.ai/av-demo/CamVid/reports/Training-Semantic-Segmentation-Models-for-Autonomous-Vehicles-A-Step-by-Step-Guide---VmlldzoyNTMyMjc4

[23] M. N. S. Jahromi et al., "Privacy-constrained biometric system for non-cooperative users," Entropy, vol. 21, no. 11, p. 1033, 2019.

[24] A. Khattar and S. Quadri, "Generalization of convolutional network to domain adaptation network for classification of disaster images on twitter," Multimedia Tools and Applications, vol. 81, no. 21, pp. 30437-30464, 2022.

[25] J. S. Talahua, J. Buele, P. Calvopiña, and J. Varela-Aldás, "Facial recognition system for people with and without face mask in times of the covid-19 pandemic," Sustainability, vol. 13, no. 12, p. 6900, 2021.

Smart Cities, Smarter Roads: A Review of Leveraging Cutting-Edge Technologies for Intelligent Event Detection from Social Media

Ebtesam Ahmad Alomari[1], Rashid Mehmood[2]
Faculty of Computer Science and Information Technology, Albaha University, Albaha, Saudi Arabia[1]
Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia[2]

***Abstract*—The rapidly evolving landscape of smart cities and intelligent transportation systems makes the timely detection of traffic events a critical element for optimizing urban mobility. Furthermore, social media emerges as a valuable source of real-time information, with users acting as active sensors who spontaneously share observations and experiences related to traffic incidents. This review paper offers a comprehensive understanding of the state-of-the-art in traffic event detection from social media. The paper explores leveraging cutting-edge technologies including machine learning, and deep learning with big data technologies and high-performance computing. The discussion unfolds with an in-depth examination of the recent approaches for event detection followed by an exploration of the techniques of spatio-temporal information extraction and sentiment analysis, which are both considered fundamental aspects in enhancing the contextual understanding of traffic events. Further, the review explores the pivotal role of big data technologies in addressing scalability challenges inherent in the vast expanse of social data. The examination encompasses how big data frameworks facilitate efficient storage, processing, and analysis of large-scale social media datasets, thereby empowering machine learning and deep learning models for robust and real-time traffic event detection. Subsequently, the challenges and future directions have been highlighted. Addressing these challenges and leveraging advanced technologies, facilitates the proactive detection and management of these events, paving the way for smart mobility systems.**

***Keywords*—Mobility; smart cities; event detection; social media; big data analytics**

I. Introduction

The importance of detecting traffic events (incidents) cannot be overstated in the context of smart mobility and cities. Efficient traffic event detection lies in its impact on traffic flow optimization, safety enhancement, urban environment functionality and sustainability as well as the overall quality of life in smart cities. Therefore, rapid identification of incidents, such as accidents, road closures, or adverse weather conditions, helps authorities to make timely decisions, minimizing the negative impact of these events and contributing to the seamless operation of transportation systems.

Moreover, social media platforms become vast and decentralized sources of information, with their active users, who function as sensors spontaneously sharing observations, experiences, and updates related to topics in different domains including traffic. Subsequently, Twitter has widely been used to enable smart mobility systems, such as for traffic congestion estimation [1], passenger flow prediction in public metro transit systems [2], understanding taxi traffic dynamics [3], and detecting traffic anomalies caused by traffic accidents, disasters, etc. [4]. Besides, other social media have been used as well, for instance, public geotagged Instagram posts are used to detect an abnormal increase or decrease of the citizen's number in a specific area at a specific time by applying a density-based clustering algorithm [5]. Furthermore, several works have been focused on detecting events using social data, which raises the need for review papers that discuss the significant previous contributions and highlight future directions.

This review paper explores the detection of traffic events through analysis of social media data leveraging artificial intelligence with a specific focus on machine learning, and deep learning techniques in conjunction with big data technologies. The paper begins with examining the approaches of events detection and then investigates the techniques of Spatio-temporal information extraction and sentiment analysis. Furthermore, the review explores using big data technologies. The importance of utilizing big data technology in analyzing social data cannot be overstated, due to the volume, velocity, and variety of data generated on social media platforms, which require scalable and efficient processing frameworks. By leveraging advanced analytics and machine learning on large datasets, big data technology empowers the development of sophisticated models for traffic event detection, ensuring timely and accurate responses to dynamic urban challenges. To the best of our knowledge, only one review paper [6] provides a systematic review of traffic event detection techniques from social data but it does not includes the recent approaches such as using big data technology.

The main contribution of our work can be summarized as follows:

1) Cover the recent approaches including using deep learning and big data technologies as well as provide a taxonomy of the approaches.

2) Discuss the techniques of spatio-temporal information extraction and sentiment analysis, which are both considered

as fundamental aspects in enhancing the contextual understanding of traffic events.

3) Highlight the challenges and future directions to facilitate the proactive detection and management of the events and pave the way for smart mobility systems.

The rest of the paper is organized as follows. Section II reviews the works on traffic-related event detection from social data. Section III discusses using big data technologies in mobility and reviews the works related to traffic event detection using big data platforms and technologies. Section IV discusses the challenges and future directions. Finally, we draw our conclusions in Section V.

II. Social Media in Traffic Event Detection

With the exponential growth of social media platforms, users actively share real-time information, offering a valuable resource for monitoring and responding to traffic-related incidents. In this section, we review the existing work on road traffic analysis and event detection from social media. Fig. 1 depicts the number of reviewed papers in the period between 2012 and 2023. The total number of papers related to traffic event detection from social media included in this review is 61, only 15 of them are using big data technologies and platforms. Moreover, Fig. 2 shows the used social media and the number of reviewed papers that used big data technologies (light blue color) and that do not use them (blue color). It can be seen that most of the works have used Twitter while only one paper has used Instagram.

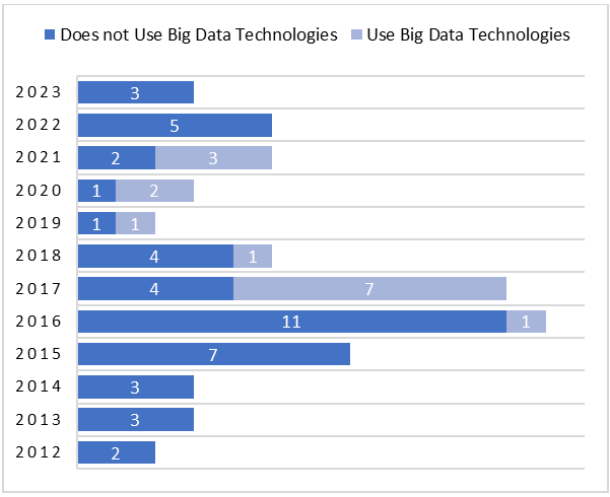


Fig. 1. Number of publications per year.

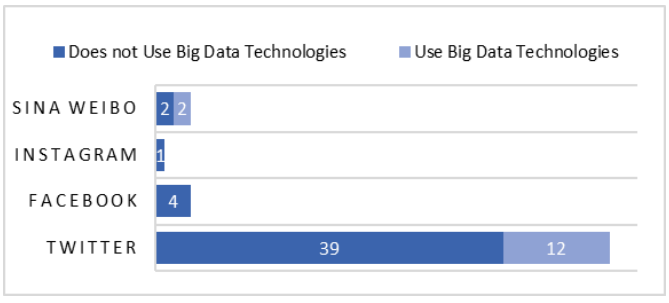


Fig. 2. Number of publications and social media platforms.

Fig. 3 shows the analysis dataset languages that have been used in the reviewed works, which include English, Arabic, Italian, Thai, Indonesian, Japanese, Malay, Korean, French, Spanish and Chinese. The blue color represents the number of papers that do not use big data technologies while the red represents the number of papers that use big data technologies. As depicted in the figure most of the works are focused on the English language.

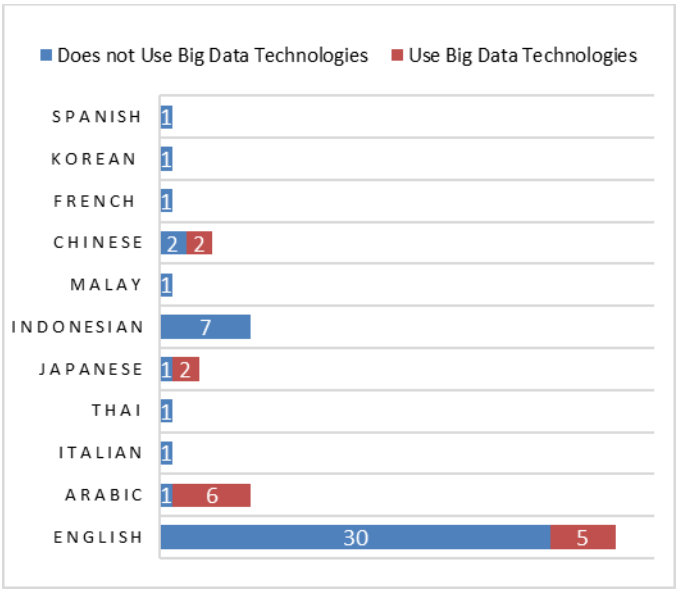


Fig. 3. Number of publications and dataset language.

Moreover, Fig. 4 describes the general workflow of event detection from social data. The main components are Data collection and filtering, Pre-processing, Event detection, Spatio-temporal information extraction, and Evaluation. Firstly, data are collected by using keywords, hashtags, accounts, geo-coordinates or a combination of these approaches. Then, the data are filtered to keep only traffic-related data. Data filtering can be done during data collection or it can be a separate step and various approaches can be used including machine learning algorithms. The next step is data pre-processing. There are common sub-components for pre-processing, which are tokenizer, normalizer, stop-words removal, and stemmer. Several tools and packages are available for pre-processing but mostly they are designed for a specific language. The next step is detecting the events and then extracting the time and location information. Subsequently, sentiment analysis is applied to understand the feelings and emotions regarding the detected events. This step is not mandatory. It depends on the interest of the researcher and the aim of the work. Finally, the tool is evaluated. The evaluation process depends on the method that has been used. For instance, if machine learning algorithms have been used to build classifiers for event detection, there are common evaluation metrics that can be used for evaluation such as accuracy, precision and recall. These are the main common steps that have been followed in literature to detect traffic events, more details about the components can be found in [7].

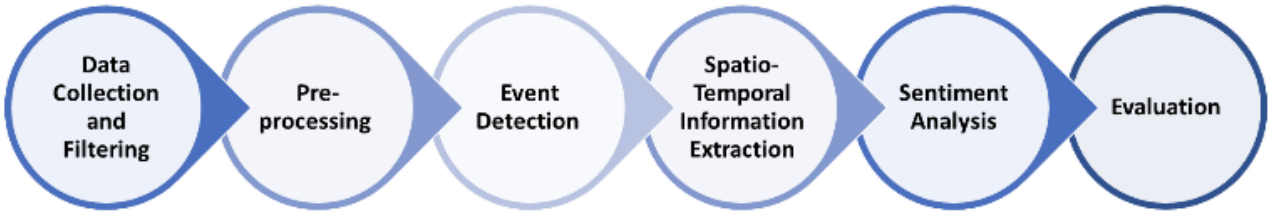


Fig. 4. General workflow of social data analysis for Traffic event detection.

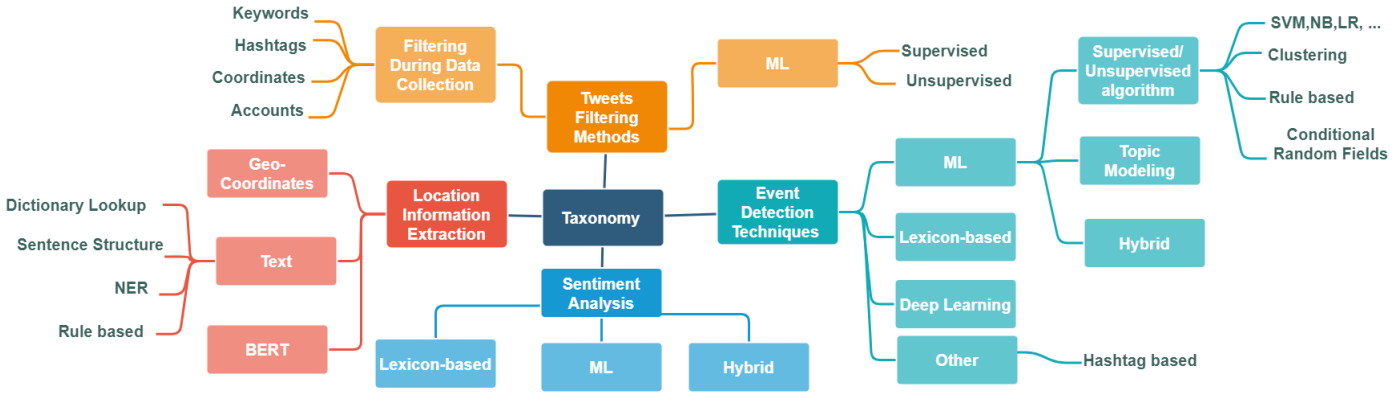


Fig. 5. Taxonomy of traffic events detection techniques from social data.

Fig. 5 shows the taxonomy of the approaches for each component in the general workflow except for the pre-processing component, which mainly depends on the language of the analyzed data and the evaluation component, which depends on the approach that has been used for the event detection. In the next subsection, we review the existing works and organize them based on the approaches that have been used. We started in section A with the event detection approaches because it is the main and the most important component. At the end of section A, we summarized the techniques for data filtering. In section B, we discuss the techniques of Spatio-temporal information extraction. Finally, section C covers the techniques of sentiment analysis

A. Event Detection Techniques

Several techniques have been used for traffic event detection including lexicon-based, machine learning (ML) and deep learning techniques. The next subsections provide a review of the existing works for each technique.

1) Lexicon-based: Istiq et al. [8] built an application to monitor road conditions. They collected road condition information from RSS fed on Facebook and stored them in the MySQL database. After Text Mining, the application categorized the information based on how much connectivity between words by categories. The information is categorized into six types including floods, traffic jams, congested roads, road damage, accidents, and landslides. Daly et al. [9] developed Dub-STAR system to extract the causes of traffic conditions from real-time tweets. They applied a simple dictionary approach to assigning tags to the input messages. The text is classified into classes such as delay, incident, roadwork, or concert. Moreover, Alomari and Mehmood [10] analyzed Arabic tweets related to traffic congestion in Jeddah city. They created custom dictionaries in SAP HANA for the common Arabic keywords about transportation and traffic congestion.

Alkouz and Alghbari [11] proposed a tool called SNSJam for traffic event detection using Arabic and English posts from Twitter and Instagram. They collected 50 million posts but after filtering they got around 2.3 million tweets and 4k Instagram posts. To identify events, they used a list of keywords.

2) Machine learning (ML)

a) Supervised and Unsupervised ML: Researchers in [12] and [13] used the SVM algorithm to classify tweets as relevant and irrelevant to traffic. But they don't detect event types. Dhavase and Bagade [14] provided a classification technique based on a machine-learning algorithm to detect crime-disaster events. Agarwal et al. [15] detected complaints tweets about road irregularities and bad road conditions by employing a rule-based classifier. They extracted the important information; such as the problem and the location, and then the tweets were categorized into useful, nearly-useful and irrelevant complaint reports.

Furthermore, Sakaki et al. [16] detected weather information and heavy-traffic information by classifying tweets into positive class means event-related and negative class means not related to events using the SVM algorithm. Klaithin and Haruechaiyasak [17] applied a machine learning classifier based on Naive Bayes Model. They classified the Thai tweets about traffic into three types: tweets about location, tweets about roads and tweets about traffic status. Kumar et al. [18] built a model to detect road hazards so they

classified the tweets into two classes namely, having negative or non-negative sentiment. They considered that all negative sentiment tweets have road hazard information. They applied three algorithms, which are Naïve Bayes (NB), K-nearest neighbors (KNN), and the Dynamic Language Model (DLM).

Moreover, D'Andrea et al. [19] applied text mining techniques and classified real-time Italian tweets into three classes, which are "traffic due to an external event", "traffic congestion or crash", and "non-traffic". They validated the detection of temporal information by their system using a generated dataset for traffic events from local newspapers and official news websites.

Kurniawan et al. [20] built a machine learning classifier to classify real-time Indonesian tweets into traffic-related and not-related. They used NB, SVM and decision tree (DT) algorithms. They counted the occurrences of all the words in a tweet as features. Then, in the feature selection process, they used only 40 selected words. Nguyen et al. [21] developed a system called TrafficWatch to monitor traffic conditions in Australia using real-time and historical tweets. During the feature selection process, they used serval features including a bag of words, Pattern recognition, Lemma, part-of-speech and a bag of tags. They introduced the NER annotation schema for detecting traffic related entities. Additionally, they trained the model based on Conditional Random Fields (CRFs). The traffic-related entities are implemented as key features to train the ML classification model. Further, they implemented an online clustering algorithm to incrementally cluster the streaming tweets.

Yazici et al. [22] used TF-IDF and NB to detect events from personal and organizational Twitter accounts. They noticed that the tweets from personal accounts do not disseminate traffic incident information in a very structured manner in terms of grammar and spelling compared with the organizational account. On the other side, personal tweets are better in reporting events that have just happened. Besides, Semwal *et al.* [23] and Tejaswin *et al.* [24] predicted the traffic incidents from social media using a random forest classifier. Anantharam et al. [25] developed a framework for extracting city-related events. They created a training set for building a conditional random field (CRF) automatically, by using dictionary-based spotting of event and location terms, to reduce manual tagging effort.

Moreover, Langa and Moeti [26] filtered real-time tweets about congestion in South Africa into traffic-congested-route and non-traffic-congested tweets using the Naïve Bayes algorithm. Subsequently, they developed a mobile application to send notifications to users based on classified tweets about traffic. Suat-Rojas et al. [27] built a classifier to detect accidents from tweets in the Spanish language. They classified tweets into accident and non-accident. For validation, they compared the detected accidents with official data from the Bogota Mobility Secretariat. Subsequently, they found that using doc2vec for feature extraction and SVM for classification helped in achieving good accuracy results.

3) Topic Modeling: One of the popular topic modeling algorithms is Latent Dirichlet Allocation (LDA). It was introduced by Blei et al. [28]. Besides, it is a statistical classification model based on the word-topic frequency distribution. Wang et al. [29] focused on exploring the relationship between traffic conditions in daily matter and urban human events in Toronto, Canada. They developed a tweet-LDA engine to classify tweets into two classes namely, "traffic-relevant" and "traffic-irrelevant". To ensure that no relevant tweets are missed, they filter out traffic-irrelevant tweets using a set of keywords. Moreover, Huang et al. [30] used DBSCAN to build spatiotemporal clusters from the geotagged tweets. Then, the LDA was implemented to the tweets under each cluster to extract the topics. Ali et al. [31] applied OLDA-based traffic event labeling. They used a well-known Web Ontology Language (OWL) tool called Prot´eg´e to develop ontology before using LDA. They generated a list of the most frequent words related to traffic events and then used them to develop ontology-based semantic knowledge.

a) Hybrid (Topic Modeling and Supervised/ Unsupervised ML): Gu et al. [32] collected historical and real-time tweets about traffic in Pittsburgh and Philadelphia Metropolitan. They used Supervised Latent Dirichlet Allocation (sLDA). In addition, the Tweets are classified by a Semi-Naive-Bayes (SNB) classifier. Lau [33] used Latent Dirichlet Allocation (LDA) topic modeling to filter traffic messages in the Chinese language. Additionally, they built SVM, KNN and NB classifiers for traffic events detection.

4) Deep Learning: Chen et al. [34] extracted Chinese traffic information from the Sina Weibo platform. They employed deep neural networks to learn the abstract features and classify them into traffic relevant and irrelevant. Ji et al.[35] measured the similarity of events between texts using meta-path in heterogeneous information networks. Then, to get the best possible meta-path weights, they employed a graph neural network for semi-supervised learning. After that, they designed a clustering algorithm to identify the traffic event categories.

Moreover, Ambastha and Desarkar [36] used different ML algorithms including SVM, Naive Bayes and Random Forest as well as deep learning algorithms including LSTM, CNN, and Universal Language Model FIne-Tuning (ULMFIT). The developed classification models were used to classify the tweet into two categories, namely, "Traffic incident related" or "Non-Traffic incident related". The Transfer Learning approach using ULMFIT was employed to enhance the performance.

Kim et al. [37] classified tweets in the Korean language using NB, RF, SVC, linear SVC, BiLSTM, and TextCNN into six classes: construction, weather, accident, traffic jam, crowded event and others. Rifqi et al. [38] used CNN+Word2Vec, CNN+FastText, and SVM to classify Indonesian tweets into 2 classes, which are tweets about traffic jams and tweets about smooth traffic. Swapnika and Vasumathi [39] applied the DNN and Harris Hawk optimization (HHO) algorithm to detect the following events from Twitter: education, transportation, environment, geospatial and water events. They claimed that they addressed

the challenge of the scalability challenge, but they did not provide a detailed explanation.

Furthermore, Hodorog et al. [42] combined AWD-LSTM and ULMFiT to detect traffic-related events in Cardiff City in the UK. They focused on several events including congestion, accidents, social gatherings, thefts, bus queues, floods and electricity charges. Additionally, they studied the relationship between the events as well as assigned a citizen satisfaction value to each of them. Mehri et al. [43] used BERT to detect subway-related events from tweets in English and French Language. They manually labeled 10381 records in English and 11008 in French to build the training dataset. The finding indicated that BERT in zero-shot surpasses the performance of the baseline models.

TABLE I. EVENT DETECTION TECHNIQUES

Ref.	Detected Event types	Event Detection Technique
Lexicon-based		
[8]	Floods, traffic jams, road damage, accidents, landslides	Based on the connectivity between words
[9]	Traffic congestion causes	Dictionary-based
[10]	Accidents, weather, road works, social events.	Dictionary-based
[11]	Accidents	Dictionary-based
ML		
[14]	Crime and disaster	NB
[15]	Useful, nearly-useful, irrelevant complaint reports	Rule-based classifier
[16]	Related to roads, related to weather	SVM
[17]	Accident, announcement, question, request, sentiment.	NB
[18]	Hazard and non-hazard	NB, KNN, and DLM
[19]	Congestion, crash, non-traffic event, external event.	SVM, NB, C4.5 Decision tree, PART, KNN,
[20]	Traffic and non-traffic	NB, SVM, DT
[21]	Roadwork, queue, accident, activities, breakdown, police.	Conditional Random Fields (CRFs) labeling method.
[22]	Traffic-relevant, traffic-irrelevant	NB, dictionary of frequently occurring words
[23]	Heavy-vehicle, traffic-jam, park-footpath autometer, wrong-side, breakdown, jump-signal, U-turn, no-parking.	Random forest classifier
[24]	Weather	Random forest classifier
[40]	Roadwork traffic jam, freight traffic, road closure, weather, accident	Dictionary, clustering algorithm
[41]	Earthquakes, forest fires, floods, and droughts	Checking a set of predefined weighed keywords, KNN algorithm
[25]	City-related events	CRF model, dictionary-based spotting
[26]	Traffic-congested and non-traffic-congested	NB
[27]	Accident and non-accident	SVM, NB, RF and Neural Networks
[29]	Traffic-relevant, traffic-irrelevant	LDA
[30]	Leisure, sports, music, movies, art, and other	LDA
[31]	Traffic, non-traffic	OLDA
[32]	Roadwork accidents, weather, special events, obstacle vehicles.	sLDA, SNB
[33]	Accidents, traffic jams, weather	LDA, SVM, NB, K-Nearest
Deep Learning		
[34]	Traffic relevant, Traffic-irrelevant	CNN
[35]	Traffic Control, Weather Anomaly, Accident, Congestion, Road Construction, Vehicle Anchorage, Official Occupation, and Normal Traffic.	BS-GCN
[36]	Traffic relevant, Traffic-irrelevant	ULMFiT model
[37]	Construction, weather, accidents, traffic jams, crowded events and others.	NB, RF, SVC, linear SVC, BiLSTM, and TextCNN
[38]	Traffic jams, smooth traffic	CNN+Word2Vec, CNN+FastText, and SVM
[39]	Education, transportation, environment, geospatial and water event	DNN and HHO
[42]	Congestion, accidents, social gatherings, thefts, bus queues, floods, and electricity charges.	AWD-LSTM and ULMFiT
[43]	Incident and non-incident	BERT

5) Other Methods: He et al. [44] proposed MetroScope system which analyzed real-time tweets to detect events related to the Washington D.C. Metro system. They developed a phrase-level algorithm that groups events with similar key phrases into a story. To prioritize urgent events, they performed sentiment analysis and then implemented a function to automatically send emails to authorities regarding emergency events.

Shekhar et al. [45] collected data about traffic conditions from Facebook and Twitter. They constructed a decision tree to display traffic-sensitive optimized routes. After that, they categorized particular streets within a city into three categories, which are Moderate Congestion, Severe Congestion, and No Congestion based on the average user's sentiment on hourly time slots. Further, they detected the possible causes of traffic congestion in a particular area and enabled users to search for the cause of congestion at a particular time. Ni et al. [2] developed a hashtag-based event detection algorithm. To detect events, they examined tweets within a specific area and probe (the subway station and two stadiums) instead of detecting the exact topic of the events.

Table I shows the type of detected events by each of the reviewed papers as well as the techniques that have been used for event detection. Table II illustrates the filtering approaches that have been used to filter out irrelevant posts before detecting the events.

TABLE II. FILTERING TECHNIQUES

Tweets Filtering Techniques		References
During Data Collection	By Keywords	[8], [16], [18], [19], [41], [22], [30],
	By accounts	[9], [10], [14], [15], [17]
	By Geo-coordinates	[2], [25], [42]
Using Machine Learning		[46], [32], [33], [40], [29], [20], [21], [34], [12], [13]

TABLE III. LOCATION INFORMATION EXTRACTION TECHNIQUES

Technique		References
Geo-Coordinates		[1], [29], [16], [47], [18], [48], [33], [49]
Text Analysis	Dictionary lookup	[50], [17], [1], [9], [12], [16]
	Sentence structure	[13], [29], [12], [45], [51], [24]
	Rule-based	[17], [13], [14]
	NER	[52], [53], [15], [40], [33], [27] [46], [54], [27]
Deep Learning	BERT	[37]

B. Spatio-temporal Information Extraction

Chaniotakis et al. [50] analyzed tweets about the flood and the evacuation in Oroville, California USA. They used the WordNet dictionary to create a corpus to detect discussions concerning the evacuation. Muhammad and Khodra [53] used the conditional random field (CRF) model for event information extracted from Indonesian tweets. They filtered the event-related tweets by combining the rule-based method with the bag of words model. Kumar *et al.* [18] extracted coordinates from geo-tagged tweets. Then, the geographic coordinates are mapped to a specific road or road segment**.** Shekhar *et al.* [45] extracted location names from the text. They assumed that the location is almost preceded by a preposition. So, they created a list of all the prepositions (e.g. in, at, on, etc.). The extracted location name is sent to Google Maps API to return the geographical coordinates. Wang et al. [29] extracted the coordinates from geo-tagged tweets and mapped them to extract location in terms of road, street, and landmark. On the other side, the location information from non-geotagged tweets is extracted either from users' profiles or from tweet content by using semantic analysis to identify the key joint words of "between...and", "from...to" and "exit to...,".

Furthermore, Wang *et al.* [1] extracted the streets, landmarks, and direction information from the text by using a gazetteer. Additionally, for traffic estimation and prediction, they extracted two types of road features, which are physical features (such as the road segment length, the number of lanes and the number of intersections) and point of interest (POI) (such as schools, hospitals, shopping mall, etc.). Sakaki et al. [16] extracted driving information from Japanese tweets and transformed geographically related terms into geographical coordinates. They created a dictionary for the place names in Japan. In addition, they collected pairs of verbs and prepositions, which are dependent on the names of places. Then, they used such pairs to extract the names of places.

Moreover, Daly et al. [9] perform a dictionary lookup to extract location names from SMS messages or tweets. Alifi and Supangkat [12] classified the tweets by using SVM to distinguish between the data that are related and not related to the traffic condition. They suggested methods for extracting location information which is using location vocabulary, using the symbol "-" or based on the structure and words in a sentence. In addition, they obtained useful information from real-time streams involving congestion causes, traffic conditions, as well as weather conditions. They suggested using three approaches: (i) the existence of location words (such as from, to, toward), (ii) the use of the symbol "-" that usually used to link two location points at once, (iii) the location vocabulary (such as street name). Klaithin and Haruechaiyasak [17] extract words or phrases related to traffic information from Thai tweets using lexicon-based and rule-based methods. They extracted traffic information such as road names, locations, and traffic accidents. Hanifah et al. [13] applied a rule-based approach and obtained information regarding the time and date, location and image. Additionally, they employed SVM model for traffic congestion detection from tweets posted in Bandung, Indonesia. The developed model filters the tweets into relevant to traffic congestion and irrelevant.

Dhavase and Bagade [14] proposed an approach to extract location from tweets. They used three different parsers: (i) named location parser (e.g., use gazetteer matching) to check for locations in tweets, (ii) NER parser (like Stanford NER). (iii) street building parser using rule-based pattern matching.

Hasby and Khodra [52] developed an information extraction module to extract time, location, condition, direction, and causes from Indonesian tweets about traffic jam.

The module consists of five elements, which are tokenizer, normalizer, Named Entity Recognition (NER), template element task, information filling and relation extraction. Similarly, Muhammad and Khodra [53] extracted event name, location, time and additional event information using an extractor module that was built up by Tokenizer, NER and POS Taggercomponent.

Moreover, Agarwal et al. [15] applied a combination of NER (Indico Text Analysis API) and GeoCoding (OpenStreetMap API) APIs to obtain the geographical entities from tweets. Gutierrez et al. [40] extracted locations from tweet messages using four NER engines, which are: Alchemy, Stanford NER, OpenCalais and NERD. Further, they used three geolocation external applications, which are: GeoNames, Google Geocoding, and Nominatim. Raymond [33] and Musaev et al. [54] applied the open-source Stanford NER tool to extract the place name. Salas et al. [46] used NER to link a concept to a unique location through a knowledge base such as Wikipedia.

Subsequently, Xu *et al.* [47] proposed a model based on crowdsourcing for describing urban emergency incidents such as storms, fires or traffic jams. They proposed a 5W model for illustrating the data, which provides five basic elements 1) When: temporal information (e.g. the starting/ending time of the event), 2) Where: spatial information (places), 3) What: the semantic information for the event, 4) Who: personal information (e.g., participatory or witness) and 5) Why: the reason information. They extracted location information from the check-in information from Weibo.

Besides, Berlingerio *et al.* [48] developed a system named SaferCity based on a new spatiotemporal clustering algorithm for incident detection from Twitter. Singh [55] extracted location from tweet text. To address the issue of the lack of location information, they suggested using the historical locations of the user to predict the probable location by applying Markov chain model. Yang et al. [51] extracted information form tweets related to the traffic conditions in Malaysia. They suggested extracting the location and direction information from the text using prepositions and words like "from…to", "along", "heading" and etc. Tejaswin et al. [24] extracted location entities using a regular expression parser. After that, entity disambiguation is applied to verify if it is a location and ensure that the address belongs to the correct city. Kim et al. [37] built an algorithm for region extraction to extract keywords from Korean tweets with the help of entity name recognition API based on BERT.

Table III summarizes the techniques for extracting location information. We grouped the approaches for extracting the location information into three main groups, which are: (i) from the coordinates attribute in the geotagged posts (ii) by extracting the location name from the text, and iii) using deep learning models such as BERT. The first approach is not always applicable since not all posts are geotagged because some users turn off location services on their smartphones to protect their privacy. For the second approach, the text will be analyzed using natural language processing (NLP) methods to extract the place name. The common methods that are applied to extract a placename from the text are as follows: i) Dictionary lookup: requires checking the text to discover place names listed in a gazetteer or glossary. ii) Sentence structure: use a list of prepositions (e.g., in, at, on, etc.) iii) Named Entity Recognition (NER): identify and categorize entities from text. iv) Rule-based pattern matching: implement the extraction based on certain written actions.

C. Sentiment Analysis

1) Lexicon-Based: Shekhar et al. in [45] and [41] categorized the users' emotions during a disaster by feeding English text from social media to sentiment analysis method. The users' emotions are sub-categorized as positive, negative, unhappy, depressed and angry. Additionally, they created a dictionary of weighted sentiment ratings for words and used SentiStrength online. SentiStrength [56] is a popular stand-alone online sentiment analysis tool. It uses a dictionary of sentiment words for assigning scores to negative and positive phrases in the text. Salas et al. [46] applied sentiment analysis to classify the tweets into positive, negative, or neutral class. They used TensiStrength for stress and relaxation strength detection.

2) Machine learning: A different sentiment classification method was applied by Kumar et al. [18] to categorize tweets into four sentiment classes: false negative, true negative, false positive and true positive. A true positive indicates that the tweet is accurately categorized as non-hazard whereas a true negative indicates that a tweet is accurately classified as a hazard. The false positive category refers to tweets that include some positive sentiment terms e.g. "awesome" and "enjoy", however, the actual sentiment is negative. On the other side, false negative refers to tweets that are incorrectly categorized as a hazard. They employed three ML algorithms, which are KNN, Naïve Bayes, and DLM.

Ohbe et al. [57] classify Japanese tweets about the local event into three categories: positive, negative, and other. They used a multinomial logistic regression analysis for the classifier. Berlingerio *et al.* [48] used Sentiment140 API for sentiment analysis. Sentiment140 [58] used a trained classifier built on large tweets with emoticons for distant supervised learning. Furthermore, Musaev et al. [54] developed a model to categorize tweets into happy or sad. They applied the Continuous Bag-of-Words and Skip-gram model. To do automatic labeling, they searched for tweets that contain ":-)" and ":-(" emoticons. After that, they utilized the Word2Vec repository to convert the tweets in the training set to their vector representations.

3) Hybrid: Adetiloye and Awasthi [59] applied sentiment analysis for traffic tweets using the lexicon of opinion words (LOWs) and the improved Naïve Bayes algorithms in [60]. Table IV summarizes the techniques for sentiment analysis that have been used in literature for event detection.

D. Big Data Tools and Platforms

The term big data refers to the extremely large amount of data that grows exponentially with time, which makes it difficult to conduct an efficient analysis using conventional IT and hardware solutions within a reasonable amount of time

[61]. Subsequently, the four main characteristics of big data are i) Volume: refers to the size of data that might be measured by Zettabytes (ZB), or Yottabytes (YB) ii)Velocity: which refers to the processing speed and considered as crucial characteristic for the performance. iii) Variety: refers to the diversity of the data, which can be structured, unstructured, or semi-structured. iv) Value: refers to valuable and reliable data.

TABLE IV. SENTIMENT ANALYSIS APPROACHES FOR EVENT DETECTION

Ref.	Lexicon	ML	Hybrid	Categories
[45], [41]	✓			Positive, Negative, Unhappy, Depressed, Angry
[18]		✓		True Negative, False Negative, True Positive, False Positive.
[46]	✓			Positive, Negative, Natural
[57]		✓		Positive, Negative, and other
[48]		✓		Positive, Negative, Natural
[59]			✓	Positive, Negative, Natural
[54]		✓		Sad, Happy

III. BIG DATA TECHNOLOGIES IN MOBILITY

Furthermore, the traditional data storage, processing, and analysis applications are insufficient to address the challenges that come from the massive continuously generated transportation and traffic-related data from various sources such as sensors, digital cameras, and social media. This raises the need for big data platforms and technologies, based on distributed data management and parallel processing. Big data storage solutions, such as NoSQL databases are ideal solutions for the storage issue since they have more flexible and adaptable data models and schemas compared to relational databases. Subsequently, big data platforms have integrated libraries including machine learning, deep learning, or data mining algorithms, which facilitate smart analysis [62].

The next subsection explains the existing approaches for traffic event detection using big data.

A. Traffic Events Detections Using Big Data Technologies

Nguyen and Jung [63] detected events by applying density-based spatial clustering. Additionally, to evaluate the proposed method, they used datasets (about 'FA Cup'' and ''Super Tuesday) employed in a previous study. They evaluated the performance using Hadoop. Khazaei *et al.* [64] proposed a big data analytics platform, named Sipresk that was built over Apache Spark to detect traffic events from different sources including social media, cameras, mobile devices, etc. Lau [33] suggested using topic model-based for text filtering then they built a classifier to identify traffic events such as traffic jams, road accidents, weather, etc. They fed the message corpus to the proposed Latent Dirichlet Allocation (LDA) model for topic learning. Subsequently, they applied the probabilistic language model to estimate the generation probabilities of the labeled message based on a mined unlabeled topic. Further, they implemented ML classifier using Spark Machine Learning (MLib) library.

Salas et al. [46] fetched real-time tweets through Kafka and Flume and stored them in HBase storage. They employed Spark machine learning library to build SVM classifier and filter the tweets into traffic or non-traffic-related tweets. The tweets are processed using Natural Language Processing (NLP) methods before they are passed to the trained classifier.

Suma *et al.* [49] used Apache Spark for spatio-temporal events detection in London City. For the data pool, they utilized the power of the Fujitsu Exabyte File System (FEFS). Further, they installed both FEFS and spark technologies on top of the HPC cluster. Pandhare and Medha [65] analyzed tweets related to traffic and accidents to detect road traffic events using Spark. They used a regular expression filter to separate unnecessary information. Then, they applied some text mining steps including tokenization, creating term frequency vectors by using HashingTF and TF-IDF to reflect the importance of a token in a document. After that, they classified the tweets by employing Logistic regression and SVM algorithms. Kousiouris *et al.* [66] identified large crowd concentration events that might affect the user journey. They used Spark, Apache AVRO and Cloud-based solutions (OpenStack Swift).

Moreover, Alomari and Mehmood [10] developed a lexicon-based approach to filter Arabic tweets about traffic congestion in Jeddah city using SAP HANA. Then, they extended their word and performed sentiment analysis [67]. Subsequently, they developed multiple big data pipelines and architectures for social text event detection using cutting-edge technologies consisting of machine learning algorithms and high-performance computing as well as Apache Spark. Furthermore, they proposed supervised [68], [7] and unsupervised [69] machine learning methods to enable smarter transportation by detecting events using social data in the Arabic language. Subsequently, they detected several events including congestion, roadwork, fire, social events, weather, government measures, and public concern. Additionally, to improve the performance of detecting events from Saudi dialectical Arabic text, they proposed a pre-processing pipeline that includes a tokenizer, irrelevant characters removal, normalizer, stop words removal, as well as an Arabic light stemmer. Also, they built a tool for spatio-temporal clustering and visualization. Furthermore, they proposed methods for validating the detected events through internal sources (Twitter data) and external sources (e.g. official newspaper websites). Moreover, to address the challenges that come from manual labeling of large datasets, they proposed an automatic labeling method [70] using predefined dictionaries for detecting events.

Chen et al. [71] proposed a semi-supervised deep-learning model for detecting traffic events. They built a multi-model feature learning architecture to transform data from sensor time series and Twitter posts into a unified multi-modal feature representation. They built two encoders: the first one to extract features from sensor data using the Recurrent Neural Network (RNN) while the second encoder is designed for social data.

IV. DISCUSSION, CHALLENGES AND FUTURE DIRECTIONS

We illustrated in Section II the general workflow for event detection. In this section, we discuss the challenges for each step in the workflow.

Firstly, for data collection, the existing API such as Twitter API[1] limits the number of collected data for free. Additionally, even though they have paid APIs with fewer restrictions on the number of fetched tweets, it is too expensive. Secondly, for data pre-processing, although there are several tools and packages, most of them are built for the English language. Thus, there is a need for more efficient tools for other languages. Subsequently, there is a need to improve pre-processing and NLP methods to work on the dialectical short text. Thirdly, posts on social media are usually short and thus may not include all the important information about the events. Therefore, finding the exact location or time of occurrence might be difficult. The information might either not exist because users disable the location service for privacy reasons, or it exists but does not reflect the time or the place where the event occurred since people can post about events in other cities and countries. Additionally, in some cases, the post carries more than one location or time information. For instance, the information attached to the posts such as the geo-coordinates and the information users wrote in the text. Thus, the existing approaches for Spatio-temporal information extraction need improvement to consider the different scenarios and extract or predict the right information. Subsequently, recent approaches need to be used including deep learning and Large Language Model (LLM) such as the work in [37].

Moreover, several approaches have been used for traffic event detection including the lexicon-based approach, ML algorithms and deep learning. We divided the works that used ML into four categories. The first category includes works that used the common supervised or unsupervised algorithms such as SVM, Naïve based, clustering, etc. The second category includes the works that used topic modeling algorithms such as LDA. The third category contains the works that applied deep learning. The last category includes the works that used other methods such as hashtags-based techniques. Each approach has it is own challenges. One of the major challenges of using supervised classification is labeling the data for training the model. Manual labeling takes significant time and effort, especially for big data.

Furthermore, supervised classification requires defining the classes and then building and training the models, which means that we need to specify the types of events that we want to detect in advance. Thus, supervised classification is not appropriate if we do not want to limit the detected event types. The other challenge in supervised classification is having an imbalanced dataset where the number of posts in the training dataset for each class label is not balanced. On the other side, topic modeling has its challenges. One of the challenges is understanding the topic and finding the category that belongs to it. Additionally, we need to test different parameters to find the best number of topics and iterations, which is a difficult process and takes a long running time, especially for big data (for more details see [69]). Furthermore, the works in literature that used deep learning for traffic event detection are very limited. Additionally, more work is required to study the feelings and emotions regarding the detected events, which could help the authorities and decision-makers understand the situation and get more involved in addressing the difficulties.

Finally, detecting events from big social data is difficult due to its daunting characteristics -- volume, variety, velocity and veracity. The state-of-the-art on using big data technology for traffic event detection from social media is limited. Therefore, many more works are needed to improve the breadth and depth of the studies in this area regarding the size and diversity of the data, as well as the applicability, accuracy, performance, and scalability of the analysis and detection methods.

V. CONCLUSION

This paper comprehensively reviews recent advancements in the fusion of AI with a specific focus on machine learning, and deep learning techniques in conjunction with big data technologies for traffic event detection from social media. Furthermore, we presented the general workflow, which includes the following steps: Data collection and filtering, Pre-processing, Event detection, Spatio-temporal information extraction, and Evaluation. Before detecting events, data are filtered to keep only traffic-related posts. We found that this process is done either during the data collection phase or after through machine learning algorithms. After that, we divided the techniques for event detection into categories and then discussed the works based on the applied technique. The first technique for event detection is lexicon-based. The second technique is using machine learning including, supervised and unsupervised algorithms, topic modeling, or hybrid. The third technique is using deep learning. The last category includes other techniques such as hashtags-based techniques.

Moreover, we grouped the techniques for extracting location information into two main groups, which are using geo-coordinates attributes in the geotagged posts and extracting location names by text analysis. The first approach is not always applicable since not all posts are geotagged. For the second approach, several methods are applied to extract location from the text including using NER, Dictionary lookup, Rule-based pattern matching, deep learning models and sentence structure by using a list of prepositions.

Furthermore, we classified the existing approaches for sentiment analysis for traffic-related events into three categories, which are lexicon-based, using machine learning, and hybrid approaches. Additionally, we reviewed the works that used big data technology for traffic event detection from social media.

However, the state of the art in this area that uses deep learning, LLM or big data technologies is limited, and thus many more works are needed to improve the breadth and depth of the studies since using them helps to improve the efficiency, scalability, performance, flexibility as well as support multilingual. Subsequently, big data technologies and platforms are very important in this domain due to the characteristics -- volume, variety, velocity, and veracity of the social data. In conclusion, this review consolidates the current state of research, offering a valuable resource for researchers, practitioners, and policymakers seeking to leverage cutting-

[1] https://developer.twitter.com/en/docs/twitter-api

edge technologies for enhancing urban mobility and smart cities.

REFERENCES

[1] S. Wang, L. He, L. Stenneth, P. S. Yu, and Z. Li, "Citywide traffic congestion estimation with social media," in Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '15, 2015, pp. 1–10, doi: 10.1145/2820783.2820829.

[2] M. Ni, Q. He, and J. Gao, "Forecasting the Subway Passenger Flow under Event Occurrences with Social Media," IEEE Trans. Intell. Transp. Syst., vol. 18, no. 6, pp. 1623–1632, 2017, doi: 10.1109/TITS.2016.2611644.

[3] F. Wu, H. Wang, and Z. Li, "Interpreting traffic dynamics using ubiquitous urban data," Proc. 24th ACM SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst. - GIS '16, pp. 1–4, 2016, doi: 10.1145/2996913.2996962.

[4] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," Acm Sigspatial, pp. 334–343, 2013, doi: 10.1145/2525314.2525343.

[5] D. R. Domínguez, R. P. Díaz Redondo, A. F. Vilas, and M. Ben Khalifa, "Sensing the city with Instagram: Clustering geolocated data for outlier detection," Expert Syst. Appl., vol. 78, pp. 319–333, 2017, doi: 10.1016/j.eswa.2017.02.018.

[6] S. Xu, S. Li, and R. Wen, "Sensing and detecting traffic events using geosocial media data: A review," Comput. Environ. Urban Syst., no. June, 2018, doi: 10.1016/j.compenvurbsys.2018.06.006.

[7] E. Alomari, I. Katib, and R. Mehmood, "Iktishaf: A Big Data Road-Traffic Event Detection Tool Using Twitter and Spark Machine Learning," Mob. Networks Appl., pp. 1–16, 2020.

[8] A. F. Istiq Septiana, Setiowati, Yuliana, Arna Fariza, "Road Condition Monitoring Application Based on Social Media With Text Mining System," Int. Electron. Symp. Road, pp. 148–153, 2016.

[9] E. M. Daly, F. Lecue, and V. Bicer, "Westland row why so slow?Fusing Social Media and Linked Data Sources for Understanding Real-Time Traffic Conditions," Proc. 2013 Int. Conf. Intell. user interfaces - IUI '13, no. March, p. 203, 2013, doi: 10.1145/2449396.2449423.

[10] E. Alomari and R. Mehmood, "Analysis of Tweets in Arabic Language for Detection of Road Traffic Conditions," in in Proceedings of the First EAI Conference on Smart Societies, Infrastructure, Technologies and Applications (SCITA 2017), Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (LNICST), pp. 98–110.

[11] B. Alkouz and Z. Al Aghbari, "SNSJam: Road traffic analysis and prediction by fusing data from multiple social networks," Inf. Process. Manag., vol. 57, no. 1, p. 102139, 2020, doi: 10.1016/j.ipm.2019.102139.

[12] M. R. Alifi and S. H. Supangkat, "Information Extraction for Traffic Congestion in Social Network," in International Conference on ICT For Smart Society, 2016, no. July, pp. 20–21.

[13] R. Hanifah, S. H. Supangkat, and A. Purwarianti, "Twitter information extraction for smart city," Proc. - 2014 Int. Conf. ICT Smart Soc. "Smart Syst. Platf. Dev. City Soc. GoeSmart 2014", ICISS 2014, pp. 295–299, 2014, doi: 10.1109/ICTSS.2014.7013190.

[14] N. Dhavase and A. M. Bagade, "Location identification for crime & disaster events by geoparsing Twitter," 2014 Int. Conf. Converg. Technol. I2CT 2014, pp. 2–4, 2014, doi: 10.1109/I2CT.2014.7092336.

[15] S. Agarwal, N. Mittal, and A. Sureka, "Potholes and Bad Road Conditions- Mining Twitter to Extract Information on Killer Roads," ACM India Jt. Int. Conf. Data Sci. Manag. Data CoDS-COMAD 2018, 2018.

[16] T. Sakaki, Y. Matsuo, T. Yanagihara, N. P. Chandrasiri, and K. Nawa, "Real-time event extraction for driving information from social sensors," in Proceedings - 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems, CYBER 2012, 2012, pp. 221–226, doi: 10.1109/CYBER.2012.6392557.

[17] S. Klaithin and C. Haruechaiyasak, "Traffic Information Extraction and Classification from Thai Twitter," Comput. Sci. Softw. Eng. (JCSSE), 2016 13th Int. Jt. Conf., pp. 1–6, 2016, doi: 10.1109/JCSSE.2016.7748851.

[18] A. Kumar, M. Jiang, and Y. Fang, "Where not to go?: detecting road hazards using twitter," in Proceedings of the 37th international ACM …, 2014, vol. 2609550, pp. 1223–1226, doi: 10.1145/2600428.2609550.

[19] E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, "Real-Time Detection of Traffic from Twitter Stream Analysis," IEEE Trans. Intell. Transp. Syst., vol. 16, no. 4, pp. 2269–2283, 2015, doi: 10.1109/TITS.2015.2404431.

[20] D. A. Kurniawan, S. Wibirama, and N. A. Setiawan, "Real-time Traffic Classification with Twitter Data Mining," in In 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), 2016, pp. 1–5, doi: 10.1109/ICITEED.2016.7863251.

[21] and F. C. Hoang Nguyen, Wei Liu, Paul Rivera, "TrafficWatch: Real-Time Traffic Incident Detection and Monitoring Using Social Media," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9651, pp. 540–551, 2016, doi: 10.1007/978-3-319-31753-3.

[22] M. A. Yazici, S. Mudigonda, and C. Kamga, "Incident Detection through Twitter Organization vs. Personal Accounts," no. 212, pp. 1–17, 2017.

[23] D. Semwal, S. Patil, S. Galhotra, A. Arora, and N. Unny, "STAR: Real-time Spatio-Temporal Analysis and Prediction of Traffic Insights using Social Media," in In Proceedings of the 2nd IKDD Conference on Data Sciences, 2015, p. 7, doi: 10.1145/2778865.2778872.

[24] P. Tejaswin, R. Kumar, and S. Gupta, "Tweeting Traffic: Analyzing Twitter for generating real-time city traffic insights and predictions," in Proceedings of the 2nd IKDD Conference on Data Sciences - CODS-IKDD '15, 2015, pp. 1–4, doi: 10.1145/2778865.2778874.

[25] P. Anantharam, P. Barnaghi, K. Thirunarayan, and A. Sheth, "Extracting City Traffic Events from Social Streams," ACM Trans. Intell. Syst. Technol., vol. 6, no. 4, pp. 1–27, 2015, doi: 10.1145/2717317.

[26] M. R. Langa and M. N. Moeti, "A Real-Time Notification System for Traffic Congestion on South African National Routes," pp. 79–91, 2022.

[27] N. Suat-rojas, C. Gutierrez-osorio, and C. Pedraza, "Extraction and Analysis of Social Networks Data to Detect Traffic Accidents," 2022.

[28] D. M. Blei, B. B. Edu, A. Y. Ng, A. S. Edu, M. I. Jordan, and J. B. Edu, "Latent Dirichlet Allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003, doi: 10.1162/jmlr.2003.3.4-5.993.

[29] D. Wang, A. Al-Rubaie, S. S. Clarke, and J. Davies, "Real-Time Traffic Event Detection From Social Media," ACM Trans. Internet Technol, vol. 18, no. 23, pp. 1–23, 2017, doi: 10.1145/3122982.

[30] W. Huang, S. Xu, Y. Yan, and A. Zipf, "An exploration of the interaction between urban human activities and daily traffic conditions: A case study of Toronto, Canada," Cities, no. July, 2018, doi: 10.1016/j.cities.2018.07.001.

[31] F. Ali, A. Ali, M. Imran, R. A. Naqvi, M. H. Siddiqi, and K. S. Kwak, "Traffic accident detection and condition analysis based on social networking data," Accid. Anal. Prev., vol. 151, no. December 2020, p. 105973, 2021, doi: 10.1016/j.aap.2021.105973.

[32] Y. Gu, Z. (Sean) Qian, and F. Chen, "From Twitter to detector: Real-time traffic incident detection using social media data," Transp. Res. Part C Emerg. Technol., vol. 67, pp. 321–342, 2016, doi: 10.1016/j.trc.2016.02.011.

[33] R. Y. K. Lau, "Toward a social sensor based framework for intelligent transportation," in 2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2017, pp. 1–6, doi: 10.1109/WoWMoM.2017.7974354.

[34] Y. Chen, Y. Lv, X. Wang, and F.-Y. Wang, "A convolutional neural network for traffic information sensing from social media text," 2017 IEEE 20th Int. Conf. Intell. Transp. Syst., pp. 1–6, 2017, doi: 10.1109/ITSC.2017.8317650.

[35] Y. Ji, J. Wang, Y. Niu, and H. Ma, "Reliable Event Detection via Multiple Edge Computing on Streaming Traffic Social Data," IEEE Access, pp. 1–14, 2021, doi: 10.1109/ACCESS.2021.3060624.

[36] P. Ambastha and M. S. Desarkar, "Incident Detection from Social Media Targeting Indian Traffic Scenario Using Transfer Learning," 2020 IEEE 23rd Int. Conf. Intell. Transp. Syst. ITSC 2020, 2020, doi: 10.1109/ITSC45102.2020.9294295.

[37] Y. Kim et al., "Regional Traffic Event Detection Using Data Crowdsourcing," Appl. Sci., vol. 13, no. 16, 2023, doi: 10.3390/app13169422.

[38] R. R. Almassar and A. S. Girsang, "Detection of traffic congestion based on twitter using convolutional neural network model," IAES Int. J. Artif. Intell., vol. 11, no. 4, pp. 1448–1459, 2022, doi: 10.11591/ijai.v11.i4.pp1448-1459.

[39] K. Swapnika and D. Vasumathi, "a Hybrid Dnn-Hho Approach for Event Detection in Big Data," Indian J. Comput. Sci. Eng., vol. 13, no. 5, pp. 1401–1411, 2022, doi: 10.21817/indjcse/2022/v13i5/221305086.

[40] C. Gutierrez, P. Figuerias, P. Oliveira, R. Costa, and R. Jardim-Goncalves, "Twitter mining for traffic events detection," Proc. 2015 Sci. Inf. Conf. SAI 2015, pp. 371–378, 2015, doi: 10.1109/SAI.2015.7237170.

[41] H. Shekhar and S. Setty, "Disaster analysis through tweets," 2015 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2015, no. August, pp. 1719–1723, 2015, doi: 10.1109/ICACCI.2015.7275861.

[42] A. Hodorog, I. Petri, and Y. Rezgui, "Machine learning and Natural Language Processing of social media data for event detection in smart cities," Sustain. Cities Soc., vol. 85, no. June, p. 104026, 2022, doi: 10.1016/j.scs.2022.104026.

[43] B. Mehri, M. Trépanier, and Y. Goussard, "Multilingual Text Classification on Social Media Data for Incident Alert in Subway Multilingual Text Classification on Social Media Data for Incident Alert in Subway Transportation Network," no. January, 2023.

[44] J. He et al., "MetroScope: An Advanced System for Real-Time Detection and Analysis of Metro-Related Threats and Events via Twitter," SIGIR 2023 - Proc. 46th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., pp. 3130–3134, 2023, doi: 10.1145/3539618.3591807.

[45] H. Shekhar, S. Setty, and U. Mudenagudi, "Vehicular traffic analysis from social media data," 2016 Int. Conf. Adv. Comput. Commun. Informatics, pp. 1628–1634, 2016, doi: 10.1109/ICACCI.2016.7732281.

[46] I. Salas, A., Georgakis, P., Nwagboso, C., Ammari, A. and Petalas, "Traffic Event Detection Framework Using Social Media," in IEEE International Conference on Smart Grid and Smart Cities, 2017, no. July, pp. 303–307, doi: 10.1109/ICSGSC.2017.8038595.

[47] Z. Xu et al., "Crowdsourcing based Description of Urban Emergency Events using Social Media Big Data," IEEE Trans. Cloud Comput., pp. 1–1, 2016, doi: 10.1109/TCC.2016.2517638.

[48] M. Berlingerio, F. Calabrese, G. Di Lorenzo, X. Dong, Y. Gkoufas, and D. Mavroeidis, "SaferCity: A system for detecting and analyzing incidents from social media," Proc. - IEEE 13th Int. Conf. Data Min. Work. ICDMW 2013, pp. 1077–1080, 2013, doi: 10.1109/ICDMW.2013.39.

[49] S. Suma, R. Mehmood, N. Albugami, I. Katib, and A. Albeshri, "Enabling Next Generation Logistics and Planning for Smarter Societies," Procedia - Procedia Comput. Sci., pp. 1–6, 2017.

[50] E. Chaniotakis, C. Antoniou, and ..., "Enhancing resilience to disasters using social media," … Technol. …, pp. 699–703, 2017, doi: 10.1109/MTITS.2017.8005602.

[51] L. C. Yang, B. Selvaretnam, P. K. Hoong, I. K. T. Tan, E. K. Howg, and L. H. Kar, "Exploration of road traffic tweets for congestion monitoring," J. Telecommun. Electron. Comput. Eng., vol. 8, no. 2, pp. 141–145, 2016, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84984848398&partnerID=40&md5=aa3409237b2ff2788facabd0f6edd723.

[52] M. Hasby and M. L. Khodra, "Optimal Path Finding based on Traffic Information Extraction from Twitter," Int. Conf. ICT Smart Soc., pp. 1–5, 2013, doi: 10.1109/ICTSS.2013.6588076.

[53] F. Muhammad and M. L. Khodra, "Event information extraction from Indonesian tweets using conditional random field," ICAICTA 2015 - 2015 Int. Conf. Adv. Informatics Concepts, Theory Appl., pp. 0–5, 2015, doi: 10.1109/ICAICTA.2015.7335383.

[54] A. M. B, Z. Jiang, S. Jones, and P. Sheinidashtegol, Detection of Damage and Failure Events of Road Infrastructure Using Social Media, vol. 10966. Springer International Publishing, 2018.

[55] J. P. Singh, Y. K. Dwivedi, N. P. Rana, A. Kumar, and K. K. Kapoor, "Event classification and location prediction from tweets during disasters," Ann. Oper. Res., pp. 1–21, 2017, doi: 10.1007/s10479-017-2522-3.

[56] M. Thelwall, K. Buckley, G. Paltoglou, and D. Cai, "Sentiment Strength Detection in Short Informal Text," Am. Soc. Informational Sci. Technol., vol. 61, no. 12, pp. 2544–2558, 2010, doi: 10.1002/asi.

[57] T. Ohbe, T. Ozono, and T. Shintani, "Developing a sentiment polarity visualization system for local event information analysis," Proc. - 2016 5th IIAI Int. Congr. Adv. Appl. Informatics, IIAI-AAI 2016, pp. 19–24, 2016, doi: 10.1109/IIAI-AAI.2016.118.

[58] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," Processing, vol. 150, no. 12, pp. 1–6, 2009, doi: 10.1016/j.sedgeo.2006.07.004.

[59] T. Adetiloye and A. Awasthi, Traffic Condition Monitoring Using Social Media Analytics, vol. 44. Springer Singapore, 2018.

[60] H. Kang, S. J. Yoo, and D. Han, "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews," Expert Syst. Appl., vol. 39, no. 5, pp. 6000–6010, Apr. 2012, doi: 10.1016/J.ESWA.2011.11.107.

[61] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," Mob. Networks Appl., vol. 19, no. 2, pp. 171–209, 2014, doi: 10.1007/s11036-013-0489-0.

[62] R. M. Ebtesam Alomari, Iyad Katib, "Big Data Technologies for Arabic Social Media Analysis to enable Smarter Transportation," King AbdualAziz University, 2021.

[63] D. T. Nguyen and J. E. Jung, "Real-time event detection for online behavioral analysis of big social data," Futur. Gener. Comput. Syst., vol. 66, pp. 137–145, 2017, doi: 10.1016/j.future.2016.04.012.

[64] H. Khazaei, R. Veleda, M. Litoiu, and A. Tizghadam, "Realtime big data analytics for event detection in highways," 2016 IEEE 3rd World Forum Internet Things, WF-IoT 2016, pp. 472–477, 2017, doi: 10.1109/WF-IoT.2016.7845461.

[65] K. R. Pandhare and M. A. Shah, "Real time road traffic event detection using Twitter and spark," 2017 Int. Conf. Inven. Commun. Comput. Technol., no. Icicct, pp. 445–449, 2017, doi: 10.1109/ICICCT.2017.7975237.

[66] G. Kousiouris et al., "An integrated information lifecycle management framework for exploiting social network data to identify dynamic large crowd concentration events in smart cities applications," Futur. Gener. Comput. Syst., vol. 78, pp. 516–530, 2018, doi: 10.1016/j.future.2017.07.026.

[67] E. Alomari, R. Mehmood, and I. Katib, "Sentiment Analysis of Arabic Tweets for Road Traffic Congestion and Event Detection," in In: Mehmood R., See S., Katib I., Chlamtac I. (eds) Smart Infrastructure and Applications: Foundations for Smarter Cities and Societies, Springer (https://www.springer.com/us/book/9783030137045), 2020, pp. 37–54.

[68] E. Alomari, R. Mehmood, and I. Katib, "Road Traffic Event Detection Using Twitter Data, Machine Learning, and Apache Spark," in 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Aug. 2019, pp. 1888–1895, doi: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00332.

[69] E. Alomari, I. Katib, A. Albeshri, and R. Mehmood, "COVID-19: Detecting Government Pandemic Measures and Public Concerns from Twitter Arabic Data Using Distributed Machine Learning," Int. J. Environ. Res. Public Health, vol. 18, no. 1, p. 282, Jan. 2021, doi: 10.3390/ijerph18010282.

[70] E. Alomari, I. Katib, A. Albeshri, T. Yigitcanlar, and R. Mehmood, "Iktishaf+: A big data tool with automatic labeling for road traffic social sensing and event detection using distributed machine learning," Sensors, vol. 21, no. 9, pp. 1–33, 2021, doi: 10.3390/s21092993.

[71] Q. Chen, W. Wang, K. Huang, S. De, and F. Coenen, "Multi-modal generative adversarial networks for traffic event detection in smart cities," Expert Syst. Appl., vol. 177, no. March, p. 114939, 2021, doi: 10.1016/j.eswa.2021.114939.

A Proposed Roadmap for Optimizing Predictive Maintenance of Industrial Equipment

Maria Eddarhri[1], Mustapha Hain[2], Jihad Adib[3], Abdelaziz Marzak[4]
Industrial Engineering Department AICSE Laboratory, ENSAM University of Hassan II, Casablanca, Morocco[1, 2]
Mathematics and Computer Sciences Department LTIM-Faculty of Sciences, University of Hassan II, Casablanca, Morocco[3, 4]

***Abstract*—Now-a-days, the maintenance management of industrial equipment, particularly in the aeronautical industry, has evolved into a substantial challenge and a critical concern for the sector. Aeronautical wiring companies are currently grappling with escalating difficulties in equipment maintenance. This paper proposes an intelligent system for the automated detection of machine failures. It assesses predictive maintenance approaches and underscores the significance of sensor selection to optimize outcomes. The integration of Machine Learning techniques with the Industrial Internet of Things (IIoT) and intelligent sensors is presented, showcasing the heightened accuracy and effectiveness of predictive maintenance, especially in the aeronautical industry. The research aims to leverage Predictive Maintenance for enhancing the performance of production machines, predicting their failures, recognizing faults, and determining maintenance dates through the analysis and processing of collected data. Employing sophisticated code, the study emphasizes real-time data collection, data traceability, and enhanced precision in predicting potential failures using Machine Learning. The findings underscore the collaboration between sensors and the synergy of Machine Learning with IIoT, ultimately aiming for sustained reliability and efficiency of predictive maintenance in aeronautical wiring companies.**

Keywords*—Predictive maintenance; intelligent system; aeronautical wiring companies; machine learning; IIoT

I. Introduction

Maintenance in aeronautical wiring companies is crucial to ensure the safety and reliability of aircraft electrical systems. Technicians conduct regular inspections, testing, and repairs to meet aviation standards. Adherence to regulations and the adoption of advanced techniques like predictive maintenance are essential for optimal performance and safety. In the previous article [1], a maintenance management model employing machine learning was implemented. This model utilizes an expert system to support diagnostics and generate action plans for repairs based on the type of failure. The expert system reduces maintenance intervention time, minimizes machine downtime, and serves as the foundation for predictive maintenance.

The main objective of this paper is to significantly improve the maintenance approach by integrating intelligent sensors. These sensors ensure seamless communication via the (IIoT) and are connected to a central server to collect real-time data, which is efficiently stored in a database. Machine learning leverages this data to predict potential breakdowns and detect patterns and trends in the state of industrial equipment. With this proactive approach, corrective measures can be swiftly taken to minimize unexpected downtime, reduce high maintenance costs, and ensure continuous reliability and efficiency in the aeronautical industry.

This paper proposes an intelligent system for automatically detecting failures in machines. The initial focus is on presenting and evaluating the maintenance predictive approaches, including the approach detailed in article [1]. Following that, an in-depth discussion ensues regarding various sensor types and their significance in optimizing and predicting outcomes. Emphasis is placed on selecting sensors that support inter-sensor communication, fostering collaboration, and leveraging IIoT to enhance predictive maintenance capabilities, particularly in the aeronautical industry.

Additionally, the integration of Machine Learning techniques with IIoT and intelligent sensors is showcased, illustrating their synergistic impact on improving the accuracy and effectiveness of predictive maintenance processes.

The main contribution of this paper is integrating intelligent diagnostic sensors with Industrial Internet of Things (IIoT) technology and machine learning analysis in aeronautical industries, specifically in wire cutting machines. This integration enables real-time data collection and storage in a centralized database. The ultimate goal is to optimize maintenance procedures, ensuring the long-term reliability and efficiency of cutting cable machines in aeronautical industries.

The structure of this paper is organized as follows: Section II provides a review of existing approaches. Section III presents materials and methods. In Section IV, the proposed model is introduced, and results and discussion are presented in Section V. Finally, the last section concludes by highlighting the main contributions and outlining future work.

II. Review of Existing Approaches

To establish a detailed context for exploration, a comprehensive analysis of predictive maintenance in the industrial sector is undertaken in the following sections. Furthermore, numerous research studies have investigated different facets of intelligent sensor integration for maintenance in smart factories [2].

Specifically, Song et al. [3] focus on smart sensors used to monitor rock bolts' condition and integrity to minimize economic and personnel losses. Jin [4] discusses multifunctional sensors suitable for industrial production, while Feng [5] delves into sensors for intelligent gas sensing. Paidi

[6] examines intelligent parking sensors that utilize machine learning to replace ultrasonic sensors. In addition, Talal and Sony [7] [8] characterize sensors for health monitoring. The concept of combining smart sensors and smart factories, a crucial aspect of Industry 4.0, is commonly found in literature reviews. Lee [9] illustrates the use of smart sensors to evaluate and diagnose individual devices in smart factories.

Expanding the review, Strozzi in [10] emphasizes the actual transition and implementation of large, intelligent factories. Pereira and Álvarez [11] [12] focus on implementing smart factory principles and highlight that effective value creation depends on the chosen method of implementation.

The implementation process, which involves managing technological and organizational changes and desirable competencies, is further addressed by Sousa [13] and Lee et al. [14]. They also draw attention to the gap between recent research and the actual level of deployment.

Regarding maintenance in intelligent factories, several literature reviews revolve around predictive maintenance. Carvalho [15] concentrates on machine learning methods, considering them a promising tool for predictive maintenance. Sakib [16] observes a shift from reactive service activities to proactive, predictive maintenance and relates it to the context of Industry 4.0.

In specific applications, Olesen and Shaker [17] study practical use in thermal power plants, while Fei [18] focuses on predictive maintenance in aircraft systems. As for the approach [1], a maintenance management model utilizing Machine Learning has been proposed to optimize and enhance reliability. The model includes a referential and expert system that aids in diagnosing breakdowns, assigning technicians, and generating detailed repair action plans. The expert system reduces intervention time, minimizing machine downtime, and facilitates predictive maintenance implementation.

So, studies focus on the use of smart sensors to monitor conditions, enhance safety, improve efficiency and implement predictive maintenance techniques, with the aim of minimizing economic and human losses. However, it should be noted that the integration of smart sensors via IIoT and real-time data collection for predictive maintenance in the aeronautical industry, particularly in cable-cutting machines, has not been discussed in these studies.

The current studies primarily concentrate on utilizing smart sensors to monitor conditions, enhance safety, improve efficiency, and implement predictive maintenance techniques, all with the objective of reducing economic and human losses. However, it's crucial to note that the aeronautical industry has not received comprehensive attention in these studies, specifically regarding the integration of smart sensors through the IIoT and real-time data collection for predictive maintenance, especially in the context of cable-cutting machines. This identified gap calls for special consideration to fully exploit the potential benefits of technology in this specific field of aviation. Further research and development in this area can lead to significant advancements and improvements in maintenance practices within the aeronautical industry. Although the proposed expert system successfully reduces technician assignment and maintenance intervention time [1], it faces a notable limitation. The data collected by technicians may be unreliable, which can have adverse consequences on the maintenance process. One major limitation of this approach is its failure to identify the specific nature of equipment failures. Human involvement is still necessary to diagnose the equipment and pinpoint the failed components. However, this manual approach to data collection slows down the maintenance process, resulting in machine immobilization.

The diagnostic aspect is crucial as it allows for accurate identification of the failure and the specific part that requires replacement. Without this information, machine downtime can be prolonged, especially if there is an inadequate stock of spare parts available. Therefore, improvements are needed to enhance data reliability and the system's ability to diagnose failures effectively.

So, the primary goal is to enhance the maintenance approach significantly by employing intelligent sensors specially created for diagnostics. These sensors will communicate through the IIoT, enabling the real-time collection and efficient storage of data in a centralized database. Subsequently, machine learning will analyze the data to predict potential breakdowns. By adopting this proactive strategy, intervention actions can be implemented before issues escalate, thereby minimizing downtime, maintenance expenses, and disturbances in the production process. The ultimate objective is to optimize maintenance, ensuring the reliability and efficiency of industrial equipment.

III. Materials and Methods

A. Introduction to Sensors

The new requirements imposed on industrial systems in their operation and in their production and in the quality of their production, require a very elaborate strategy in the control of these installations. The difficulty is to have relevant and reliable information that allows to generate an effective corrective action. Sensors are these sources of information.

The development of information processing capabilities allows the control and automation of increasingly complex systems.

The calculation possibilities of the controlling parties seem to be limited only by the quantity and quality of the data provided to them. So Sensors play an important role in the automation of any kind of application by measuring and processing the data collected to detect changes in physical things. Fig. 1 illustrates the sensor operation.

Intelligent sensors are multi-component measuring devices that are self-calibrating, self-optimized, and simple to integrate into the environment for high connectivity, according to Eifert [19]. Furthermore, intelligent sensors possess process intelligence and are capable of generating multidimensional data information.

A sensor is a device, module, machine, or subsystem that detects events or changes in its environment and relays the information to other electronics, most commonly a computer processor [20]. A sensor converts the physical parameters into

a measurable digital signal, which can then be presented, read or processed.

Several classifications of sensors exist, established by different authors and experts. There are some that are very simple and others that are very complex. In this classification of sensors, they are divided into two categories: active and passive. Active sensors are those that require an external excitation signal or power signal. In contrast, passive sensors do not need an external power supply and are able to generate an output response directly.

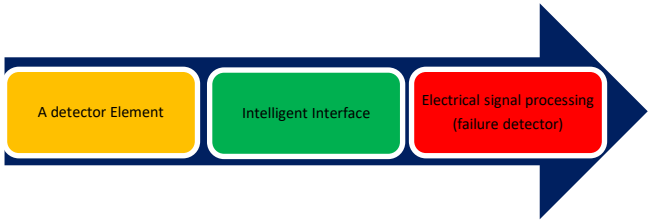


Fig. 1. Illustration view on the working of sensor.

B. *Different types of Sensors*

There are different types of sensors varying from the simplest to the most complex. Sensors are classified according to their specifications, their conversion method, the type of material used, the physical phenomenon of detection, the properties of what it measures and the application field [21].

- Proximity Sensors: Without making direct physical touch, proximity sensors make it simple to determine the location of any adjacent object. It detects the existence of an object by simply looking for any fluctuation in the return signal after emitting electromagnetic radiation, such as infrared. Numerous types of proximity sensors, including inductive, capacitive, ultrasonic, photoelectric, magnetic, and others, are available and are aimed at various purposes. This specific sort of sensor is frequently utilized in applications that demand efficiency and security. This sort of sensor has several applications, including object detection, item counting, rotation measurement, object positioning, material detection, movement direction measurement, parking sensors, and others. The best applications for proximity sensors are found in a variety of industries [21] [22] [23].

- Position Sensors: By sensing motion, the position sensor determines the presence of humans or objects in a specific region. It can be used in home security to track the doors and windows of rooms and appliances from any location. It informs them of the open or closed state at all times and can follow intruders while they are away. It can be used in health care monitoring to track the location of patients, nurses, and doctors in a hospital [24], as well as in agriculture to track the whereabouts of animals [25].

- Pressure Sensors: Liquid or other forms of pressure are used in a variety of devices. These sensors enable the creation of IoT systems that monitor pressure-driven systems and devices. Any variation from the typical pressure range alerts the system administrator to any issues that need to be addressed [26][27] The use of these sensors is beneficial not only in production but also in the maintenance of complete water and heating systems since it is simple to detect any pressure fluctuations or decreases [20] [28] [29].

- Temperature sensors: It primarily used to control air conditioning, freezers and other environmental control devices. Now they are used in manufacturing, agriculture and healthcare. Since a defined ambient and device temperature is required for most equipment in the manufacturing process, this type of measurement can still be used to improve the production process. Soil temperature, on the other hand, is crucial for crop growth in agriculture. It helps the plants to grow properly, which results in optimal results[20].

- Chemical Sensors: In the world, there are various types of chemical sensors, some of which are used to measure the chemical composition of the environment. By monitoring chemical plumes in the environment, a wireless chemical sensor network can monitor air quality [30].

- Occupancy sensors: The presence sensor, also known as the occupancy sensor, detects the presence of individuals or items in a specific area. The sensor can be used to detect numerous characteristics such as temperature, humidity, light, and air from a distance. The authors provide a similar use of these types of sensors in [31].

- With regard to the choice of intelligent sensors in the industrial field for predictive maintenance and specifically for aeronautical cable cutting machines as discussed in the previous article [1], specific criteria for collecting the data needed for early detection of failures need to be taken into account.

- The various intelligent sensors to be installed (temperature, vibration, pressure, etc.) must be able to communicate with each other to share the data collected. In addition, it is essential to check that the sensors are capable of collecting data at an appropriate frequency to enable real-time monitoring and rapid detection of anomalies. It is essential that these intelligent sensors support IIoT-compatible communication technologies, making it easier to collect and analyze data remotely, as well as integrating them into more comprehensive predictive maintenance systems.

It is also important to ensure that these sensors provide useful information that can be easily interpreted to analyze the state of the machine. The data collected must be relevant for predicting breakdowns and contributing effectively to the maintenance planning process.

Another crucial aspect is to ensure that the sensors can be easily integrated into the existing cable-cutting machine without disrupting its normal operation. Successful integration ensures that the sensors fulfil their role without compromising the overall performance of the machine.

C. About PLCs

PLCs are industrial automation devices that perform two main functions: measuring process parameters using sensors and controlling equipment according to specific programs [32].

These specialized systems solve sequential and combined logic problems using programmable configurations. PLCs are used in a wide range of applications due to their ease of programming, accessibility and high reliability. They are equipped with digital or analogue input and output cards. PLCs are programmed using various communication protocols and specialized software [32]. PLC is considered as the brain of the automation setup, controlling and coordinating various industrial processes. The PLC interprets the sensor data to trigger appropriate actions. The seamless integration of sensors with PLCs empowers industries to monitor and regulate processes with enhanced accuracy, efficiency, and safety.

IV. Proposed Work

- Using IIoT, a traceability system has been developed to ensure communication between the PLC and the server through the MQTT communication protocol. Through this integration, a high-performance traceability system has been implemented to ensure real-time data collection and monitoring via a dashboard.
- Firstly, connecting the sensors to the IIoT network allows continuous data collection on various monitored parameters such as temperature, pressure, vibrations, etc. This data is then transferred to a centralized server for storage and analysis.

Data traceability makes it possible to track the complete history of measurements made by the sensors. This provides a complete and detailed overview of how parameters evolve over time, making it easier to identify trends and anomalies.

Real-time data collection enables proactive monitoring of equipment. Sensor values are constantly updated, enabling unusual variations to be detected quickly. This real-time information can be used to make rapid decisions and avoid expensive failures. With an intuitive dashboard, users can view collected data, trends, alerts and predictive maintenance information. The dashboard can display graphs, customized dashboards and notifications when significant thresholds or deviations from normal values are exceeded.

V. Results and Discussion

By using Machine Learning techniques, the collected data can be analyzed to predict potential failures.

Machine Learning algorithms are trained to recognize patterns and warning signals that indicate imminent failure. These predictions enable maintenance activities to be planned proactively, avoiding production interruptions and reducing the costs associated with unexpected breakdowns.

A model will be created by integrating IIoT and artificial intelligence. this model uses sensor connectivity, data traceability, real-time data collection and predictive analysis to improve predictive maintenance. This optimizes operations, minimizes unplanned downtime and delivers significant savings in maintenance and repair costs.

As shown in Fig. 2, the proposed model for maintenance management demonstrates the key role that IIOT & AI play in failure prediction and production line optimization.

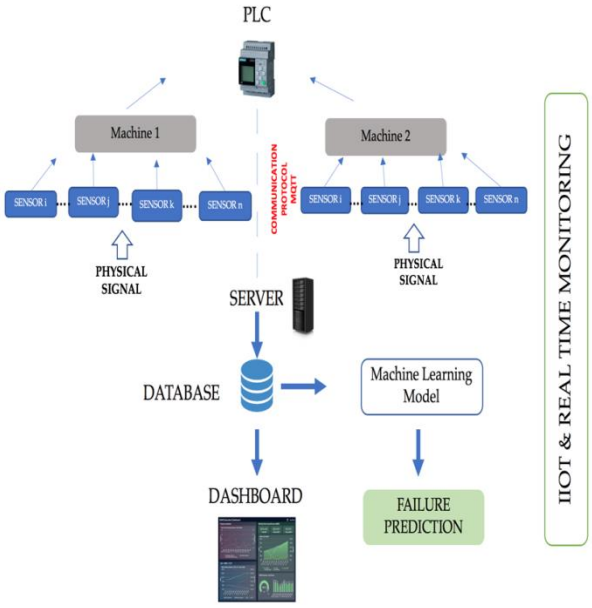


Fig. 2. Modeling of the maintenance process using intelligent sensors.

To implement the maintenance optimization approach, extensive and sophisticated code has been developed. This code plays a central role in the collection, processing and analysis of data from intelligent sensors interconnected via the IIoT (Industrial Internet of Things).

The first step in development was to create a communication interface that enables the sensors to transmit their measurements in real time to the designated data storage server. With this seamless connection, a huge amount of data can be collected from different pieces of industrial equipment, providing a global view of their operating status.

Once the data had been collected, the core of the code was based on the use of machine learning. A predictive model has been trained using advanced machine learning algorithms, capable of detecting patterns, trends, and anomalies in the sensor data. This model is continually refined and updated as new data is collected, allowing it to improve in terms of accuracy and reliability.

The pseudocode that schematically illustrates the main steps in this approach is shown in Figure 3. The pseudocode describes the general logic of the algorithm, including data pre-processing steps, model training, failure prediction techniques, and maintenance decision making.

Using this extensive code and machine learning approach enables the prediction of potential failures with enhanced precision. Anticipating problems before they occur allows proactive planning and execution of maintenance interventions, minimizing unplanned downtime and maximizing the availability and efficiency of industrial equipment.

```
M1.Listsensors=[S1...Sn]
M1.sensorAttributes=[temperature,pressure,...vibration] #n_feature
M2.Listsensors=[S1...Sm]
M2.sensorAttributes=[temperature,pressure,... vibration]  #m_feature


for sensor in M1.Listsensors:
    collect M1.sensorAttributes[sensor] from M1.Listsensors[sensor]


for sensor in M2.Listsensors:
    collect M2.sensorAttributes[sensor]  from M2.Listsensors[sensor]


DATA={"M1":M1.sensorAttribute ;"M2":M2.sensorAttribute}

Send DATA  to Server
#PLC -SERVER COMMUNICATION USING MQTT PROTOCOL
Stock DATA in DATABASE
Schematize and Monitore DATA by a realtime Dashboard
Analyse collected DATA BY ML MODEL
Make failures prediction
Display on Dashboard
```

Fig. 3. Proposed pseudocode.

In summary, the extended code and the methodology illustrated by the pseudocode in Fig. 3 provide a powerful decision-support tool, enabling the maximization of the reliability and performance of these industrial assets while reducing the costs associated with corrective maintenance.

VI. CONCLUSION AND FUTURE SCOPE

In conclusion, the adoption of smart sensors and the Industrial Internet of Things (IIoT) to enhance the maintenance approach represents a significant advancement in the industry. This proactive strategy, based on real-time data analysis through machine learning, provides the opportunity to anticipate potential failures and take action before they escalate into major issues. For aerospace companies, especially those utilizing cable cutting machines, this approach ensures not only optimal performance of their equipment but also precise traceability and real-time monitoring of each operation.

In future work, implementation of predictive maintenance using the proposed model is planned. This approach will further enhance the ability to foresee and prevent failures, thereby contributing to more efficient industrial operations management. By combining technological expertise with a proactive approach, the goal is to ensure continuous and reliable production while optimizing costs and maximizing the lifespan of equipment.

REFERENCES

[1] M. Eddarhri, J. Adib, M. Hain, and A. Marzak, “Towards predictive maintenance: the case of the aeronautical industry,” Procedia Computer Science, vol. 203, pp. 769–774, Jan. 2022, doi: 10.1016/j.procs.2022.07.115.

[2] M. Pech, J. Vrchota, and J. Bednář, “Predictive Maintenance and Intelligent Sensors in Smart Factory: Review,” Sensors, vol. 21, no. 4, p. 1470, Feb. 2021, doi: 10.3390/s21041470.

[3] G. Song, W. Li, B. Wang, and S. C. M. Ho, “A Review of Rock Bolt Monitoring Using Smart Sensors,” Sensors (Basel), vol. 17, no. 4, p. 776, Apr. 2017, doi: 10.3390/s17040776.

[4] X. Jin et al., “Review on exploration of graphene in the design and engineering of smart sensors, actuators and soft robotics,” Chemical Engineering Journal Advances, vol. 4, p. 100034, Dec. 2020, doi: 10.1016/j.ceja.2020.100034.

[5] S. Feng et al., “Review on Smart Gas Sensing Technology,” Sensors, vol. 19, no. 17, Art. no. 17, Jan. 2019, doi: 10.3390/s19173760.

[6] V. Paidi, H. Fleyeh, J. Håkansson, and R. G. Nyberg, “Smart parking sensors, technologies and applications for open parking lots: a review,” IET Intelligent Transport Systems, vol. 12, no. 8, pp. 735–741, 2018, doi: 10.1049/iet-its.2017.0406.

[7] M. Talal et al., “Smart Home-based IoT for Real-time and Secure Remote Health Monitoring of Triage and Priority System using Body Sensors: Multi-driven Systematic Review,” J Med Syst, vol. 43, no. 3, p. 42, Jan. 2019, doi: 10.1007/s10916-019-1158-z.

[8] M. Sony, S.S. Naik, “Ten Lessons for Managers While Implementing Industry 4.0 | IEEE Journals & Magazine | IEEE Xplore.” https://ieeexplore.ieee.org/document/8704884 (accessed Sep. 25, 2023).

[9] G.-Y. Lee et al., “Machine health management in smart factory: A review,” J Mech Sci Technol, vol. 32, no. 3, pp. 987–1009, Mar. 2018, doi: 10.1007/s12206-018-0201-1.

[10] F. Strozzi, C. Colicchia, A. Creazza, and C. Noè, “Literature review on the ‘Smart Factory’ concept using bibliometric tools,” International Journal of Production Research, vol. 55, no. 22, pp. 6572–6591, Nov. 2017, doi: 10.1080/00207543.2017.1326643.

[11] A. C. Pereira and F. Romero, “A review of the meanings and the implications of the Industry 4.0 concept,” Procedia Manufacturing, vol. 13, pp. 1206–1214, Jan. 2017, doi: 10.1016/j.promfg.2017.09.032.

[12] I. L. Bahena-Álvarez, E. Cordón-Pozo, and A. Delgado-Cruz, “Social Entrepreneurship in the Conduct of Responsible Innovation: Analysis Cluster in Mexican SMEs,” Sustainability, vol. 11, no. 13, Art. no. 13, Jan. 2019, doi: 10.3390/su11133714.

[13] M. J. Sousa, R. Cruz, Á. Rocha, and M. Sousa, “Innovation Trends for Smart Factories: A Literature Review,” Á. Rocha, H. Adeli, L. P. Reis, and S. Costanzo, Eds., in Advances in Intelligent Systems and Computing, vol. 930. Cham: Springer International Publishing, 2019, pp. 689–698. doi: 10.1007/978-3-030-16181-1_65.

[14] S. Lee, J.-Y. Kim, and W. Lee, “Smart Factory Literature Review and Strategies for Korean Small Manufacturing Firms,” Journal of Information Technology Applications and Management, vol. 24, no. 4, pp. 133–152, 2017, doi: 10.21219/jitam.2017.24.4.133.

[15] T. P. Carvalho, F. A. A. M. N. Soares, R. Vita, R. da P. Francisco, J. P. Basto, and S. G. S. Alcalá, “A systematic literature review of machine learning methods applied to predictive maintenance,” Computers & Industrial Engineering, vol. 137, p. 106024, Nov. 2019, doi: 10.1016/j.cie.2019.106024.

[16] N. Sakib and T. Wuest, “Challenges and Opportunities of Condition-based Predictive Maintenance: A Review,” Procedia CIRP, vol. 78, pp. 267–272, Jan. 2018, doi: 10.1016/j.procir.2018.08.318.

[17] J. Fausing Olesen and H. R. Shaker, “Predictive Maintenance for Pump Systems and Thermal Power Plants: State-of-the-Art Review, Trends and Challenges,” Sensors, vol. 20, no. 8, Art. no. 8, Jan. 2020, doi: 10.3390/s20082425.

[18] X.Fei,“Literature Review: Framework of Prognostic Health Management for Airline Predictive Maintenance | IEEE Conference Publication | IEEE Xplore.” https://ieeexplore.ieee.org/document/9164546 (accessed Sep. 25, 2023).

[19] T. Eifert, K. Eisen, M. Maiwald, and C. Herwig, “Current and future requirements to industrial analytical infrastructure—part 2: smart sensors,” Anal Bioanal Chem, vol. 412, no. 9, pp. 2037–2045, Apr. 2020, doi: 10.1007/s00216-020-02421-1.

[20] M. Javaid, A. Haleem, S. Rab, R. Pratap Singh, and R. Suman, “Sensors for daily life: A review,” Sensors International, vol. 2, p. 100121, Jan. 2021, doi: 10.1016/j.sintl.2021.100121.

[21] D. Sehrawat and N. S. Gill, "Smart Sensors: Analysis of Different Types of IoT Sensors," in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India: IEEE, Apr. 2019, pp. 523–528. doi: 10.1109/ICOEI.2019.8862778.

[22] K. T. V. Grattan and T. Sun, "Fiber optic sensor technology: an overview," Sensors and Actuators A: Physical, vol. 82, no. 1, pp. 40–61, May 2000, doi: 10.1016/S0924-4247(99)00368-4.

[23] Th. Kwaaitaal, "The fundamentals of sensors," Sensors and Actuators A: Physical, vol. 39, no. 2, pp. 103–110, Nov. 1993, doi: 10.1016/0924-4247(93)80205-U.

[24] S. K. Dhar, S. S. Bhunia, and N. Mukherjee, "Interference Aware Scheduling of Sensors in IoT Enabled Health-Care Monitoring System," presented at the 2014 Fourth International Conference of Emerging Applications of Information Technology (EAIT), IEEE Computer Society, Dec. 2014, pp. 152–157. doi: 10.1109/EAIT.2014.50.

[25] M. S. Mekala and V. Perumal, A novel technology for smart agriculture based on IoT with cloud computing. 2017, p. 82. doi: 10.1109/I-SMAC.2017.8058280.

[26] Y. Zang, F. Zhang, C. Di, and D. Zhu, "Advances of flexible pressure sensors toward artificial intelligence and health care applications," Mater. Horiz., vol. 2, Oct. 2014, doi: 10.1039/C4MH00147H.

[27] R. Kumar, S. Rab, B. D. Pant, and S. Maji, "Design, development and characterization of MEMS silicon diaphragm force sensor," Vacuum, vol. 153, pp. 211–216, Jul. 2018, doi: 10.1016/j.vacuum.2018.04.029.

[28] S. Rab et al., "Development of hydraulic cross floating valve," Review of Scientific Instruments, vol. 90, p. 085102, Aug. 2019, doi: 10.1063/1.5089953.

[29] R. Kumar, S. Rab, B. Pant, S. Maji, and R. Mishra, "FEA-Based Design Studies for Development of Diaphragm Force Transducers," MAPAN, vol. 34, Nov. 2018, doi: 10.1007/s12647-018-0292-2.

[30] A. Tapashetti and T. Ogunfunmi, IoT-enabled air quality monitoring device: A low cost smart health solution. 2016, p. 685. doi: 10.1109/GHTC.2016.7857352.

[31] N. Nesa and I. Banerjee, "IoT-Based Sensor Data Fusion for Occupancy Sensing Using Dempster–Shafer Evidence Theory for Smart Buildings," IEEE Internet of Things Journal, vol. PP, pp. 1–1, Jul. 2017, doi: 10.1109/JIOT.2017.2723424.

[32] S. G. Robert, N. Bizon, and M. Oproescu, "The importance of PLC in the predictive maintenance of electronic equipment," 2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), pp. 1–5, Jun. 2018, doi: 10.1109/ECAI.2018.8679026.

Research on 3D Target Detection Algorithm Based on PointFusion Algorithm Improvement

Jun Wang, Shuai Jiang, Linglang Zeng, Ruiran Zhang
School of Mechanical and Electrical Engineering, Jiangxi University of Science and Technology, Ganzhou, China

***Abstract*—With the continuous development of automatic driving technology, the requirements for the accuracy of 3D target detection in complex traffic scenes are getting higher and higher. To solve the problems of low recognition rate, long detection time, and poor robustness of traditional detection methods, this paper proposes a new method based on PointFusion model improvement. The method utilizes the PointFusion network architecture to input 3D point cloud data and RGB image data into the PointNet++ and ResNeXt neural network structures, respectively, and adopts a dense fusion method to predict the spatial offsets of each input point to each vertex in the 3D selection box point by point, to output the 3D prediction box of the target. Experimental results on the KITTI dataset show that compared with the PointFusion network model, the improved PointFusion-based model proposed in this paper improves the 3D target detection accuracy in three different difficulty modes (easy, medium, and hard) and performs best in the medium difficulty mode. These findings highlight the potential of the method proposed in this paper to be applied in the field of autonomous driving, providing a reliable basis for navigating self-driving cars in complex environments.**

***Keywords*—Neural network; target detection; autonomous driving; PointFusion; deep learning**

I. INTRODUCTION

With the rapid development of computer vision and deep learning technology, driverless vehicles are moving towards the practical stage. However, in intricate road conditions and uncertain traffic scenarios, the safety of automated driving technology is becoming more and more important. How to recognize obstacles efficiently and accurately has become an important challenge in the field of autonomous driving.

Target detection plays an important role in autonomous driving [1, 2]. A large number of methods have been proposed to solve the obstacle recognition problem in autonomous driving. In the field of 3D target detection, depending on the modality of the sensor data used in 3D detection networks, they can be broadly categorized into detection methods based on image, point cloud, and bimodal information fusion of image and point cloud. The SMOKE network model proposed by Liu et al. [3] is an image-based 3D target detection that utilizes feature point estimation and 3D spatial variable regression to determine the spatial location of the target, which has the advantage of a simple data preprocessing stage that improves the detection speed, while the network model solves the effect of noise introduced due to redundancy of 2D detection networks. The 3D-SSD network model proposed by Luo et al. [4] is a one-stage network for target detection based on depth information, and the prediction is realized with multi-scale mapping, the method has a good improvement in small target detection, as well as excellent performance in depth estimation against images. The Mono3D (Monocular 3D) model proposed by CHEN et al. is based on the improvement of the 3DOP model [5], which generates 3D candidate frames, then scores them with 2D image features and classifies and regresses the candidates with high scores, where the 2D image features are generated based on the information of semantic segmentation, instance segmentation, and location a priori. The practice has shown that the accuracy of the 3DOP model is insufficient relative to the estimation of depth, and some scholars have found that more accurate depth information can be obtained by utilizing parallax estimation. In 2019, LI et al. proposed a Stereo R-CNN model [6], which, relative to the 3D-SSD and the 3DOP model, utilizes the parallax estimation method to obtain more accurate depth information. In point cloud-based 3D target detection methods, many researchers rasterize the point cloud and convert it into voxel form representation to easily handle irregular point cloud data. SIMON et al. [7] proposed Complex-YOLO, a point cloud-based 3D real-time target detection network based on YOLO, where the pose of an object is estimated by adding an imaginary and a real number to the regression network in a specific Euler-Region-Proposal Network (E-RPN). The voxel size setting is a difficult problem to be solved in point cloud voxelization. In 2020, M.Y et al. [8] proposed a hybrid voxel network (HVNet) for mixing point clouds with voxels, which solves this problem by fusing voxel features of different scales at the point-level of the point cloud and projecting them into multiple pseudo-image feature maps. Deng et al. proposed a two-stage voxel-based framework Voxel R-CNN network [9], which first generates region proposals based on a bird's-eye view, and then extracts region-of-interest features directly from voxel features using a designed voxel RoI pool. However, the process of transforming the point cloud into either a voxel or bird's eye view projection map form causes some loss of point cloud data. To minimize or avoid this loss, some scholars have investigated the direct processing of the original point cloud. Qi et al. proposed the PointNet network [10], which ensures that the order of the points in the point cloud remains unchanged during the processing, and the structural connection between the points is preserved completely. Lehner et al. introduced a two-stage model, which consists of two VoxelNet [11] based networks, the Region Proposal Network (RPN) and the Local Refinement Network (LRN), for accurate detection and localization of 3D targets from point cloud data. Although different modal data have obvious effects when used individually in some specific scenarios [12], however, a sensor can only acquire a single modal data in the environment, and a large number of

experiments have proved that there are obvious inherent deficiencies in the environment sensing tasks accomplished by relying on only single modal data. The fusion of LiDAR point cloud data and camera image data with complementary relationships to improve the detection of targets in autonomous driving environment sensing tasks has been the focus of researchers [13]. Xu et al. proposed a PointFusion network structure [14], which is one of the typical pre-fusion structures. PointFusion first generates 2D selection frames on the image with a 2D detector projects the point cloud to the image plane, and selects the appropriate target region in the point cloud using the 2D selection frames region, and the selected 2D image data and the 3D point cloud data are used for feature extraction with ResNet [15] and PointNet networks for feature extraction to predict the location of the target in space. The advantage of this method is that the point cloud is directly input as raw data so that the information is preserved. However, the network is limited to dense point cloud data and is poor for sparse point clouds. Inspired by fusion methods such as PointFusion, Sindagi et al. proposed the MVX-Net network architecture for hybrid fusion based on earlier fusion [16], which utilized the Voxel Net network structure introduced at the time to combine two modal data, RGB images, and point clouds. Wang et al. proposed the F-ConvNet network [17], which, unlike the method of fusing voxelized point clouds with images, consists of a set of view cones proposed to be generated from 2D checkboxes, which are used to group the local point clouds, and the view cone features are formed through feature extraction. The VeloFCN network proposed by Li et al. [18] draws on the experience of 2D image detection by projecting a 3D point cloud to a front view similar to a camera image, and the data obtained from this processing is converted to a 2D image form, and the image is detected with a 2D target detector, but the point cloud in this method has multiple points that overlap in the process of projecting to the front view, resulting in loss of information.

In recent years, with the continuous development of computer vision and deep learning, the detection of targets using deep learning techniques has become a popular research direction. Saranya. K.C et al. [19] proposed YOLO v3 to detect pedestrians, using this method reduces the computational resources and speeds up the computation speed based on guaranteeing the detection accuracy, but there will still be misdetection and omission problems for the targets occluding each other, overlapping and so on. Ren S [20] proposed to utilize neural networks instead of selective search and proposed the concept of anchor frames, the highlight of this method is the integration of subsequent steps such as feature extraction in the same network, which leads to an improvement in the overall performance.

However, the current target detection algorithms are still unable to meet the practical needs, and many problems still need to be solved and improved. On the one hand, most of the autonomous driving scenarios are outdoor open scenarios, containing a large number of static and dynamic targets, and the traffic situation is complex; on the other hand, most of the sensors used for target detection in autonomous driving vehicles are more than three types, and the currently designed target detection algorithms have a single task on the network, and the fused sensor data types are fewer, which will deplete the limited arithmetic power of the vehicle control unit when carrying out multiple tasks at the same time. When performing multiple tasks at the same time, it consumes the limited arithmetic power of the vehicle control unit. Therefore, how to recognize obstacles efficiently and accurately is still of great significance in the field of autonomous driving.

The rest of the paper is organized as follows: Section II describes the theoretical approach and optimization process of the proposed improved model, including the relevant parameter settings of the PointNet++ network and ResNeXt network. Section III examines the accuracy of this paper's model on the KITTI dataset, followed by a comprehensive analysis and discussion of the experimental results. Section IV summarizes the main contributions of this paper's model and proposes future research directions.

II. Proposed Method

A. Pointfusion Network Model Optimization

In this paper, we improve the PointFusion architecture based on the PointFusion architecture, optimize the point cloud feature extraction module and the image feature extraction module, introduce the better performing PointNet++ and ResNeXt neural network structures instead of the PointNet module and the ResNet module, and use only the dense fusion structure to predict the 3D selection box point by point from each input point to the 8 corners (i.e., each vertex) of the spatial offsets, and in this way outputs the 3D prediction frame of the target. The structure is shown in Fig. 1.

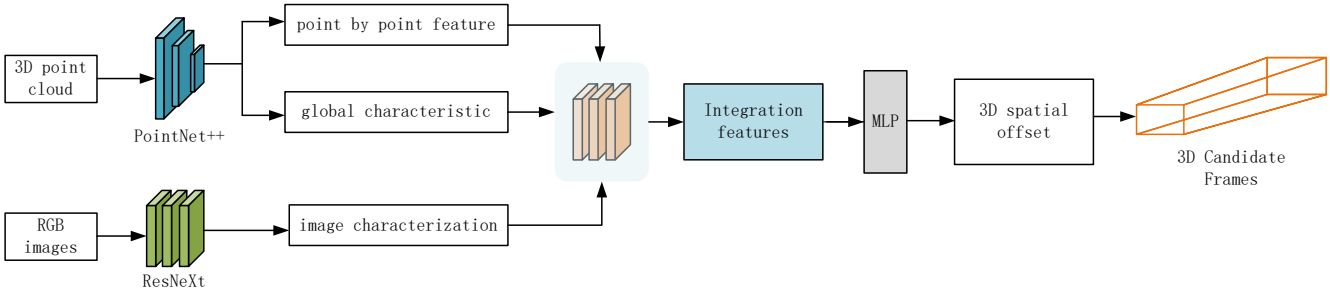


Fig. 1. Improved converged architecture based on PointFusion.

The PointFusion network architecture is a pre-fusion two-stage network structure. The network architecture performs target detection on the image data with a 2D target detector and enhances the point cloud information with the detected 2D target image information. Then the processed RGB image and point cloud image are used as input information, and the corresponding image and point cloud feature extraction network are used to extract features from the input data. Finally, the points in the point cloud of the target area are used as localization anchor points in space and predicted to obtain the 3D candidate frame of the target.

PointFusion uses heterogeneous network architecture to process the input. 3D point cloud data and RGB image data are fed into different branches for feature extraction. A variant model of the PointNet network architecture is used to process the raw point cloud data directly, avoiding the lossy input preprocessing caused by converting the point cloud data into Range maps or voxel forms. However, the PointNet network itself has a poor ability to process sparse point cloud data, which leads to weak detection of sparse point cloud targets in PointFusion itself. In the architecture of PointFusion, on the one hand, the data enhancement of point cloud data is performed with a 2D target detection method before data input, which makes the input point cloud information a cropped dense point cloud, and its network architecture is also trained in the dense region of point cloud, which has a weak adaptive ability to sparse point cloud; on the other hand, when utilizing the point cloud as a spatial localization point, the dense point cloud can be very On the other hand, when using point clouds as spatial localization points, dense point clouds can predict the spatial location of the target, while sparse point clouds cannot predict the spatial location of the target very well using this method.

B. PointNet++ Network

PointNet++ is a deep optimization improvement based on PointNet.PointNet++ is a neural network structure consisting of a combination of multiple individual layers, which applies the PointNet network in feature extraction of the input point set. The PointNet++ network, by calculating the spatial distances between points in space, can follow the change in size between two neighboring layers to extract local features. Moreover, each ensemble sampling layer can be adapted to sample point cloud regions with different densities and can automatically combine feature information at different scales.

PointNet++ employs hierarchical point set feature learning with a hierarchy consisting of multiple ensemble sampling layers. At each ensemble sampling layer, the point cloud sets within a region are first grouped and sampled, then undergo feature extraction and move to the next ensemble sampling layer. In the new ensemble sampling layer, each set of features from the previous layer is combined into a new point cloud element, and the previous operation is repeated until all the features are extracted. The ensemble sampling layer is the core layer and consists of the sampling layer, the grouping layer, and the point network layer. The sampling layer uses the farthest point sampling (FPS) method for sampling, which randomly takes a set of input points as the center of the region, then calculates the distances between other points and that point, and determines the adjacent points based on the distances between the points. The grouping layer constructs the neighborhood of each point cloud based on the distances between spatial points. The PointNet layer focuses on feature extraction of point clouds in the domain, using a simplified version of the PointNet network to extract features from different groups of point clouds in the grouping layer. The PointNet++ structure is shown in Fig. 2.

The input to the ensemble sampling layer is N × (d + C), where d represents the coordinate dimension of the point cloud, C represents the feature vector dimension of each point in the point cloud, and N represents the number of points in each point cloud ensemble. The output of this layer is $N'\times$（d+C'）, where the coordinate dimensions of the points are unchanged, C'represents the eigenvector dimension of each point in the point cloud collection in the input to the next ensemble sampling layer, and N' represents the number of points in each point cloud collection in the input to the next ensemble sampling layer.

A set of points$\{x_1, x_2, \ldots, x_n\}$ is input to the sampling layer, and a set of points $\{x_{i_1}, x_{i_2}, \ldots, x_{i_m}\}$ is found using constant iterative sampling to determine that x_{i_j} is the maximum distance from the set $\{x_{i_1}, x_{i_2}, \ldots, x_{i_{j-1}}\}$. This sampling ensures that all points in the point cloud collection are utilized.

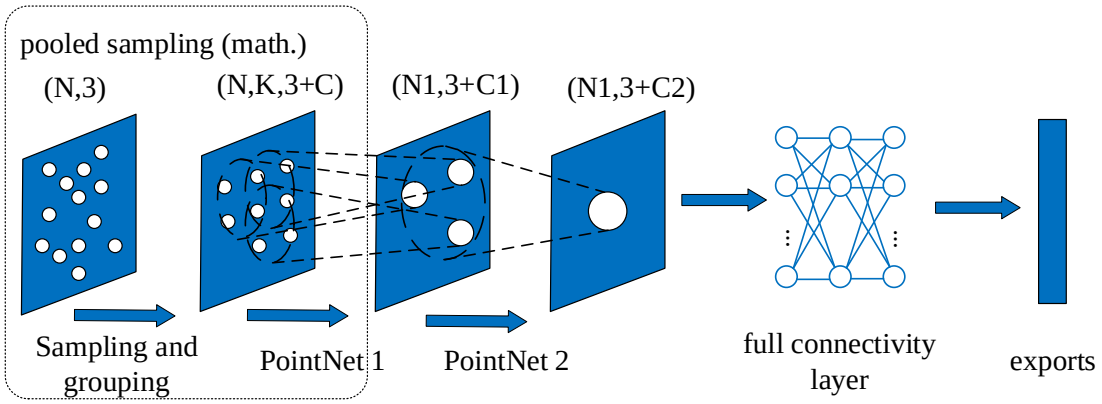


Fig. 2. PointNet++ network structure.

The grouping layer inputs are a point cloud collection and a center of mass collection. The point cloud collection contains the coordinate dimension of the point cloud, the eigenvector dimension of each point, and the number of points in the point cloud; the center of mass collection contains a set of center of mass coordinates. The output of this layer is N'×K×(d+C), with N' denoting the number of neighbors in the set, and K being the center-of-mass points selected in the neighborhood with variable size.

In the PointNet layer, the input is a localized region of N' points of size N'×K×（d+C）, and each localized region in the output is abstracted by its center of mass and local features encoding the neighborhood of the center of mass, with a data size of N'×K×（d+C'）. The PointNet layer is to extract the features of the point cloud data within each neighborhood partitioned by the previous layer, represented by a feature vector of uniform size. The input to this layer is N'×K×（d+C ） and the output is N'×K×（d+C'）.

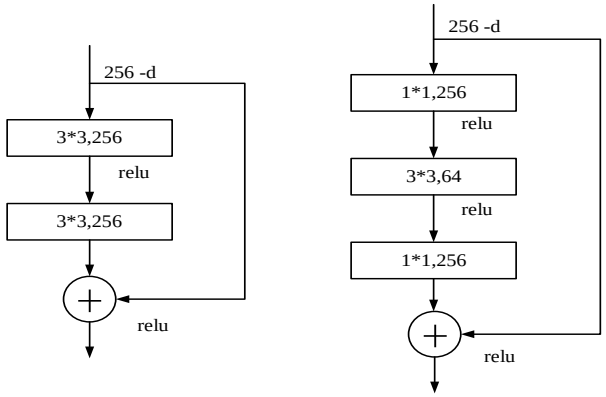
The PointNet++ network structure specifically addresses the problem that when the point cloud is inhomogeneous, the features learned in the dense region may not be suitable for the sparse region, and proposes a multiscale grouping (MSG), where the point cloud data with different densities are grouped into multiple local neighborhoods with different sizes, and these neighborhoods are first feature extracted, and then later the local neighborhood features are extracted by a point cloud feature extraction network to obtain the global features.

C. ResNeXt Network

In the image feature extraction branch, the ResNeXt neural network structure is used, which is developed based on ResNet and combines the experience of classic network structures such as VGG, ResNet, and Inception. It is composed of stacking multiple residual blocks of similar structure and in each block, three convolutional layers are used to realize various transformations from "dimensionality reduction-transformation-upgrading".

The ResNeXt network has been used as the core of many advanced networks for its performance in speed and accuracy in many image vision tasks. Although Resnet can be subdivided into 18, 34, 50, and 101 layers, the main body of the network structure is the same. The starting input of the Resnet network structure is a convolutional layer with 7*7 convolutional kernels, followed by a maximum pooling downsampling layer with 3*3 convolutional kernels, followed by a residual structure with multiple blocks stacked in layers conv2~conv5, and the last layer is an average pooling downsampling layer with 3*3 pooling kernels. The first input of the Resnet network structure is a convolutional layer with 7*7 convolutional kernels, followed by a maximally pooled downsampling layer with 3*3 convolutional kernels, followed by conv2~conv5 layers which are residual structures stacked with multiple blocks, and the last layer is an average pooled downsampling and fully-connected layer, and the result is outputted by softmax. The structure of the ResNeXt network is shown in Fig. 3.

Fig. 3(a) shows the residual module of ResNet-18/34, and Fig. 3(b) shows the residual module of ResNet-50/101/152. Each module consists of a main branch and a shortcut branch, which are output after the convolution operation of the main branch on one hand, and the shortcut branch is directly the input, and then the results of the two branches are summed up and processed by the activation function, so the size and depth of the feature matrices of the shortcut branch and the output of the main branch should be the same. The size of the kernel of convolution and the input/output of the residual module of ResNet-18/34 are the same at each layer, and the size of the kernel of convolution and the input/output of ResNet-50/101/152 is the same, while the size of the convolution kernel of the convolutional layers in the residual module of ResNet-50/101/152 is not consistent. In Fig. 3(b) 1*1 convolution kernel is used for downscaling and upscaling, and it can be seen from the figure that the input of 256 channels in the first layer is output down to 64 channels, while the 64 channels input in the second layer is increased to 256 channels by the output of the third layer. In Fig. 3(b), this structure serves to save more number of convolution kernels. Taking the input of 256 channels as an example, after calculation, there are 1179648 convolutional kernels in Fig. 3(a) and 69632 convolutional kernels in Fig. 3(b), compared to the reduction of 111,016 convolutional kernels, which saves resources and improves the efficiency at the same time.



(a) The residual module of ResNet-18/34 (b) The residual module of ResNet-50/101/152

Fig. 3. PointNet++ network structure.

The improved network uses a dense fusion network structure. The network uses the input spatial points as spatial anchors and predicts the spatial positional offsets of each spatial point to each vertex of the neighboring preselected boxes. An unsupervised approach is utilized to predict a score for each spatial point, and the point with a high prediction score receives a high confidence level. The high confidence point is used as the final prediction point. The unsupervised scoring loss function is:

$$L_{score} = \frac{1}{N}\sum\left(L^{i}_{offset} \cdot c_i - w \cdot log(c_i)\right) + L_{stn} \quad (1)$$

where, w is the weight coefficient, c_i is the confidence level, and L^{i}_{offset} is the spatial angular offset loss at the first spatial point. L_{stn} is the spatial transformation regularization loss. The loss function of the dense fusion network is:

$$L_{dense} = \frac{1}{N}\sum_i smothL\left(x^{i^*}_{offset,} x^{i}_{offset}\right) + L_{score} + L_{stn} \quad (2)$$

where, N is the number of input points, $x_{offset}^{i^*}$ is the offset between the true checkbox vertex position and the ith spatial point, and x_{offset}^{i} is the spatial position offset between the predicted checkbox vertex position and the ith spatial point.

III. Experimental Results

A. Experimental Parameter setting

1) PointNet++ parameter settings: This experiment uses the MSG network (Multi-Scale Network) of the PointNet++ network to set up three ensemble sampling layers, and some of the parameters of each ensemble sampling layer are shown in Table I.

TABLE I. Parameter Settings for each Ensemble Sampling Layer

Ensemble sampling layer	Number of input points	Number of sampling points	Sampling radius	characteristic channel
SA1	1024	512	[0.1,0.2,0.4]	-
SA2	512	128	[0.2,0.4,0.8]	320
SA3	128	-	-	640+3

During model training, batch is set to 24 and epoch is set to 200. The MSG model has three layers SA1, SA2, and SA3 used for point cloud data feature extraction.

1) Ensemble Sampling Layer: The MSG network model has three layers, sa1, sa2, and sa3, which are used for point cloud data feature extraction. The output of this layer mainly has seven-dimensional features. The first-dimensional feature is the number of sampled points; the second-dimensional feature is the radius size, a total of three radius parameters are set, [0.1,0.2,0.4] for the first layer, [0.2,0.4,0.8] for the second layer; the third-dimensional feature is the number of points in the group corresponding to the radius; the fourth-dimensional feature is the number of features or the channel size of the input points; and the fifth to the seventh-dimensional features are corresponding to the three radiuses respectively. The fifth to seventh-dimensional features are the number of feature dimensions corresponding to each of the three radii.

2) ResNeXt parameter settings: The ResNeXt101 network is used in this experiment, where the batch is set to 32, num_workers is set to 4, and epoch is set to 100. The optimizer is ADAM, and the learning rate is 0.001. The parameters of conv2~conv5 of the ResNeXt residual structure are set to {3, 4, 23, 3}.

B. Analysis and Discussion of Results

To test the model precision on the KITTI dataset, to better compare with other target detection methods, the experiment uses the evaluation method provided by the official KITTI dataset, sets the thresholds for cars, cyclists, and pedestrians at 0.7 respectively, and divides the test set among the KITTI dataset into three modes: simple, medium and difficult. The accuracy and recall in the three modes are calculated to obtain the average precision as the evaluation performance metric. According to the method of PointFusion, Fast-Rcnn is used as a 2D target detector with ResNeXt-101 to extract the input image feature information and average over the feature map positions. For each input 2D frame, the image is cropped and resized to 224 × 224, and up to 400 3D point clouds are randomly sampled as input for training and evaluation.

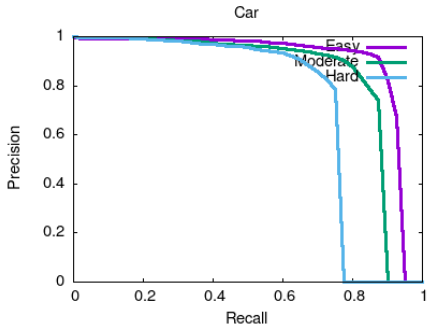
The P-R curves obtained from the experiment are shown in Fig. 4, and the final experimental AP result statistics are shown in Table II.

Table II results show the improved fusion network model detection results. The evaluation results for 3D detection in both evaluation criteria show that the detection accuracy for the automobile class is improved to 79.97% in the simple model; the detection accuracy for pedestrians is still below fifty percent. Some of the visualized test results are shown in Fig. 5.

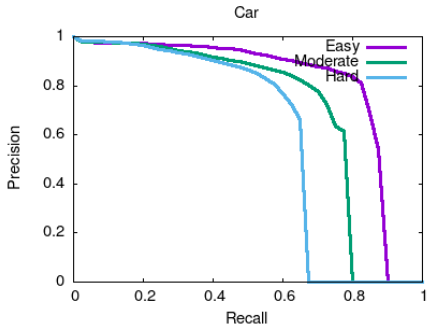
TABLE II. Model Test Results

objectives	AP_{BEV}(%)			AP_{3D}(%)		
	Easy	**Medium**	**Difficult**	**Easy**	**Medium**	**Difficult**
vehicle	89.10	83.14	71.41	79.97	69.13	58.57
pedestrians	56.52	47.80	43.63	48.65	40.11	35.99
cyclist	71.92	57.99	51.55	65.51	51.33	45.05

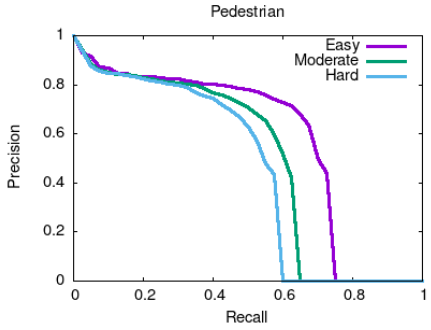
The PointFusion model before and after improvement is compared with other state-of-the-art models in Table III. The experimental results show that the improved PointFusion-based model proposed in this paper performs well for the target detection task in general and improves the performance compared to the original model. Among them, the highest improvement is 6.13% in the medium mode, 2.05%, and 5.3% in the easy and difficult modes, respectively. In the medium and difficult modes, the proportion of accuracy improvement before the improvement relative to that after the improvement is larger, indicating that the generalization ability of the improved model to the original model has been improved. The method proposed in this paper reaches or approaches the results of the current state-of-the-art methods in all three difficulty modes, but there is an obvious gap in accuracy compared to AOVD and F-PointNets in the difficulty level. The comparison reveals that both the original model and the improved model based on this paper have poor performance in the difficulty modes relative to the above two models. This indicates that the detection of large occluded targets and small targets needs to be improved.



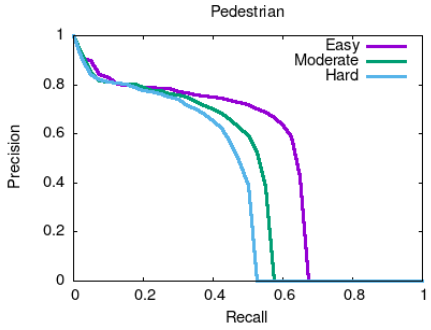
(a) PR curve for Bev_Car



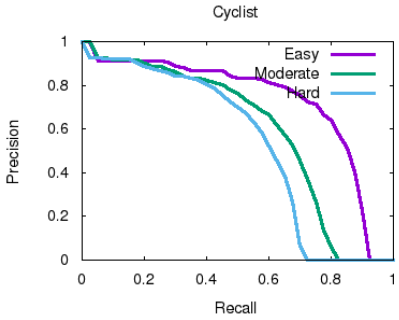
(b) PR curve for 3D_Car



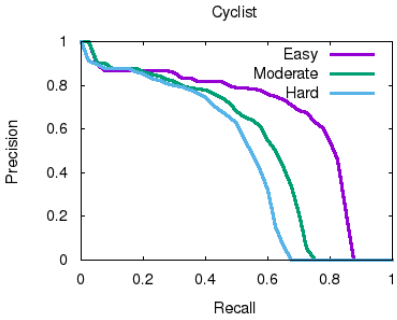
(d) PR curve for 3D_Pedestrian



(c) PR curve for Bev_Pedestrian

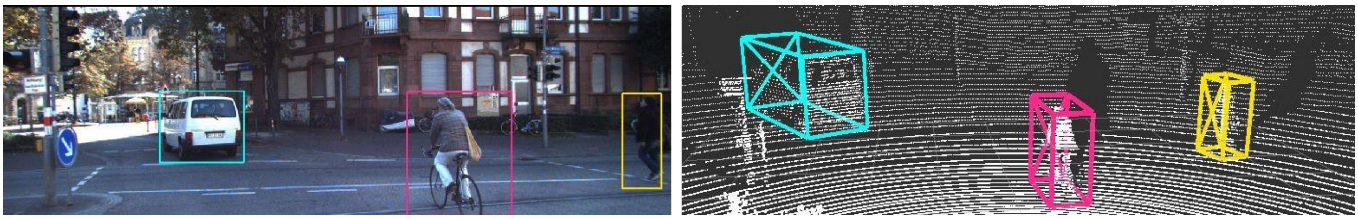


(e) PR curve for Bev_Cyclist

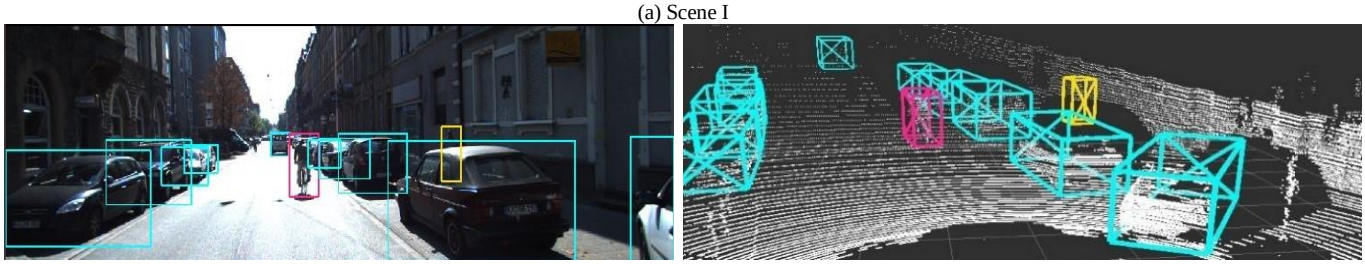


(f) PR curve for 3D_Cyclist

Fig. 4. Improved P-R curves of Disp R-CNN on bird's eye view (Bev) and 3D detection tasks.



(a) Scene I



(b) Scene II

Fig. 5. Visualization of some of the test results.

TABLE III. MODEL TEST RESULTS

Network model	APBEV(%)			AP3D(%)		
	Easy	Medium	Difficult	Easy	Medium	Difficult
F-PointNets	91.17	84.67	74.77	82.19	69.79	60.59
AOVD	89.75	84.95	78.32	76.39	66.47	60.23
MV3D	86.62	78.93	69.80	74.97	63.63	54.00
PointFusion	-	-	-	77.92	63.00	53.27
paper model	89.10	83.14	71.41	79.97	69.13	58.57

IV. CONCLUSION

In this paper, we propose to optimize the PointFusion architecture based on the improved PointFusion fusion algorithm by replacing the PointNet module and ResNet module in the point cloud feature extraction module and image feature extraction module with PointNet++ and ResNeXt network structures, respectively, to achieve more accurate 3D target detection. The key difference with the existing methods is that we consider the problems of poor generalization ability of the PointNet network and poor feature extraction in sparse regions of the point cloud. We feature extract image and point cloud information from the input data and use a dense fusion approach to obtain the final prediction, which effectively reduces the number of hyperparameters in the image feature extraction module and improves the computing speed without affecting the accuracy of the final prediction. Finally, we used experiments on the KITTI dataset to demonstrate the effectiveness of the method, and compared with the original prediction model, the model in this paper has the highest improvement among the medium modes, reaching 6.13%. Most of the current neural networks for autonomous driving perception tasks are designed for a single task, which generates a large amount of arithmetic power consumption for the in-vehicle main controller that works on multiple tasks at the same time. How to integrate the networks of different tasks into one main network, optimizing the arithmetic power, and improve the cooperative work of each task will be a future research direction.

REFERENCES

[1] Balasubramaniam and S. Pasricha, "Object detection in autonomous vehicles: Status and open challenges," arXiv preprint arXiv:2201.07706, 2022.

[2] N. M. A. A. Dazlee, S. A. Khalil, S. Abdul-Rahman, and S. Mutalib, "Object detection for autonomous vehicles with sensor-based technology using yolo," International Journal of Intelligent Systems and Applications in Engineering, vol. 10, no. 1, pp. 129-134, 2022.

[3] Liu Z, Wu Z, Tóth R. Smoke: Single-stage monocular 3d object detection via keypoint estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 996-997.

[4] Luo Q ,Ma H ,Tang L , et al. 3D-SSD: Learning hierarchical features from RGB-D images for amodal 3D object detection[J]. Neurocomputing,2020,378.

[5] XIAOZHI CHEN, KAUSTAV KUNDU, ZIYU ZHANG, et al. Monocular 3D Object Detection for Autonomous Driving[C]. //29th IEEE Conference on Computer Vision and Pattern Recognition: 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 26 June – 1 July 2016, Las Vegas, Nevada.:Institute of Electrical and Electronics Engineers, 2016:2147-2156.

[6] PEILIANG LI, XIAOZHI CHEN, SHAOJIE SHEN. Stereo R-CNN based 3D Object Detection for Autonomous Driving[C]. //2019 IEEE/CVP Conference on Computer Vision and Pattern Recognition: CVPR 2019, Long Beach, California, USA, 15-20 June 2019, [v.11].:Institute of Electrical and Electronics Engineers, 2019:7636-7644.

[7] MARTIN SIMON, STEFAN MILZ, KARL AMENDE, et al. Complex-YOLO: An Euler-Region-Proposal for Real-Time 3D Object Detection on Point Clouds[C]. //Computer Vision - ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part I.:Springer, 2019:197-209.

[8] M. Y ,S. X ,T. C . HVNet: Hybrid Voxel Network for LiDAR Based 3D Object Detection[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2020.

[9] Deng J, Shi S, Li P, et al. Voxel r-cnn: Towards high performance voxel-based 3d object detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(2): 1201-1209.

[10] Qi C R, Su H, Mo K, et al. Pointnet: Deep learning on point sets for 3d classification and segment-ation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 652-660.

[11] Lehner J, Mitterecker A, Adler T, et al. Patch Refinement--Localized 3D Object Detection[J]. arXiv preprint arXiv:1910.04093, 2019.

[12] LI, YING, MA, LINGFEI, ZHONG, ZILONG, et al. Deep Learning for LiDAR Point Clouds in Autonomous Driving: A Review[J]. IEEE transactions on neural networks and learning systems,2021,32(8):3412-3432. DOI:10.1109/TNNLS.2020.3015992.

[13] CUI, YAODONG, CHEN, REN, CHU, WENBO, et al. Deep Learning for Image and Point Cloud Fusion in Autonomous Driving: A Review[J]. 2022,23(2):722-739. DOI:10.1109/TITS.2020.3023541.

[14] Xu D, Anguelov D, Jain A. Pointfusion: Deep sensor fusion for 3d bounding box estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 244-253.

[15] KAIMING HE, XIANGYU ZHANG, SHAOQING REN, et al. Deep Residual Learning for Image Recognition[C]. //29th IEEE Conference on Computer Vision and Pattern Recognition: 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 26 June – 1 July 2016, Las Vegas, Nevada.:Institute of Electrical and Electronics Engineers, 2016:770-778.

[16] Sindagi V A, Zhou Y, Tuzel O. Mvx-net: Multimodal voxelnet for 3d object detection[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 7276-7282.

[17] Wang Z, Jia K. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 1742-1749.

[18] Li B, Zhang T, Xia T. Vehicle detection from 3d lidar using fully convolutional network[J]. arXiv preprint arXiv:1608.07916, 2016.

[19] C K S ,Thangavelu A . Vulnerable Road User Detection using YOLO v3[J]. International Journal of Advanced Computer Science and Applications (IJACSA),2019,10(12).

[20] REN, SHAOQING, HE, KAIMING, GIRSHICK, ROSS, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149.DOI:10.1109/TPAMI.2016.2577031.

Beyond the Norm: A Modified VGG-16 Model for COVID-19 Detection

Shimja M, K.Kartheeban
Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Tamil Nadu, India

***Abstract*—The outbreak of Coronavirus Disease 2019 (COVID-19) in the initial days of December 2019 has severely harmed human health and the world's overall condition. There are currently five million instances that have been confirmed, and the unique virus is continuing spreading quickly throughout the entire world. The manual Reverse Transcription-Polymerase Chain Reaction (RT-PCR) test is time-consuming and difficult, and many hospitals throughout the world do not yet have an adequate number of testing kits. Designing an automated and early diagnosis system that can deliver quick decisions and significantly lower diagnosis error is therefore crucial. Recent advances in emerging Deep Learning (DL) algorithms and emerging Artificial Intelligence (AI) approaches have made the chest X-ray images a viable option for early COVID-19 screening. For visual image analysis, CNNs are the most often utilized class of deep learning neural networks. At the core of CNN is a multi-layered neural network that offers solutions, particularly for the analysis, classification, and recognition of videos and images. This paper proposes a modified VGG-16 model for detection of COVID-19 infection from chest X-ray images. The analysis has been made among the model by considering some important parameters such as accuracy, precision and recall. The model has been validated on publicly available chest X-ray images. The best performance is obtained by the proposed model with an accuracy of 97.94%.**

***Keywords*—Covid-19; coronavirus; artificial intelligence; deep learning; transfer learning; VGG-16; performance metrics**

I. Introduction

Since the beginning of December 2019, the Coronavirus Disease 2019 (COVID-19) outbreak has put enormous pressure on the entire world [1]. According to the World Health Organization (WHO), more than five million people have been infected globally to date, and there have been about three lakh confirmed cases of death. A respiratory disease, COVID-19 is brought on by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and is characterized by symptoms such as fever, myalgia, dry cough, headache, sore throat, and chest pain [2]. The infected person may not exhibit all symptoms for up to 14 days.

A severe respiratory condition known as COVID-19 can be cured without the use of antibiotics in affected individuals. Chronic medical diseases like diabetes, chronic respiratory illnesses, and cardiovascular issues increase the chance of spreading this virus. According to the WHO, the COVID-19 symptoms, which include fever, tiredness, a dry cough, shortness of breath, aches, pains, and sore throat, are similar to those of the ordinary flu [3]. Because of these common symptoms, it is difficult to detect the virus in its early stages. There is no chance that antibiotics, which are used to treat bacterial or fungal infections, will be able to halt this as it is a virus.

In medical, scientific, and healthcare laboratories, the coronavirus has been identified using a variety of diagnostic techniques. Indirect methods evaluate antibodies against the virus in a host that has been exposed, whereas direct tests identify the contamination directly by detecting the viral RNA. During a pandemic, a clinical test method should be accurate and sensitive enough to quickly make the right clinical recommendations. The various diagnostic methods for the covid-19 detection are shown in Fig. 1.

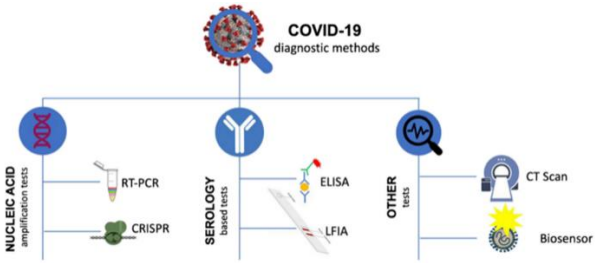


Fig. 1. Diagnostic method for Covid-19 detection [4].

The most dependable and popular technique for COVID-19 detection among the several methods available is RT-PCR [5]. There is sometimes a lack of supplies during pandemic events, including PCR kits. Therefore, having a variety of diagnostic method choices is crucial. Equally crucial are other testing methods and accessories that may be produced locally, even on a small scale. Those platforms would be suitable in environments with restricted resources as well. Because the currently used RT-PCR techniques are expensive, many nations, especially those with poor incomes, cannot afford enough COVID-19 tests to screen a broader population. Screening asymptomatic people throughout the incubation phase still has significant gaps. It is also difficult to accurately predict live virus shedding in healed patients in order to make de-isolation decisions. Several COVID-19 test kits have been recalled in various nations due to questionable quality. The rigorous evaluation of the diagnostic precision of the recently introduced SARS-CoV-2 assays is hampered by the lack of a recognized reference standard, the use of various sample collection and preparation techniques, and a flawed understanding of viral interplay across the time phase of infection.

There is currently no specific medication or treatment for this illness. However, RT-PCR is the technique that is most frequently used to diagnose COVID-19. It has recently been

discovered that medical imaging methods like X-rays and Computed Tomography (CT) are essential for testing COVID-19 instances. The chest radiography images (chest X-ray or CT images) have been extensively investigated since the virus typically infects the lungs [6]. Radiologists manually evaluate these images to look for any visual signs of COVID-19 infection. These visual cues can be used as an alternative screening technique for infected patients.

The WHO has approved the RT-PCR technique as a coronavirus testing method, in which short DNA or RNA sequences are evaluated and replicated or amplified. But it takes more than one test to totally rule out coronavirus in some people. WHO guidelines for laboratory testing state that negative results do not necessarily rule out the possibility that the person is virus-infected. The lack of screening workstations and testing kits to identify COVID-19 makes it incredibly challenging for medical staff and professionals to address the issue. It is incredibly challenging for medical practitioners in this situation to promptly and precisely detect possible COVID-19 cases. Numerous tests are also necessary because of the cases' exponential growth in order to fully comprehend the circumstances and come up with the finest decisions.

Despite the availability of numerous imaging modalities, chest radiography is believed to have a low sensitivity for major clinical findings. Doctors frequently use X-ray imaging technologies to identify pneumonia because they are an essential part of healthcare systems all over the world. The use of chest X-ray equipment is time-consuming in the absence of screening workbenches and kits notwithstanding their simplicity in locating COVID-19 cases. Additionally, there may be situations where individuals receive imaging for a different reason and the results of their scans point to COVID-19. The findings from the chest X-ray images strongly suggest that even in the early stages of COVID-19, the effect can be seen in the lungs, particularly in the lower lobes and posterior segments, with peripheral and subpleural distribution. The lesions diffuse more and more as time passes. The biggest problem, though, is that analyzing each chest X-ray image and determining what information is most important takes a lot of time and the presence of medical professionals. Therefore, to assist in the detection of COVID-19 cases using chest X-ray images, medical personnel need computer assistance.

Although the traditional diagnosing process has sped up somewhat, it still puts medical professionals at great risk. Additionally, it is expensive, and there are only a few diagnostic test kits available. On the other hand, screening based on medical imaging techniques (such X-ray and CT) is often safe, quick, and available. X-ray imaging, as opposed to CT imaging, has been utilized extensively for COVID-19 screening because it is less expensive, requires less imaging time, and is more readily accessible even in rural areas than CT imaging. The larger-scale visual examination of X-ray images by radiologists is time-consuming, labor-intensive, and could result in an incorrect diagnosis because of ignorance of the virus-infected regions. Therefore, the development of automated approaches to acquire a quicker and more precise COVID-19 diagnosis is highly necessary. The most recent automated methods sought to lessen the labor of radiologists while enhancing the power of X-ray imaging using cutting-edge Artificial Intelligence (AI) technologies [7]. Convolutional Neural Networks (CNN), a type of DL model, in particular, have proven to be more efficient than conventional AI techniques and are frequently used to analyze a variety of medical images. CNN has recently been used to successfully identify COVID-19 in chest X-ray images. The major contribution of this paper includes:

- An effective detection model for Covid-19.
- A modified VGG-16 model for Covid-19 detection from chest X-ray images.
- Classification of chest X-ray images into Covid-19 and Normal.

This paper seems to have the following format. The details of the currently used COVID-19 detection techniques are covered in Section II. The methodology is covered in Section III. The experimental findings and conclusions are presented in Sections IV and V respectively. Finally, the future works are presented in Section VI.

II. Literature Review

Saul Calderon-Ramirez et al. [8] introduced COVID-19 detection from chest X-ray images using semi- supervised deep learning methods. Using a semi-supervised deep learning framework built on the Mix Match architecture, chest X-rays are classified into Covid-19, pneumonia, and healthy cases in this paper. When there is a lack of high-quality labeled data, performance levels for Covid-19 identification can be enhanced with the use of the semi-supervised framework. Shamima Akter et al. [9] introduced COVID-19 detection using deep learning approaches. The paper suggested a deep learning-based automated classification model based on a convolutional neural network that exhibits a high COVID-19 detection rate. The COVID-19 symptoms were initially identified utilizing the dataset by applying eleven pre-existing CNN models. A confusion matrix was used to demonstrate how well the models performed. Zehra Karhan and Fuat Akal [10] discussed Covid-19 Classification Using Deep Learning in Chest X-Ray Images. The ResNet50 model was utilized for Covid-19 classification from chest X-ray images. Artificial intelligence allows for fast analysis of chest X-ray images and identification of diseased individuals. It can also be applied when RT-PCR testing and other options are insufficient. Sami Bourouis et al. [11] introduced Bayesian Learning of Shifted-Scaled Dirichlet Mixture Models for Covid-19 detection. In order to distinguish between individuals who are either negative or positive for particular types of viruses and pneumonia, this research introduced a novel statistical framework. The success and reliability of this mixing model in recent image processing applications is encouraging. The established Bayesian framework has the benefit of accounting for uncertainty to precisely estimate the model parameters and the capability to address the overfitting issue. Lightweight deep learning models for detecting COVID-19 from chest X-ray images were discussed by Stefanos Karakanis and Georgios Leontidis [12]. In this paper, a new approach was proposed to detect COVID-19 via exploiting a conditional generative adversarial network to generate synthetic images

for augmenting the limited amount of data available. The study focused on both binary classification for COVID-19 vs Normal cases and multi-classification that includes a third class for bacterial pneumonia.

In study [13], Covid-19 detection using majority voting classifiers are used. This work introduced an automated COVID screening (ACoS) method that identifies healthy, suspicious, and COVID-19-infected patients using radiomic texture descriptors taken from CXR images. The proposed system employs a majority vote-based classifier ensemble comprising five benchmark supervised classification algorithms in a two-phase classification approach (normal vs. abnormal and COVID-19 vs. pneumonia). Tulin Ozturk et al. The study in [14] proposed an automated detection of COVID-19 cases using deep neural networks with X-ray images. This paper presented a new model for automatically detecting COVID-19 from raw chest X-ray images. The suggested approach is designed to deliver precise diagnostics for multi-class classification and binary classification. The DarkNet model was used as a classifier for the real-time You Only Look Once (YOLO) object identification system. Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms was proposed by Morteza Heidari et al. [15]. The purpose of this study is to create and evaluate a new computer-aided diagnostic method using chest X-ray images to identify pneumonia caused by the coronavirus. A histogram equalization technique and a bilateral low-pass filter are used to process the original image in the first two image preparation steps of the CAD method. Then, a pseudo-color image is created using the original image, two filtered images, and the original image. This image is fed into three input channels of a transfer learning-based convolutional neural network (CNN) model. Muhammad Ilyas et al. [16] discussed various methods used for Covid-19 detection using Artificial Intelligence. This paper outlines the difficulties we are now encountering as well as the various methods utilized to detect COVID-19. To stop the spread of the virus through contact, an automatic detection system must be created. For the detection of COVID-19, a number of deep learning architectures, including ResNet, Inception, Googlenet, etc., are being used. In [17], proposed an automatic Covid-19 detection method. To assess and contrast the created models, three distinct experiments are conducted in accordance with three preprocessing approaches. The objective is to assess the effects of data preprocessing on the outcomes and how well they can be explained. Similarly, a critical evaluation of various variability issues that could threaten the system and its impacts is carried out.

Shayan Hassantabar et al. [18] proposed a diagnosis and detection of infected tissue of COVID-19 patients based on lung X-ray image using convolutional neural network approaches. Three deep learning-based techniques were employed in this study to identify and diagnose COVID-19 patients using X-ray images of the lungs. A convolutional neural network (CNN) technique was introduced, which uses the lung images and deep neural network (DNN) methods based on the fractal characteristic of images, for the diagnosis of the condition. In study [19], an automatic detection of COVID-19 infection using chest X-Ray images through transfer learning was proposed. It employs various convolutional neural network (CNN) architectures that have been trained on ImageNet and modifies them to function as feature extractors for the X-ray images. Then, the CNNs are integrated with consolidated machine learning techniques, including support vector machine, k-Nearest Neighbor, Bayes, Random Forest, and multilayer perceptron (MLP). Govardhan Jain et al. [20] proposed a deep learning approach to detect Covid-19 coronavirus with X-Ray images. Using accessible resources and cutting-edge deep learning techniques, an alternative diagnostic tool to find COVID-19 patients is suggested in this work. The proposed approach is put into practice in four stages: data augmentation, preprocessing, creating stage-I and stage-II deep network models. In order to improve the generalization of the model and avoid the model from overfitting, this work used web resources of 1215 images that have been improved further by data augmentation techniques, bringing the total number of images in the dataset to 1832. Aras M. Ismael and Abdulkadir Şengür [21] introduced deep learning based covid-19 detection method. In order to categorize COVID-19 and normal (healthy) chest X-ray images, deep-learning-based approaches, including deep feature extraction, fine-tuning of pretrained convolutional neural networks (CNN), and end-to-end training of a constructed CNN model, have been used in this study. Pretrained deep CNN models (ResNet18, ResNet50, ResNet101, VGG16, and VGG19) were utilized for deep feature extraction. The Support Vector Machines (SVM) classifier was used to categorize the deep features using a variety of kernel functions, including linear, quadratic, cubic, and gaussian. The fine-tuning process also made use of the previously mentioned pre-trained deep CNN models. In [22], a novel CNN model called CoroDet for automatic detection of COVID-19 by using raw chest X-ray and CT scan images has been proposed. With the development of CoroDet, an accurate diagnostic tool for COVID and Normal, COVID, Normal, and Non-COVID Pneumonia, and 4 Class Classification was available (COVID, Normal, non-COVID viral pneumonia, and non-COVID bacterial pneumonia).

While deep learning methodologies and chest X-ray images have been widely employed in COVID-19 detection research, there remains a crucial requirement to evaluate the adaptability and resilience of these models across diverse imaging scenarios and population demographics. The absence of investigations into real-world variables, including demographic variances, imaging device fluctuations, and variations in data acquisition settings, raises legitimate concerns about the external reliability and validity of existing models. This underscores the urgency for more comprehensive and interdisciplinary research endeavors to enhance the practicality of these diagnostic tools.

The block diagram of the proposed method is given in Fig. 2. The proposed method is carried out using COVID-19 X-ray Dataset [23]. The dataset contains two folders: test and train. The dataset includes NORMAL and COVID-19 images with .jpg format. The models are trained and tested using the complete dataset. 30% of the dataset's data were used for testing, and the remaining 70% were used for training. The input images are preprocessed before applying to the pre-

trained model. In order to ensure numerical stability in CNN systems, normalization of data is a crucial step. Normalization increases the likelihood that a CNN model will learn more quickly and that the gradient descent will be stable. As a result, in this study, the input image pixel values have been normalized to fall between 0 and 1. The grayscale images used in the datasets under consideration were rescaled by multiplying the pixel values by 1/255. The CNN models have demonstrated to perform better on larger datasets and need a significant amount of data for optimal training. The dataset only contains a very small number of training X-ray images. Since it is difficult to gather medical data, this has been a major concern when doing analysis of medical images using DL algorithms. Data augmentation approaches, which enable to increase the number of images via a set of modifications while keeping class labels, have been frequently used to address this issue. Additionally, augmentation makes the images more variable and acts as a dataset. The preprocessing includes normalization and augmentation. The preprocessed images are used to train the modified VGG-16 model. Finally, the model classifies the chest X-ray images. The sample and augmented chest X-ray images are shown in Fig. 3 and Fig. 4 respectively.

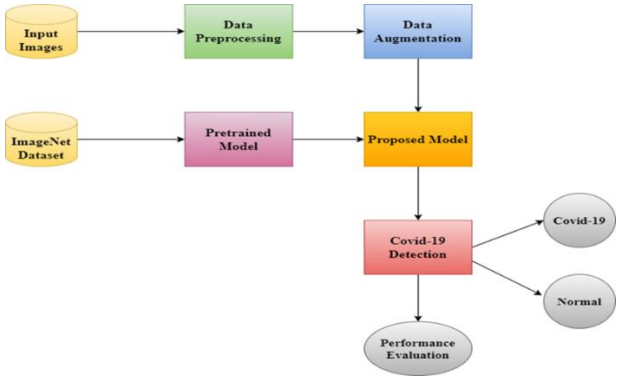


Fig. 2. Block diagram of the proposed method.

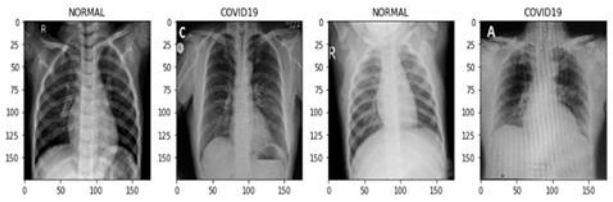


Fig. 3. Sample CXR images.

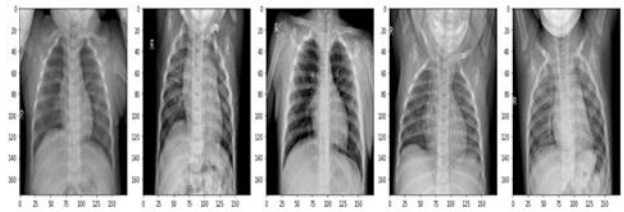


Fig. 4. Augmented CXR images.

III. METHODOLOGY

For the study of visual images, CNNs are the most widely used class of neural networks in deep learning. A multi-layered neural network that provides solutions, notably for the analysis, classification, and recognition of images and videos, is the fundamental aspect of CNN. CNN is made up of different layers. They are the convolution layer, the pooling layer and the fully connected layer. In essence, the convolutional and pooling layers provide the model's learning while the full connection layer handles categorization. The convolutional layer is the central part of CNN architectures. The feature map is created by applying high- and low-level filters to the input image. The pooling layer is often positioned between the convoluted layers. Its main objective is to reduce the size of the feature map in order to reduce the amount of computational resources required to construct the model. It also efficiently trains the model by removing its dominating and invariant characteristics. Although there are many different pooling methods, the maximum and average pooling layers are the most often used ones. Each neuron in the entire connection layer is connected to every neuron from the layer above. Depending on its topology, CNN architecture may have fewer or more complete connection layers. The output layer comes after the last full connection layer. It is now possible to create output distributions using SoftMax regression by obtaining probability distributions for the output classes in classification experiments.

Transfer learning is a machine learning method that applies knowledge that a CNN has learned from a batch of related data to a separate but related problem. The fundamental element of transfer learning is that knowledge can be acquired by transferring it from one related task to another. With pretrained models, the transfer learning design method is widely applied. These pretrained models are based on deep convolutional neural networks. Initial CNN training for a classification problem using sizable training datasets is required for this deep learning technique. Since a CNN model can learn to extract significant components of the image, the availability of data for initial training is a vital element of efficient training.

A. VGG-16

The Visual Geometric Group (VGG) Network is a simple model. Its integration with CNN models is common due to its deeper structure, which is followed by layers of associated double- or triple-convolution layers, which is the most noticeable difference between it and prior models. The model architecture is given in Fig. 5. It is possible to utilize the model to extract pertinent features from suitable new images. The ImageNet dataset has the ability to extract attributes from images, including brand-new ones that might not yet exist or that might be found in dataset categories that are completely unrelated to one another [24]. Therefore, it is advantageous to use pretrained models as effective feature removers.

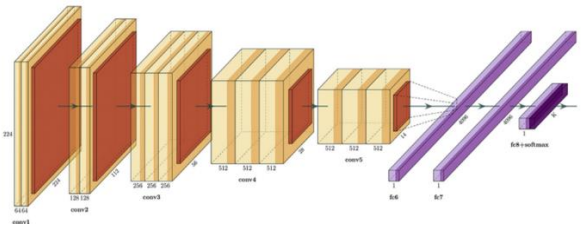


Fig. 5. VGG-16 architecture.

The VGG16 design uses three convolution filters with a total of 13 convolution layers for feature extraction. Each ReLU layer has a maximum pooling layer for sampling and is followed by a convolution layer. Its final classification layer is composed of 1000 units that are equivalent to the image categories in the ImageNet database, and it has three fully connected classification layers, two of which are hidden layers. This design simulates a larger filter while maintaining the benefits of smaller filter sizes. VggNet has been shown to perform better with fewer parameters. Furthermore, two ReLU layers rather than only one were used for the two convolution layers. The depth of the volumes increases as the number of filters increases since the convolution and partnering layers reduce the spatial size of the input volumes in each layer.

B. Modified VGG-16

The modified VGG-16 design reduces the number of parameters by decreasing the network depth in contrast to the original VGG-16 architecture in order to avoid problems with under and overfitting during training. The original architecture of the VGG16 convolutional network was preserved by doing feature extraction with two consecutive small convolutional kernels as opposed to a single large one. This expedites training while retaining the depth of the network by reducing the number of parameters while maintaining the VGG16 perceptual effects.

The input image size was altered initially, after which the hidden layer was divided into five blocks, each of which contained two convolutional layers and a pooling layer. Using 32 randomly generated 3x3 convolutional kernels, each convolutional layer extracted features, while the pooling layer compressed the image. Convolutional kernels in blocks 3 through 5 were all the same size (3x3), but there were, respectively, 64, 128, and 64 kernels in each block. When compared to VGG16, the number of parameters needed was decreased by decreasing the convolutional kernels. After that, the pooling layer reduced the size of the image, the flattened layer reduced feature mappings to one dimension, and three fully connected layers combined output features into two classes. Finally, the performance of the model is evaluated based on different metrics.

C. Performance Evaluation

Accuracy is one of the measures that is most frequently used while classifying data. Accuracy is a parameter measuring how well a model performs across all classes. When all classes are treated equally, it is advantageous. It is calculated by dividing the total number of predictions by the number of predictions that came true. The formula below can be used to assess a model's accuracy.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

The ratio of Positive samples that were correctly identified is used to calculate precision. How precisely the model can classify a sample as positive depends on its precision. The capacity to correctly classify all positive samples as positive while avoiding incorrectly labeling a negative sample as positive is known as precision.

$$Precision = \frac{(TP)}{(TP+FP)} \quad (2)$$

By dividing the total number of Positive samples by the recall rate, the percentage of Positive samples that were correctly identified as Positive is obtained. Recall quantifies how well a model can find Positive samples. The recall increases as more positive samples are found.

$$Recall = \frac{(TP)}{(TP+FN)} \quad (3)$$

IV. Results and Discussion

A. Hardware and Software Setup

The system consists of a single NVIDIA GeForce GTX 1080 Ti GPU 2760 4MB, an Intel Core i7-6850K 12-core processor operating at 3.60 GHz, and additional parts. Google Collaboratory serves as the testing and training platform. The hyperparameters utilized for this study are tabulated in Table I.

TABLE I. Hyperparameters

Batch Size	32
Activation Function	ReLu
Optimizer	Adam
Loss Function	Binary Crossentropy

B. Experimrntal Results

In Table II, classification report for the modified VGG-16 model is tabulated. The model's precision, recall, and accuracy are 97.94%, 98.23%, and 95.58%, respectively.

Fig. 6 shows a plot of the modified VGG-16 model's accuracy. As the number of epochs increases, training and testing accuracy constantly improves.

Fig. 7 displays the modified VGG-16 model's loss. With more epochs, the training and testing loss always decreases.

TABLE II. Classification Report of Proposed Model

Parameters	Obtained Results
Accuracy	97.94 %
Precision	98.23 %
Recall	95.58 %

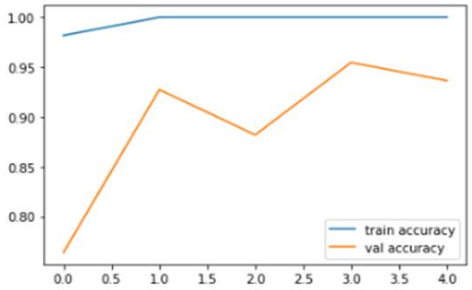


Fig. 6. Accuracy plot of proposed model.

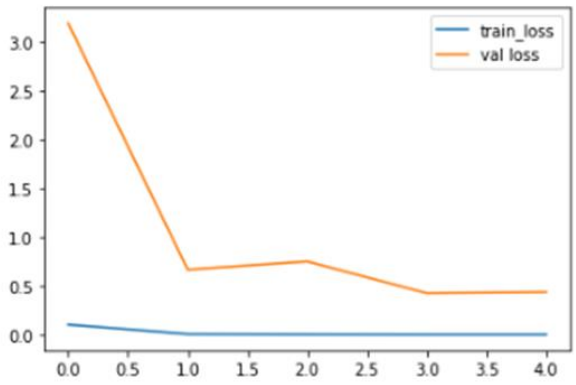


Fig. 7. Loss plot of proposed model.

A graph between numerous possible learning rates and the validation loss has been plotted in order to discover the learning rate for the model. The plot between learning rate and loss obtained for the model is shown below. The learning rate is a hyperparameter that determines how much to alter the model each time when model weights are updated in response to the predicted error. The plot between learning rate and loss of the model is illustrated in Fig. 8.

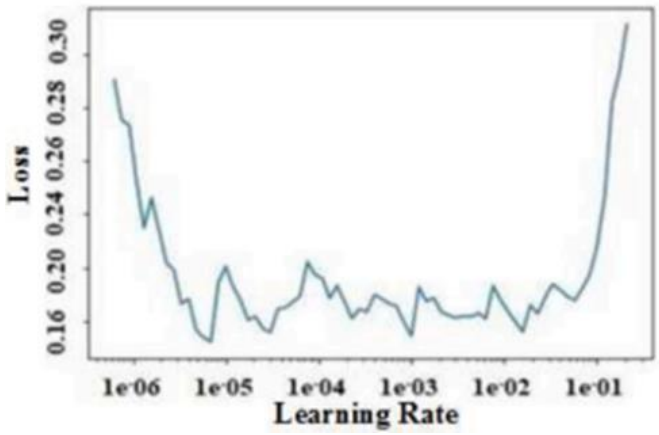


Fig. 8. Plot between learning rate and loss of the model.

It can be difficult to choose the learning rate since a number that is too little could lead to a lengthy training process that could become stuck, but a value that is too large could lead to learning a suboptimal set of weights too quickly or to an unstable training process. How quickly the model adapts to the situation is determined by the learning rate. Given the smaller changes to the weights made with each update, smaller learning rates necessitate more training epochs, whereas bigger learning rates produce quick changes and necessitate fewer training epochs.

The chest X-ray images were classified by the model as COVID-19 or NORMAL. Fig. 9 is an example of the image classification. The models can detect what category an image belongs to when given a randomly selected image as input.

The imaging similarities between normal and COVID-19 infection cases may have led to the misclassification, which is shown in Fig. 10.

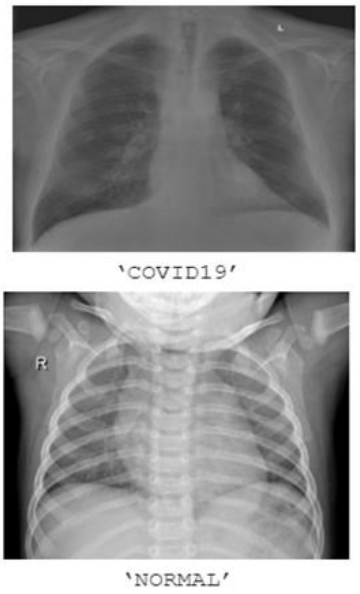


Fig. 9. Sample output of Covid-19 detection.

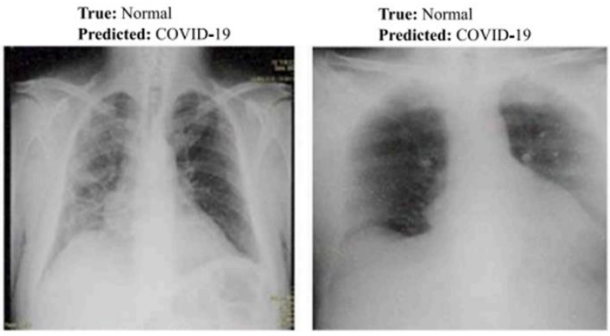


Fig. 10. Illustration of misclassification.

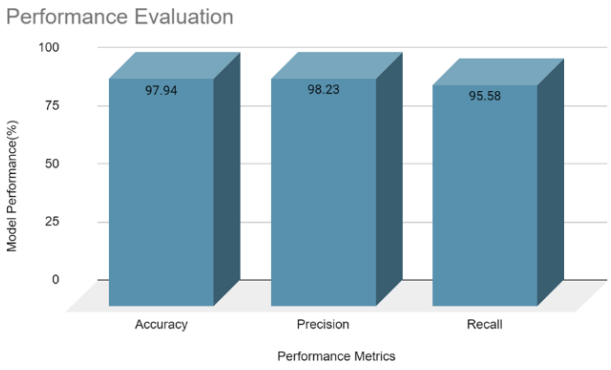


Fig. 11. Performance evaluation of proposed model.

The model's performance evaluation is graphically represented in Fig. 11. The x- axis and y- axis represent the performance metrics and model performance (%) respectively. The model obtained 97.94% accuracy, which outperforms other existing pre-trained models. The model also provides better precision and recall rate.

C. Discussions

The simulation results for the COVID-19 detection model using a modified VGG-16 architecture are promising, indicating a high overall accuracy of 97.94%. Precision, which measures the model's ability to correctly identify positive cases among the predicted positives, is notably high at 98.23%. The recall, representing the model's capability to identify all actual positive cases, is also strong at 95.58%. These metrics collectively suggest that the modified VGG-16 model exhibits a robust performance in distinguishing COVID-19 cases from normal cases.

In comparison with traditional detection methods, the high accuracy, precision, and recall values of the modified VGG-16 model underscore the potential of deep learning approaches in enhancing COVID-19 detection. Traditional methods often rely on manual interpretation or rule-based algorithms, which may lack the complexity and adaptability demonstrated by deep learning models. The precision and recall values above 95% indicate a low rate of false positives and false negatives, crucial for reliable COVID-19 diagnostics. The model's performance may be influenced by the diversity of the dataset, variations in image quality, and the representativeness of the COVID-19 cases. Moreover, the interpretability of deep learning models poses a challenge, making it crucial to ensure the clinical relevance and trustworthiness of the predictions.

In conclusion, the modified VGG-16 model exhibits a commendable performance in COVID-19 detection, outperforming traditional methods in terms of accuracy, precision, and recall. While these results are promising, further validation on diverse datasets and real-world clinical settings is necessary to establish the model's robustness and generalizability for practical implementation in the field of COVID-19 diagnostics. Additionally, efforts should be directed towards addressing interpretability concerns and ensuring seamless integration with existing healthcare practices.

V. Conclusion

According to the World Health Organization (WHO), more than five million people have been infected globally to date, and there have been about three lakh confirmed cases of death. Therefore, it is crucial to identify COVID-19 as soon as possible in order to stop its spread and lower its mortality. Currently, RT-PCR is the benchmark for diagnosing COVID-19. In this test, viral nucleic acid from sputum or a nasopharyngeal swab is found. This testing mechanism has a few drawbacks. The complicated and painful nature of this testing technique for the patients is another issue. As a result of these concerns, there is a huge need for alternative diagnostic techniques that deal with these issues. In this work a modified VGG-16 model was used for COVID-19 detection from chest X-ray images. This model classified the chest X-ray images into two categories: COVID-19 and NORMAL images. Finally, the performance of the proposed model is evaluated. The model has been validated on publicly available chest X-ray images. The best performance is obtained by the proposed model with an accuracy of 97.94%.

VI. Future Work

Therefore, in future work, the performance of the suggested methodology for multiclass classification issues will be established. In order to create a model that is more dependable, we also intend to investigate the application of optimization methods in conjunction with the DL models utilized in this study. In order to efficiently provide doctors with precise focal area detection during diagnosis and to facilitate early illness identification and prevention, the proposed model will be integrated with object detection.

References

[1] Wu YC, Chen CS, Chan YJ. The outbreak of COVID-19: An overview. Journal of the Chinese medical association. 2020 Mar;83(3):217.

[2] Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. International journal of antimicrobial agents. 2020 Mar 1;55(3):105924.

[3] Dewangan V, Sahu R, Satapathy T, Roy A. The Exploring of Current Development status and the unusual Symptoms of coronavirus Pandemic (Covid-19). Research journal of Pharmacology and Pharmacodynamics. 2020;12(4):172-6.

[4] Giri B, Pandey S, Shrestha R, Pokharel K, Ligler FS, Neupane BB. Review of analytical performance of COVID-19 detection methods. Analytical and bioanalytical chemistry. 2021 Jan;413(1):35-48.

[5] Teymouri M, Mollazadeh S, Mortazavi H, Ghale-Noie ZN, Keyvani V, Aghababaei F, Hamblin MR, Abbaszadeh-Goudarzi G, Pourghadamyari H, Hashemian SM, Mirzaei H. Recent advances and challenges of RT-PCR tests for the diagnosis of COVID-19. Pathology-Research and Practice. 2021 May 1;221:153443.

[6] McAdams HP, Samei E, Dobbins III J, Tourassi GD, Ravin CE. Recent advances in chest radiography. Radiology. 2006 Dec;241(3):663-83.

[7] Russell SJ, Norvig P. Artificial intelligence a modern approach. London; 2010.

[8] Calderon-Ramirez S, Giri R, Yang S, Moemeni A, Umana M, Elizondo D, Torrents-Barrena J, Molina-Cabello MA. Dealing with scarce labelled data: Semi-supervised deep learning with mix match for covid-19 detection using chest x-ray images. In2020 25th International Conference on Pattern Recognition (ICPR) 2021 Jan 10 (pp. 5294-5301). IEEE.

[9] Akter S, Shamrat FJ, Chakraborty S, Karim A, Azam S. COVID-19 detection using deep learning algorithm on chest X-ray images. Biology. 2021 Nov 13;10(11):1174.

[10] Karhan Z, Fuat AK. Covid-19 classification using deep learning in chest X-ray images. In2020 Medical Technologies Congress (TIPTEKNO) 2020 Nov 19 (pp. 1-4). IEEE

[11] Bourouis, S., Alharbi, A., & Bouguila, N. (2021). Bayesian learning of shifted-scaled dirichlet mixture models and its application to early COVID-19 detection in chest X-ray images. Journal of Imaging, 7(1), 7.

[12] Karakanis S, Leontidis G. Lightweight deep learning models for detecting COVID-19 from chest X-ray images. Computers in biology and medicine. 2021 Mar 1;130:104181

[13] Chandra TB, Verma K, Singh BK, Jain D, Netam SS. Coronavirus disease (COVID-19) detection in chest X-ray images using majority voting-based classifier ensemble. Expert systems with applications. 2021 Mar 1; 165:113909.

[14] Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Acharya UR. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Computers in biology and medicine. 2020 Jun 1; 121:103792

[15] Heidari M, Mirniaharikandehei S, Khuzani AZ, Danala G, Qiu Y, Zheng B. Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms. International journal of medical informatics. 2020 Dec 1; 144:104284

[16] Ilyas M, Rehman H, Naït-Ali A. Detection of covid-19 from chest x-ray images using artificial intelligence: An early review. arXiv preprint arXiv:2004.05436. 2020 Apr 11.

[17] Arias-Londoño JD, Gomez-Garcia JA, Moro-Velazquez L, Godino-Llorente JI. Artificial intelligence applied to chest X-ray images for the automatic detection of COVID-19. A thoughtful evaluation approaches. Ieee Access. 2020 Dec 14;8:226811-27.

[18] Hassantabar S, Ahmadi M, Sharifi A. Diagnosis and detection of infected tissue of COVID-19 patients based on lung X-ray image using convolutional neural network approaches. Chaos, Solitons & Fractals. 2020 Nov 1;140:110170.

[19] Ohata EF, Bezerra GM, das Chagas JV, Neto AV, Albuquerque AB, De Albuquerque VH, Reboucas Filho PP. Automatic detection of COVID-19 infection using chest X-ray images through transfer learning. IEEE/CAA Journal of Automatica Sinica. 2020 Sep 24;8(1):239-48.

[20] Jain G, Mittal D, Thakur D, Mittal MK. A deep learning approach to detect Covid-19 coronavirus with X-Ray images. Biocybernetics and biomedical engineering. 2020 Oct 1;40(4):1391-405.

[21] Ismael AM, Şengür A. Deep learning approaches for COVID-19 detection based on chest X-ray images. Expert Systems with Applications. 2021 Feb 1;164:114054.

[22] Hussain E, Hasan M, Rahman MA, Lee I, Tamanna T, Parvez MZ. CoroDet: A deep learning based classification for COVID-19 detection using chest X-ray images. Chaos, Solitons & Fractals. 2021 Jan 1;142:110495.

[23] https://www.kaggle.com/datasets/khoongweihao/covid19-xray-dataset-train-test-sets.

[24] Sahinbas K, Catak FO. Transfer learning-based convolutional neural network for COVID-19 detection with X-ray images. InData science for COVID-19 2021 Jan 1 (pp. 451-466). Academic Press.

Optimizing Crack Detection: The Integration of Coarse and Fine Networks in Image Segmentation

Hoanh Nguyen*, Tuan Anh Nguyen
Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam

***Abstract*—In recent years, the automation of detecting structural deformities, particularly cracks, has become vital across a wide range of applications, spanning from infrastructure maintenance to quality assurance. While numerous methods, ranging from traditional image processing to advanced deep learning architectures, have been introduced for crack segmentation, reliable and precise segmentation remains challenging, especially when dealing with complex or low-resolution images. This paper introduces a novel method that adopts a dual-network model to optimize crack segmentation through a coarse-to-fine strategy. This model integrates both a coarse network, focusing on the global context of the entire image to identify probable crack areas, and a fine network that zooms in on these identified regions, processing them at higher resolutions to ensure detailed crack segmentation results. The foundation of this architecture lies in utilizing shared encoders throughout the networks, which highlights the extraction of uniform features, paired with the introduction of separate decoders for different segmentation levels. The efficiency of the proposed model is evaluated through experiments on two public datasets, highlighting its capability to deliver superior results in crack detection and segmentation.**

***Keywords*—Deep learning; crack segmentation; coarse-to-fine strategy; image segmentation**

I. Introduction

The structural integrity and safety of infrastructures, such as buildings, roads, dams, and bridges, are vital to the well-being of societies across the world. One of the earliest and most common indicators of deteriorating structural health is the appearance of cracks. Crack segmentation, an essential branch of computer vision and structural health monitoring, focuses on the accurate identification and tracking of cracks in various materials and surfaces. The primary objective is to detect cracks as early as possible, ensuring timely maintenance, prevention of potential catastrophic failures, and extension of the lifespan of structures. With the rapid development of deep learning [1], [2], [3], [4] and the emergence of segmentation models such as SegNet [5], UNet [6], FCNs [7], and the DeepLab series [8], [9], general semantic segmentation tasks and crack segmentation tasks, in particular, have achieved significant improvements [10], [11], [12]. However, crack segmentation still poses numerous challenges that need addressing [13], [14], [15]. First, cracks can appear in various shapes, ranging from fine lines to wide gaps, and may display in different depths, lengths, and orientations. This makes it challenging for models to generalize across all possible crack presentations. Second, the surface on which a crack appears often possesses its texture, which can resemble a crack, making it difficult to differentiate between actual defects and background patterns. Third, uneven and dynamic lighting conditions, as well as external factors such as dirt, moisture, or staining can either obscure cracks or create shadows that might be mistaken for cracks. In recent years, with the rapid advancement of convolutional neural networks (CNNs), many methods have been proposed to address these challenges. In study [16], the authors proposed a novel unsupervised multi-scale fusion crack detection algorithm for pavement images, which addresses challenges posed by intensity inhomogeneity, topology complexity, and other factors without the need for training data. This method integrates a windowed minimal intensity path-based technique for candidate crack extraction, cross-scale crack correspondence, and a multivariate statistical hypothesis test for crack evaluation. In study [17], the authors introduced a cutting-edge network architecture called feature pyramid and hierarchical boosting network tailored for pavement crack detection. This network integrates context information into low-level features through a feature pyramid approach, and introduces a unique nested sample reweighting process, along with a new measurement method, the average intersection over union for enhanced crack detection accuracy. Yue et al. [18] presented CrackNet-V, an enhanced deep network tailored for pixel-level crack detection in 3D asphalt pavement images. This advanced network, building upon the foundational principles of CrackNet, features a deeper structure with fewer parameters, thereby offering superior accuracy and computational efficiency, while also incorporating novel features like the leaky rectified tanh activation function for precise shallow crack detection. Zhengxin et al. [19] proposed a semantic segmentation neural network designed for road area extraction, seamlessly merging the capabilities of residual learning and the U-Net architecture. This model incorporates residual units for simplified deep network training while its rich skip connections streamline information propagation, resulting in a leaner yet more performative network. In study [20], the authors introduced the Crack Transformer network (CrackFormer), a specialized solution tailored for fine-grained crack detection, integrating innovative attention mechanisms within a SegNet-inspired encoder-decoder framework. CrackFormer features unique self-attention modules and efficient positional embedding, while also incorporating new scaling-attention modules, emphasizing semantic crack features and reducing non-semantic interferences.

Although the above methods have addressed many challenges of crack segmentation, some difficulties remain, especially with thin cracks. As illustrated in Fig. 1, thin cracks are often more difficult to detect, particularly in low-resolution images. Furthermore, thin cracks can easily be mistaken for the

natural texture of asphalt, especially when the asphalt surface is rough or granular. Thin cracks can also appear darker or fainter depending on the lighting conditions and the angle of image capture, posing challenges for consistent detection. To address these issues, this paper introduces a dual-network model that employs a coarse-to-fine strategy for enhanced crack segmentation. The model consists of two networks: the coarse network captures global image context to identify potential crack areas, followed by the fine network, which focuses on these identified regions at a high resolution to achieve precise segmentation. Both networks share an encoder, but use separate decoders to process images, ensuring high-quality crack detection results. The proposed model is evaluated on two public datasets, including CrackTree260 and DeepCrack537 datasets. Experimental results show that the proposed model delivers superior results in crack detection and segmentation.

Fig. 1. Some images illustrate the challenges faced when performing crack segmentation.

II. Proposed Model

Fig. 2 provides a detailed visualization of the multi-stage segmentation process implemented in the proposed method. The proposed pipeline consists of both a coarse and a fine network. While both networks use a shared encoder, denoted as E, they each have their own decoders: D_c for the coarse network and D_f for the fine network. The main objective of the coarse network is to capture global contextual information from the entire image and subsequently highlight regions potentially containing cracks. Based on the predictions from the coarse network, the fine network then zooms into these identified regions, focusing specifically on local patches considered to have cracks, to achieve high-resolution crack segmentation. An input image represented mathematically as $I \in R^{W \times H \times 3}$ undergoes a downsampling step first. This optimizes its size for an initial analysis without sacrificing key details. Following this, the coarse network processes the downsized image to identify regions that may have cracks, resulting in a coarse crack map. This map essentially serves as a probability output, given by $P = Sigmoid(D_c(E(I)))$, highlighting pixels with a higher probability of being part of a crack. To refine this output, a thresholding technique is applied, producing a more defined coarse crack mask. This mask is then employed as a guide for the subsequent fine network. The fine network delves deeper, cropping and zooming into the previously identified regions. Its primary task is to work on these local patches, operating at a higher resolution to accurately segment the cracks, leading to a precise segmentation result.

A. Shared Encoder

We use ResNet-101 architecture in [21] as the shared encoder of our model. ResNet-101, a variant of the Residual Network (ResNet) family, is a deep convolutional neural network architecture known for its excellent performance on a variety of computer vision tasks. At its core, the design philosophy behind ResNet-101 is the introduction of "residual blocks" which address the vanishing gradient problem encountered in very deep neural networks. The network begins with a single convolutional layer with a 7×7 kernel, stride of 2, followed by a max pooling layer. The majority of the network consists of sequences of residual blocks. These blocks allow the model to learn identity functions that ease the training of deeper networks by providing shortcut connections across layers. Specifically, a residual block contains a skip connection that bypasses one or more layers. ResNet-101 contains four main groups of residual blocks. The first group has 3 blocks, the second group has 4 blocks, the third group comprises a significant 23 blocks, and the fourth group has 3 blocks. Each block within these groups consists of 3 layers (i.e., a 1×1 convolution, a 3×3 convolution, and another 1×1 convolution), with the exception of the first block of each group, which adjusts the number of channels and downsamples using a stride of 2. For crack segmentation task, we remove fully connected layers at the end of the architecture to maintain spatial information throughout the network. In addition, to stabilize the activations and speed up training, batch normalization and ReLU (Rectified Linear Unit) activation functions are applied after each convolution within the blocks. The details of ResNet-101 structure used in this paper are shown in Table I. The deep nature of ResNet-101 enables it to capture a wide range of features at different scales, while its residual connections ensure that training remains stable and efficient even with its impressive depth.

B. Segmentation Decoders

For both coarse and fine segmentation decoders, this paper employs the Atrous Spatial Pyramid Pooling (ASPP) decoder in [22]. ASPP is a prominent module designed to capture multi-scale contextual information without the need for multiple input scales or exhaustive downsampling. Originally proposed for semantic image segmentation in the context of the DeepLab series of architectures, ASPP is specifically designed to handle the challenges posed by objects of varying scales in images. The core concept behind ASPP is to apply parallel dilated (or atrous) convolutions on the feature map, each having a different dilation rate. This results in capturing spatial information from different field-of-views without substantially increasing the number of parameters or the computational burden. The multi-scale feature maps resulting from these parallel operations are then concatenated. Specifically, the ASPP module comprises: A 1×1 convolution which captures the image's immediate context, three 3×3 convolutions but with varying dilation rates (e.g., 6, 12, and 18), which allow the network to capture spatial information from different ranges without downsampling, and a global average pooling layer to

process the feature map, capturing the holistic context of the image. The resulting features are then upsampled and concatenated with the other components. After concatenating the outputs from these parallel operations, the combined feature map passes through another convolution to produce the final enhanced feature map that fuses multi-scale contextual information. This feature map is then fed into the ASPP decoder for semantic segmentation, as shown in Fig. 3. In the ASPP decoder, the enhanced feature maps are first bilinearly upsampled by a factor of 4 and then concatenated with the corresponding low-level feature map from the backbone (i.e., Conv2 layer of ResNet-101). To implement the concatenation operation, a 1×1 convolution layer is applied to the low-level feature map to reduce the number of channels. After concatenation, two 3×3 convolution layers are used to refine the features, followed by another simple bilinear upsampling by a factor of 4. The refined feature map is finally fed into the segmentation head. The ASPP decoder enables the network to effectively segment objects of various sizes and scales in images, making it a powerful choice for many semantic segmentation tasks.

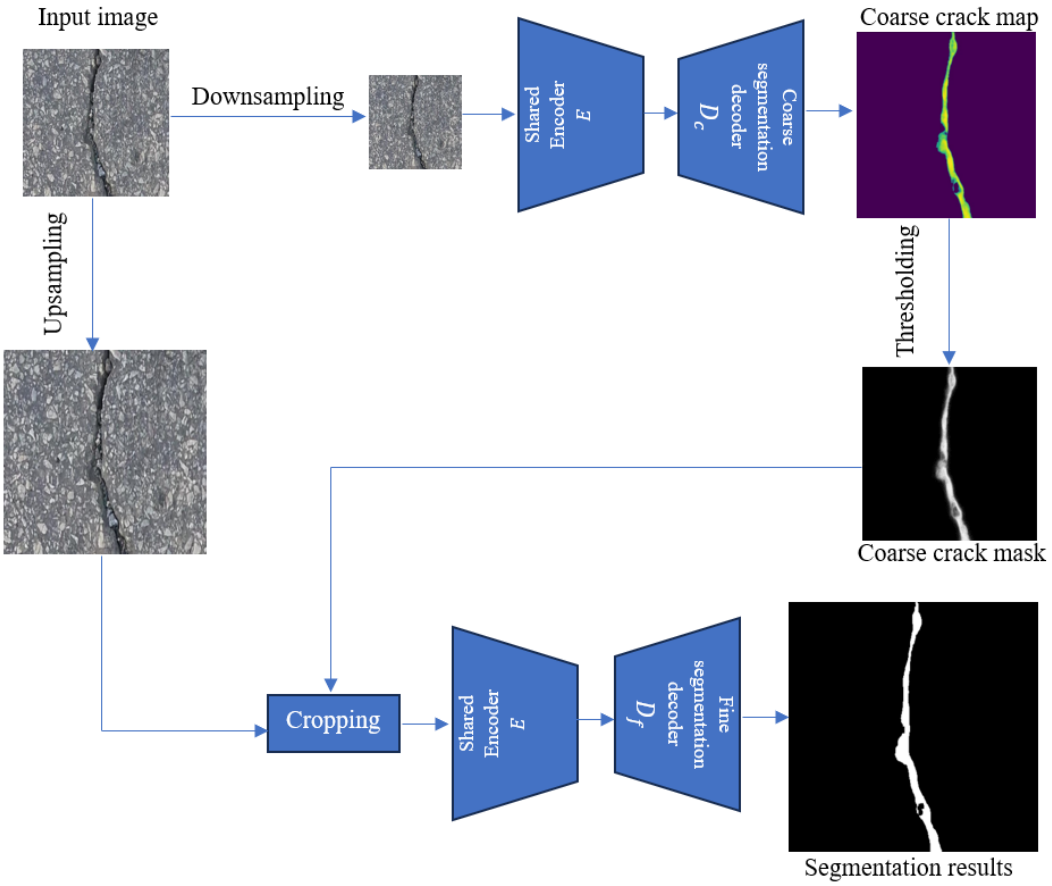


Fig. 2. The structure of the proposed model for crack segmentation.

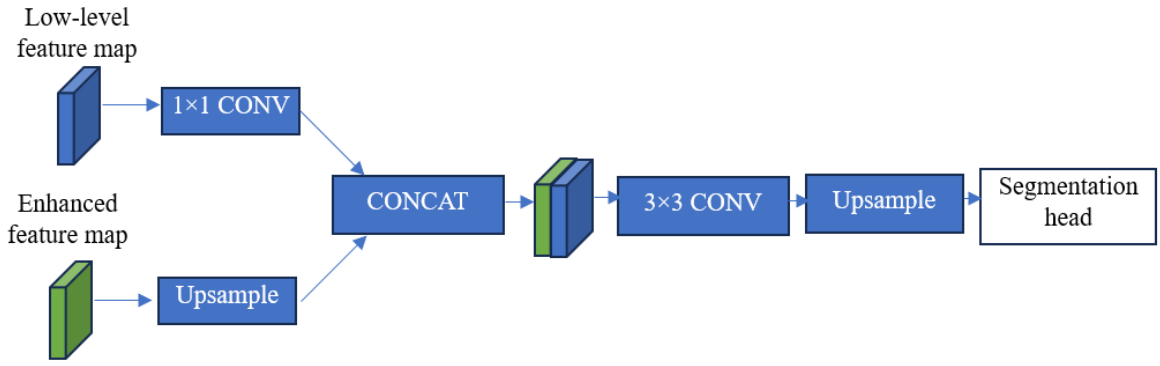


Fig. 3. The pipeline of the ASPP decoder.

C. Cropping

To extract high-resolution crack patches from the original image based on the coarse crack map, we first convert the coarse crack probabilities generated by the coarse network into a binary mask using a hard-thresholding method. This will produce a binary mask indicating the presence of cracks. To ensure that all potential details of the crack, even those that might have been slightly missed by the coarse network, are captured, we apply a dilation operation on the binary mask. This slightly enlarges the mask region. Next, we compute the bounding box for each contiguous region in the dilated binary mask. To generate these bounding boxes for crack regions, we select all the pixels with a corresponding density mask value of "1". We then merge the eight-neighbor connected pixels into a large candidate region. Finally, we use the circumscribed rectangle of the candidate region to crop the original image. This box encapsulates the region potentially containing the crack. Furthermore, we filter out crops with resolutions below the density threshold to eliminate noise and reject low-resolution patches. This step is crucial because crack segmentation doesn't perform well on low-resolution patches. After extracting the corresponding high-resolution patches from the original image, we feed each one into the fine network for detailed segmentation. As this network focuses only on smaller regions containing potential cracks, it can pay more attention to details, yielding better segmentation quality. Once the fine module processes the patches, it generates high-resolution segmentation for each patch. These segmented patches are then projected back to their original positions in the full-resolution image to obtain the final segmented result. By using this approach, the model benefit from both the global context provided by the coarse module and the local detail-centric approach of the fine module, ensuring accurate segmentation of cracks even in high-resolution images.

TABLE I. THE DETAILS OF RESNET-101 STRUCTURE USED AS THE SHARED ENCODER IN THIS PAPER

Layer type	Output size	Details
Input	H × W × 3	
Conv1	H/2 × W/2 × 64	7×7 convolution, stride 2
Max Pooling	H/4 × W/4 × 64	3×3 max pool, stride 2
Conv2_x (3 blocks)	H/4 × W/4 × 256	[1×1, 64], [3×3, 64], [1×1, 256] for each block
Conv3_x (4 blocks)	H/8 × W/8 × 512	[1×1, 128], [3×3, 128], [1×1, 512] for each block
Conv4_x (23 blocks)	H/16 × W/16 × 1024	[1×1, 256], [3×3, 256], [1×1, 1024] for each block
Conv5_x (3 blocks)	H/32 × W/32 × 2048	[1×1, 512], [3×3, 512], [1×1, 2048] for each block

D. Training Loss

In the domain of crack segmentation, the objective is to classify each pixel as either being part of a crack or not. This pixel-wise classification task can be effectively formulated and optimized using the Binary Cross Entropy (BCE) loss as follow:

$$L = \frac{1}{N}\sum_{ij} BCE(y_{ij}, \hat{y}_{ij}) \quad (1)$$

$$BCE(y, \hat{y}) = -ylog(\hat{y}) - (1-y)\log(1-\hat{y}) \quad (2)$$

where, y is the true label of the pixel, $\hat{y}$ is the predicted probability of the pixel belonging to the class labeled as foreground class.

The BCE loss quantifies the divergence between the predicted probability distribution and the actual distribution of the pixel labels. Specifically, for every pixel in the segmented image, the model predicts a probability score representing its confidence that the pixel belongs to the crack class. The BCE loss then computes the logarithmic difference between these predicted probabilities and the ground truth labels. When the model's prediction aligns closely with the actual label, the BCE loss approaches zero, indicating a perfect prediction. Conversely, if the model's prediction deviates significantly from the ground truth, the loss value increases. This property makes BCE loss particularly suitable for the crack segmentation task, as it penalizes misclassifications heavily, thereby driving the model to improve its pixel-wise classification accuracy. By minimizing the BCE loss during training, the segmentation model is guided to produce predictions that closely match the true crack structures in the images, leading to precise and reliable segmentation results. We use the BCE loss to both coarse and fine networks. The final loss L is the sum of the two:

$$L = \lambda_c L_c + \lambda_f L_f \quad (3)$$

where, λ_c and λ_f are coefficients for each loss, L_c is the BCE loss for the coarse network, L_f is the BCE loss for the fine network.

III. EXPERIMENTS

A. Dataset and Metrics

The proposed model is trained and evaluated on two public crack datasets: the CrackTree260 [23] and DeepCrack537 [24].

CrackTree260 is a dataset comprising 260 road pavement images, which is an extended version of the dataset from [23]. Images are captured using an area-array camera under visible-light illumination. This study utilized 200 of these images for training, 20 for validation, and 40 for testing. To enhance the training set, data augmentation techniques were employed. Specifically, each image was rotated at nine distinct angles ranging from 0 to 90 degrees in 10-degree increments. Following rotation, each image was then flipped both vertically and horizontally. From each flipped variant, five sub-images of 512×512 pixels were extracted – four from the corners and one from the center. As a result of this augmentation process, the training set accumulates a total of 35,100 images.

DeepCrack537 comprises 537 images, each having resolution of 544×384 pixels. These images depict a variety of cracks. Unlike other crack datasets, the diversity of cracks in DeepCrack537 is notable. Examples include top-down views, tilted views, cracks on both concrete and asphalt surfaces, variations in crack width from wide to thin, and instances where the cracks are partially occluded. For the purposes of this study, 300 images were utilized for training and 237 for testing.

For evaluation metrics, we use Precision (P), Recall (R), $F-measure$, mean intersection over union ($mIoU$) to evaluate the proposed model. Precision evaluates how many of the detected/segmented cracks (positive predictions) are actually real cracks.

$$P = \frac{TP}{TP+FP} \quad (4)$$

where, TP (True Positives) are the correctly detected cracks, FP (False Positives) are the wrongly detected cracks (i.e., detected cracks that are not real).

Recall measures how many of the actual cracks have been detected by the segmentation model.

$$R = \frac{TP}{TP+FN} \quad (5)$$

where, FN (False Negatives) are the actual cracks that the model failed to detect.

$F-measure$ is the harmonic mean of Precision and Recall. It provides a balance between the two. If either Precision or Recall is low, the $F-measure$ will also be low. It's especially useful when the class distribution is uneven.

$$F-measure = \frac{2\times P\times R}{P+R} \quad (6)$$

Mean intersection over union ($mIoU$) is a popular metric for segmentation tasks. It evaluates the overlap between the ground truth segmentation and the predicted segmentation.

$$mIoU = \frac{1}{2}\left(\frac{TP}{TP+FP+FN} + \frac{TN}{TN+FP+FN}\right) \quad (7)$$

B. Experimental Settings

The shared encoder based on ResNet-101 architecture is initialized with weights pre-trained on the ImageNet dataset [25] to take advantage of the extensive pretrained insights and rich feature representations it offers. The entire network was trained using the Adam optimizer [26] with a learning rate of 0.0001, which was reduced by a factor of 10 whenever the validation loss stabilized. Data augmentation techniques, including random rotations, zooms, and horizontal flips, were applied to prevent overfitting and enhance the model's generalization capabilities. The model is trained for 100k iterations with a batch size of 4. All experiments were conducted on a machine equipped with an NVIDIA RTX 4080 GPU, 64 GB RAM, and ran on the PyTorch framework.

C. Performance Evaluation

To demonstrate the effectiveness and superiority of the proposed method, we adopted eight existing and popular methods to compare to the proposed model, including Unet [6], TransUNet [27], FCNs [7], SegNet [5], and DeepCrack [28]. Table II shows results on the CrackTree260 dataset. From Table II, when examining the performance metrics of various models on the CrackTree260 dataset, the proposed model demonstrates superior performance across all metrics. With P of 0.892, R of 0.886, F-measure of 0.897, and mIoU of 0.894, the proposed model surpasses the other models in the ability to detect and segment cracks accurately. Notably, while DeepCrack comes closest to the proposed model with an F-measure of 0.852 and mIoU of 0.865, the proposed model still offers improvements, particularly in capturing the true positive rate as indicated by the highest recall. UNet and SegNet also demonstrate competitive results, with F-measures of 0.847 and 0.844 respectively. However, FCNs, with an F-measure of 0.463 and mIoU of 0.612, is the least effective among the mentioned models. TransUNet, despite being a transformer-based model, shows a relatively moderate performance with an F-measure of 0.771. Overall, the results suggest that the two-step approach of the proposed model is highly effective in detecting and segmenting cracks on the CrackTree260 dataset. For the DeepCrack537 dataset, the results are shown in Table III. From Table III, it's evident that the proposed model delivers the most impressive results in terms of crack detection and segmentation. With P of 0.891, R of 0.846, F-measure of 0.875, and mIoU of 0.878, the proposed model surpasses the other evaluated models. The DeepCrack model is the closest competitor with an F-measure of 0.847 and mIoU of 0.861, indicating a relatively high accuracy. SegNet also exhibits strong results with an F-measure of 0.840. In comparison, UNet and FCNs, while displaying reasonable performances, fall slightly behind with F-measures of 0.815 and 0.812, respectively. It is apparent from these results that the bi-level approach of the proposed model, involving a global context capture followed by focused high-resolution segmentation, is effective in the context of the DeepCrack537 dataset.

The visualization results of the model across datasets are depicted in Fig. 4. In this figure, the rows labeled "Input image" display images of various surfaces, each with distinct crack morphologies. The "Ground-truth labels" rows accurately depict the actual cracks present in the images. In contrast, the "Segmentation results" rows present the model's predictions. For most of the images, the segmentation results align closely with the ground-truth labels, indicating a high degree of accuracy in crack detection. This demonstrates the strength of our method in capturing both the global context through the coarse network and then refining the details through the fine network. Particularly in areas with intricate crack patterns or smaller fissures, the fine network provides superior results in segmenting these challenging features. However, there are instances, especially in the second set of images, where the segmentation results show a slightly broader crack outline than the ground-truth, suggesting a potential overestimation by the model in those cases. Overall, the visualization validates the effectiveness of our approach. While there are minor deviations in a few cases, the proposed pipeline demonstrates robust performance in segmenting cracks across diverse scenarios.

TABLE II. RESULTS ON THE CRACKTREE260 DATASET

Model	Metrics			
	P	R	F-measure	mIoU
UNet	0.860	0.834	0.847	0.861
TransUNet	0.797	0.746	0.771	0.803
FCNs	0.519	0.418	0.463	0.612
SegNet	0.851	0.837	0.844	0.858
DeepCrack	0.871	0.834	0.852	0.865
Proposed model	0.892	0.886	0.897	0.894

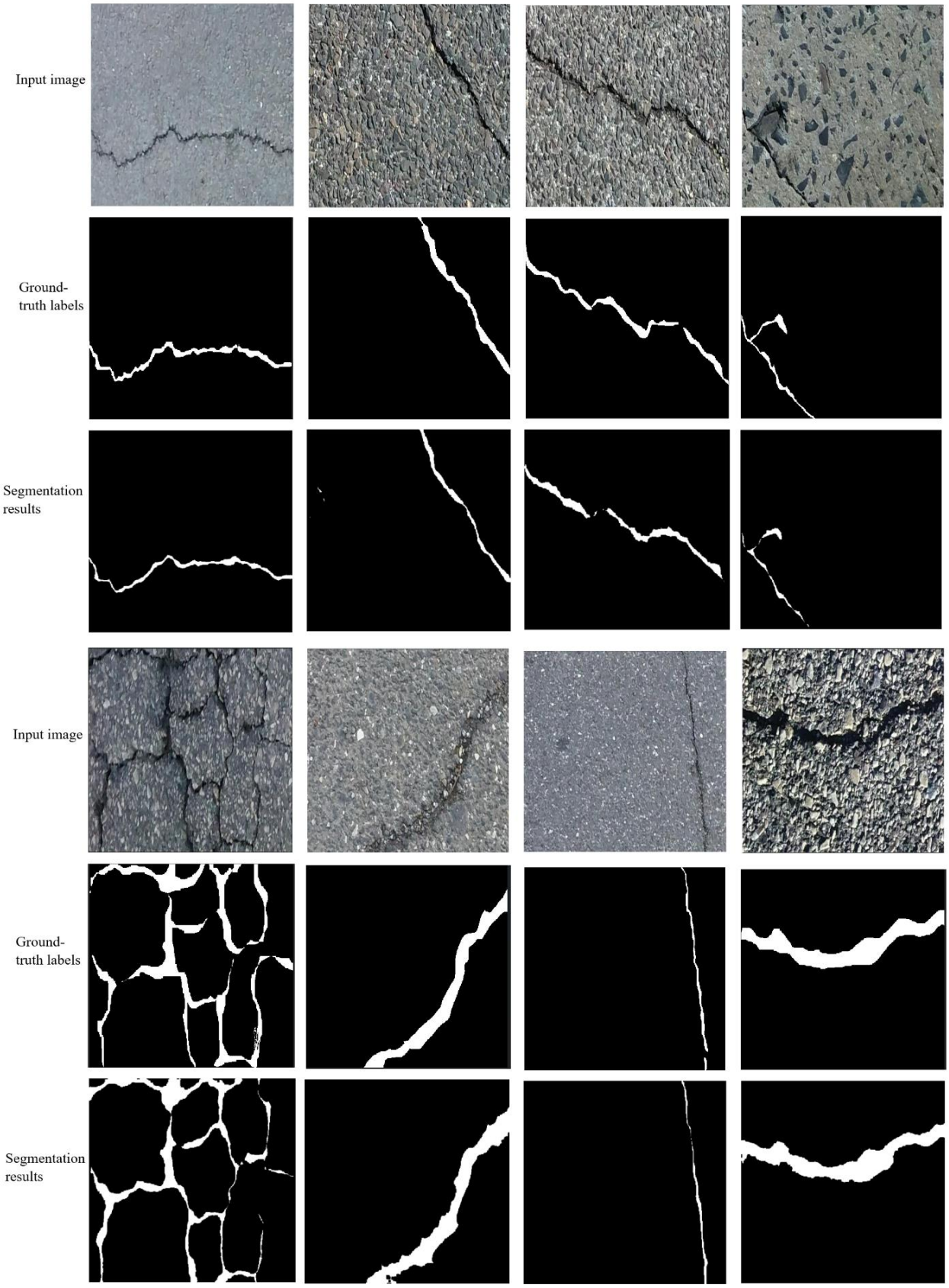


Fig. 4. Visualization results of the model on the datasets.

TABLE III. RESULTS ON THE DEEPCRACK537 DATASET

Model	Metrics			
	P	R	F-measure	mIoU
UNet	0.841	0.791	0.815	0.837
FCNs	0.829	0.796	0.812	0.833
SegNet	0.857	0.824	0.840	0.851
DeepCrack	0.876	0.819	0.847	0.861
Proposed model	0.891	0.846	0.875	0.878

IV. CONCLUSION

This study presents a dual-network model that optimally leverages the combined strengths of both a coarse and a fine network to enhance crack segmentation. Each network utilizes a shared encoder, with separate decoders tailored to their specific roles: the coarse network captures a holistic view of the image, emphasizing regions that potentially contain cracks, while the fine network focuses on these highlighted regions for precise high-resolution crack segmentation. The integration of the coarse network with the fine network was effective in addressing the challenges of crack detection. The experimental results on two public datasets underscore the robustness and effectiveness of the proposed approach in diverse scenarios. However, our model may struggle with extremely subtle cracks or those obscured by environmental factors, such as shadows or debris. Furthermore, the model's performance might vary depending on the quality and resolution of the input images. In future work, we plan to incorporate additional data augmentation techniques and explore the integration of advanced sensors for better crack detection in challenging conditions.

REFERENCES

[1] He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In *Proceedings of the IEEE international conference on computer vision*, pp. 2961-2969. 2017.

[2] Zhou, Bolei, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. "Semantic understanding of scenes through the ade20k dataset." *International Journal of Computer Vision* 127 (2019): 302-321.

[3] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems* 28 (2015).

[4] Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. "Ssd: Single shot multibox detector." In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 21-37. Springer International Publishing, 2016.

[5] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39, no. 12 (2017): 2481-2495.

[6] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234-241. Springer International Publishing, 2015.

[7] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431-3440. 2015.

[8] Chen, Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." *IEEE transactions on pattern analysis and machine intelligence* 40, no. 4 (2017): 834-848.

[9] Chen, Liang-Chieh, George Papandreou, Florian Schroff, and Hartwig Adam. "Rethinking atrous convolution for semantic image segmentation." *arXiv preprint arXiv:1706.05587* (2017).

[10] Ayomide, Kabirat Sulaiman, Teh Noranis Mohd Aris, and Maslina Zolkepli. "Improving Brain Tumor Segmentation in MRI Images through Enhanced Convolutional Neural Networks." *International Journal of Advanced Computer Science and Applications* 14, no. 4 (2023).

[11] Gunawan, Alexander Agung Santoso, Ilma Arifiany, and Edy Irwansyah. "Semantic segmentation of aerial imagery for road and building extraction with deep learning." *ICIC Express Letters* 14, no. 1 (2020): 43-51.

[12] Farooqui, Mehwash, Atta-ur Rahman, Roaa Alorefan, Mariam Alqusser, Lubna Alzaid, Sara Alnajim, Amal Althobaiti, and Mohammed Salid Ahmed. "Food Classification Using Deep Learning: Presenting a New Food Segmentation Dataset." *Mathematical Modelling of Engineering Problems* 10, no. 3 (2023).

[13] Mohan, Arun, and Sumathi Poobal. "Crack detection using image processing: A critical review and analysis." *alexandria engineering journal* 57, no. 2 (2018): 787-798.

[14] Adhikari, R. S., O. Moselhi, and A. Bagchi. "Image-based retrieval of concrete crack properties for bridge inspection." *Automation in construction* 39 (2014): 180-194.

[15] Prasanna, Prateek, Kristin J. Dana, Nenad Gucunski, Basily B. Basily, Hung M. La, Ronny Salim Lim, and Hooman Parvardeh. "Automated crack detection on concrete bridges." *IEEE Transactions on automation science and engineering* 13, no. 2 (2014): 591-599.

[16] Li, Haifeng, Dezhen Song, Yu Liu, and Binbin Li. "Automatic pavement crack detection by multi-scale image fusion." *IEEE Transactions on Intelligent Transportation Systems* 20, no. 6 (2018): 2025-2036.

[17] Yang, Fan, Lei Zhang, Sijia Yu, Danil Prokhorov, Xue Mei, and Haibin Ling. "Feature pyramid and hierarchical boosting network for pavement crack detection." *IEEE Transactions on Intelligent Transportation Systems* 21, no. 4 (2019): 1525-1535.

[18] Fei, Yue, Kelvin CP Wang, Allen Zhang, Cheng Chen, Joshua Q. Li, Yang Liu, Guangwei Yang, and Baoxian Li. "Pixel-level cracking detection on 3D asphalt pavement images through deep-learning-based CrackNet-V." *IEEE Transactions on Intelligent Transportation Systems* 21, no. 1 (2019): 273-284.

[19] Zhang, Zhengxin, Qingjie Liu, and Yunhong Wang. "Road extraction by deep residual u-net." *IEEE Geoscience and Remote Sensing Letters* 15, no. 5 (2018): 749-753.

[20] Liu, Huajun, Xiangyu Miao, Christoph Mertz, Chengzhong Xu, and Hui Kong. "Crackformer: Transformer network for fine-grained crack detection." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3783-3792. 2021.

[21] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[22] Chen, Liang-Chieh, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801-818. 2018.

[23] Zou, Qin, Yu Cao, Qingquan Li, Qingzhou Mao, and Song Wang. "CrackTree: Automatic crack detection from pavement images." *Pattern Recognition Letters* 33, no. 3 (2012): 227-238.

[24] Liu, Yahui, Jian Yao, Xiaohu Lu, Renping Xie, and Li Li. "DeepCrack: A deep hierarchical feature learning architecture for crack segmentation." *Neurocomputing* 338 (2019): 139-153.

[25] Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248-255. Ieee, 2009.

[26] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[27] Chen, Jieneng, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. "Transunet: Transformers make strong encoders for medical image segmentation." *arXiv preprint arXiv:2102.04306* (2021).

[28] Zou, Qin, Zheng Zhang, Qingquan Li, Xianbiao Qi, Qian Wang, and Song Wang. "Deepcrack: Learning hierarchical convolutional features for crack detection." *IEEE transactions on image processing* 28, no. 3 (2018): 1498-1512.

Enhanced Land Use and Land Cover Classification Through Human Group-based Particle Swarm Optimization-Ant Colony Optimization Integration with Convolutional Neural Network

Dr. Moresh Mukhedkar[1], Dr. Chamandeep Kaur[2], Divvela Srinivasa Rao[3], Shweta Bandhekar[4],
Dr. Mohammed Saleh Al Ansari[5], Maganti Syamala[6], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[7]
Assistant Professor, D Y PATIL UNIVERSITY, Pune, India[1]
Lecturer, Department of Computer Science, Jazan University, Jazan, Saudi Arabia[2]
Sr.Assistant Professor, Department of AI & DS, Lakireddy Bali Reddy College of Engineering, Mylavaram[3]
Assistant Professor, Department of Computer Science and Engineering,
Rungta College of Engineering and Technology, Bhilai, C.G, India[4]
Associate Professor, College of Engineering, Department of Chemical Engineering, University of Bahrain, Bahrain[5]
Assistant Professor, Dept.of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur Dist., Andhra Pradesh - 522302, India[6]
Faculty of Informatics and Computing, UniSZA University, Malaysia[7]

Abstract*—Reliable classification of Land Use and Land Cover (LULC) using satellite images is essential for disaster management, environmental monitoring, and urban planning. This paper introduces a unique method that combines a Convolutional Neural Network (CNN) with Human Group-based Particle Swarm Optimization (HPSO) and Ant Colony Optimization (ACO) algorithms to improve the accuracy of LULC classification. The suggested hybrid HPSO-ACO-CNN architecture effectively solves the issues with feature selection, parameter optimization, and model training that are present in conventional LULC classification techniques. During the initial phases, HPSO and ACO are crucial in identifying the best hyperparameters for the CNN model and fine-tuning the selection of critical spectral bands. ACO modifies the CNN's hyperparameters (learning rate, batch size, and convolutional layers), whereas HPSO finds the optimal selection of spectral bands. This optimization technique reduces the probability of overfitting while substantially enhancing the model's ability to generalize. Utilizing the selected spectral bands and optimum parameter configuration, the CNN algorithm is trained in the second phase. With Python implementation, this method uses both the spatial and spectral characteristics that the CNN detects to reach an outstanding 99.3% accuracy in LULC classification. The hybrid approach outperforms traditional methods like Deep Neural Network (DNN), Multiclass Support Vector Machine (MSVM), and Long Short-Term Memory (LSTM) in experiments using benchmark satellite image datasets, demonstrating a significant 10.5% increase in accuracy. This hybrid HPSO-ACO-CNN architecture transforms accurate and dependable LULC classification, offering an advantageous instrument for remote sensing applications. It enhances the area of satellite imagery evaluation by combining the advantages of deep learning techniques with optimization algorithms, enabling more accurate mapping of land use and cover for sustainable land management and environmental preservation.

Keywords*—Land use and land cover; human group-based particle swarm optimization; ant colony optimization; convolutional neural network; satellite image

I. Introduction

A crucial challenge that has significant implications across a range of regions is the accurate categorization of land cover and land use using satellite images. The primary focus of the position is the organized classification and labelling of the surface of the Earth, which serves as a fundamental perspective for understanding and managing the planet's changing landscapes. The designation of urban regions, the identification of infrastructural requirements, and the reinforcement of well-informed choices on land utilization allocation all have been made possible by the LULC categorization, which is crucial for urban planning [1]. This allows for the establishment of effective and environmentally responsible cities. It is a vital instrument in the field of management of the environment for determining how ecosystems are changing, detecting deforestation, and keeping track of the condition of ecosystems in their natural state. In addition, LULC categorization in agriculture provides farmers with knowledge about different crop categories, production, and farming methods, permitting targeted farming methods and boosting food security [2]. Accurate mapping of LULC can help with evaluating susceptibility, organizing for minimizing disaster risks, and adapting quickly to emergencies throughout disaster reconstruction and prevention operations. The capacity of satellite imaging to take wide-ranging images of the exterior of the planet from orbit is crucial for LULC categorization. These images provide us an unusual perspective from which can observe the intricate and constantly shifting topography of the earth [3]. The investigation has access to a variety of data on the Earth's surface, such as specifics about human behaviours, landscape

characteristics, and the surrounding environment, by using the camera's array of satellite sensors. A specific component of this categorization, known as land utilization, deals with the numerous ways individuals utilize and communicate with the land, including metropolitan regions, agricultural areas, transportation systems, manufacturing regions, and more. In contrast, land cover describes the physical properties of the Earth's surface independent of human activity, including forests, marshes, lakes and rivers, deserts, and arid areas. Together, the two distinct aspects related to land cover and land use provide an accurate representation of the Earth's surface and provide information regarding the complex interactions between the activities of humans and the surrounding ecosystem [4]. The level of accuracy of assessments made in a variety of disciplines is strongly impacted by how well LULC categorization is done. In urban planning, accurate regulatory control, infrastructure optimization, and support for ecologically friendly techniques are all aided by the definition of land utilization classifications. The capacity to distinguish between diverse kinds of land covers in management of the environment enables investigators to observe wildlife migratory patterns, follow habitat changes, and determine the effects of warming temperatures on ecosystems. In terms of agriculture, LULC categorization enables farmers to engage in decisions based on information, enabling them to select crops more effectively, manage irrigation more effectively, and lessen the impact of diseases and pests. Quick and precise LULC mapping is crucial for response to disasters in order to evaluate destruction, identifying impacted people, and efficiently arrange relief activities [5]. The key component for solving some of the most important issues confronting the global community, from development and deterioration of the environment to food availability and disaster resilience, is proper LULC categorization.

Satellite imagery is now more widely available and of higher quality than ever due to notable technological breakthroughs in the area of remote sensing in recent years. The latest phase of Earth observations has begun as a result of the growth in gathering information, providing an unusual viewpoint on the globe from orbit. Researchers have been able to collect data about the outermost layer of the planet and its changing operations at a degree of complexity never before possible due to the installation of innovative Earth-observing satellites with modern sensors. These satellites continually gather enormous volumes of information that cover a wide range of spectral data, temporal frequencies, and geographical resolutions [6]. Because of this, the field of remote sensing today is distinguished by an extensive collection of extensive and varied satellite imagery, which serves as a significant resource for a wide range of scientific, ecological, and social purposes. Even if the amount of available imagery from satellites is increasing exponentially, there are still many difficult problems it raises. For the information to be used effectively, it requires advanced approaches due to their enormous number and complexity. The fact that these images are multi-spectral and hyperspectral, indicating that they collect data from a broad variety of wavelengths, which include those outside the visible spectrum, presents one of the main obstacles [7]. This spectral variety adds a degree of complexity that necessitates sophisticated analytical methods capable of understanding the subtle differences in the information. Conventional LULC categorization methods suffer to handle this complexity because they are unable to capture the complicated patterns seen in multi-spectral and hyperspectral data. These methods are frequently founded upon manual characteristic engineering and rule-based systems. Traditional LULC categorization techniques frequently depend on hand-made characteristics and pre-established criteria, which may not be sufficient to capture the entire range of variability inherent in satellite images. These methods can be laborious and frequently need expertise in the area for extraction of features. Additionally, rule-based systems' low capacity for adapting to various and changing environments limits their usefulness. The immense prospective of deep learning methods, particularly CNNs, has, in comparison, emerged progressively more understood in the context of the analysis of satellite imagery [8]. CNNs are exceptionally effective at gathering pertinent characteristics from unprocessed information, which enables them to find complex spatial and spectral correlations that can resist manual characteristic engineering. They therefore provide a potential way to improve the accuracy and efficiency of LULC categorization using the vast amount of available satellite information.

The present article introduces a novel method that makes advantage of the interaction between algorithms for optimization and deep learning approaches to address the significant issues provided by the complexities of satellite images and the rising need for precise land use and land cover categorization. In particular, this innovative method combines Convolutional Neural Networks with two potent optimization algorithms—Human Group-based Particle Swarm Optimization and Ant Colony Optimization—to create a hybrid structure designed exclusively for the accurate and reliable categorization of LULC according to satellite imagery [9]. This integrative approach's primary driving force is to handle choosing characteristics and hyperparameter optimization, two crucial aspects of LULC categorization. The correct interpretation of satellite images depends heavily on identifying features, which involves choosing the most significant spectral bands or channels. Each of the categories are equally significant in the context of multi-spectral and hyperspectral imaging, and choosing a suitable combination of channels is essential for lowering distortion and redundancies while enhancing the approach capacity to discriminate between distinct land cover classifications. Human Group-based PSO intelligently selects the most relevant spectral bands to improve the standard of data given into the CNN using a collaborative procedure of optimization motivated by social group characteristics. The subsequent crucial issue the hybrid system addresses is hyperparameter optimization. A wide range of hyperparameters, including learning rates, batch sizes, and the number of convolutional layers, are included in CNNs as algorithms for deep learning. The effectiveness of the simulation is significantly impacted by these hyperparameters, therefore determining the optimum setup is extremely important [10]. ACO is used to adjust these hyperparameters, in order to ensure that the CNN performs at its highest level. It aims to minimize overfitting while

optimizing categorization accuracy by balancing model complexities and generalization. This combined strategy transcends the constraints of conventional approaches that depend on manual characteristic engineering and rule-based systems, signalling an important change in LULC categorization. This structure aims to enhance the accuracy and resilience of satellite based LULC categorization, permitting the efficient usage of the extensive and complicated information contained within satellite data. It does this by integrating the effectiveness of optimization methods into deep learning [11].

A crucial and fundamental stage in the field of satellite imagery evaluation, especially in the broader context of classifying land use and land cover, involves characteristic selection. The selection of the appropriate subset of spectral bands is crucial for a number of explanations, not the least of which is that not all spectral bands contributed similarly to the categorization process. Initially reducing data noise is accomplished by carefully choosing the spectral bands. Noise in imagery from satellites can come from a variety of places, such as air interference, sensor constraints, and changes in surface reflectance [12]. Feature selection eliminates or reduces the influence of noisy information by selecting the most pertinent bands, producing more accurate and precise categorization outcomes. This noise reduction improves the categorization model's general durability, making it less sensitive to incorrect classifications carried on by external influences. The process of categorization becomes quicker and more resource-efficient because to the reduction in redundancy, which also improves computing effectiveness. The present study uses Human Group-based Particle Swarm Optimization, a method informed by the combined intelligence of social networks, to carry out the process of characteristic selection effectively [13]. PSO replicates the cooperative behaviour of members of a group, where each member represents a possible mixture of spectral bands. These "particles" move around the spectral band subset search space, continuously modifying their placements in accordance with their individualized and shared understanding. Particles may successfully explore and utilize the search space thanks to PSO's cooperative characteristic by combining spectral bands in techniques that improve categorization accuracy while reducing noise and redundancies.

Convolutional Neural Network architecture tuning of hyperparameters is crucial for obtaining optimal results and strong adaptation as well as to characteristics selection. Hyperparameters include important factors that control whether deep-learning algorithms develop and are built, such as learning rates, size of batches, and the quantity of convolutional layers. These hyperparameters have a substantial impact on how well a CNN can recognize and understand complicated patterns in the information being processed. For example, during optimization, the learning rate determines the phase size and might affect the algorithm's convergence rate and quality. The batch size influences both computational effectiveness and generalization by affecting how the system procedures and modifies parameters during training [14]. Additionally, both the complexity and depth of the CNN is directly determined by the quantity of convolutional layers, with a greater number possibly permitting the collection of more complicated data. Therefore, it is essential to optimize these hyperparameters to ensure that the CNN performs at its optimal level while minimizing the danger of overfitting, which occurs when the algorithm develops excessively specific to the information used for training. This study presents a complete technique that integrates feature selection and hyperparameter optimization, two essential components of satellite image evaluation. The resultant hybrid method, which incorporates CNNs, ACO, and HPSO, has the possibility to transform satellite image processing. The combination of PSO for characteristic selection and ACO for hyperparameter optimization results in an integrated structure that makes use of both the representational strength of deep learning systems and the collective knowledge of optimization algorithms. This innovative method improves categorization accuracy while also strengthening resilience against complicated or noisy satellite imaging information. The hybrid PSO-ACO-CNN strategy that has been developed marks a substantial advancement in the effort to fully use satellite images for important applications in a variety of fields. Land cover and land use categorization skills, which are essential for environmental monitoring, urban planning, agriculture, and disaster management, are set to become more precise and dependable as a result of this technology. The study advances the latest developments in satellite image evaluation by demonstrating the efficacy and effectiveness of this framework via thorough investigations and findings. The potential significance of this study extends beyond the limits of research by providing real-world details that can enable experts to reach better decisions about how to manage the resources of the planet and deal with difficult problems. In short, the study represents a crucial step toward releasing satellite imagery's hidden potential for tackling pressing problems that the planet is currently and in future generations will be confronting.

The Key Contribution of the paper is given as follows:

- The study introduces the EuroSAT dataset, a sophisticated collection of satellite images created specifically for categorizing land cover and usage. This collection includes imagery from satellites with labeled data covering, different categories of land cover and usage across thirteen spectral bands. The massive dataset will be a valuable resource for the remote sensing and computer vision research as it allows for the investigation of deep learning and multimodal fusion techniques.
- The paper presents a unique hybrid optimization approach that combines Ant Colony Optimization for convolutional neural network hyperparameter optimization with Human Group-based Particle Swarm Optimization for feature selection. This technique addresses two important aspects of satellite image processing: selecting the appropriate spectral bands to employ and optimizing the CNN model for best results. The accuracy and robustness of the classification of land use and land cover are increased when these

optimization approaches are combined with deep learning.

- The research employs effective image pre-processing techniques, such normalization and histogram equalization, to enhance the quality of the acquired satellite images. These techniques reduce noise and improve system performance while ensuring that the input data to the classification model is of exceptionally high quality.
- By simultaneously optimizing numerous CNN model parameters, such as batch size and learning rate, the research expands on Ant Colony Optimization. The multi-parameter optimization technique ACO-DL enables the CNN to operate at peak efficiency, achieve ideal generalization, and avoid issues such as overfitting. It facilitates the training of the model more easily and leads to improved classification results.
- A comprehensive method that integrates data collection, picture pre-processing, feature selection, hyperparameter optimization, and CNN-based classification into a unified architecture is presented in the study. This comprehensive technique offers an effective way to accurately categorize land cover and use utilizing satellite data, and it has the possibility of revolutionizing the area of satellite image processing for a number of applications, such as urban planning and environmental monitoring.

The rest of the section is organised as shown below. Section II illustrates literature works on Land Use and Land Cover categorization. Section III gives the Problem Statement. Section IV covers the proposed technique for categorization of Land Use and Land Cover from satellite images. Section V illustrates the performance measures and summarises the findings and compares the method's performance to previous techniques. Section VI summarises the conclusion.

II. Related Works

The positive consequences of merging Sentinel-1 and Sentinel-2 imagery in the context of land use land cover categorization with U-Net and an evolving understanding of the combinatorial benefits of multi-sensor information fusion are highlighted. The benefits of using both Sentinel-1's radar data and Sentinel-2's optical information for improved LULC categorization have been studied in this field. Sentinel-1's radar information is useful for assessing land surfaces in a variety of environmental circumstances since it can operate in all weather conditions and can be observed through cloud cover. Contrarily, Sentinel-2's optical data offers high-resolution, multispectral data that specializes at catching specific spectral fingerprints, notably in differentiating between different plant varieties and urban characteristics. A potential method has evolved for combining these complementary information sources: U-Net, a deep learning architecture renowned for its capacity for semantic segmentation. In addition to increasing categorization accuracy, it also increases the resilience of LULC mapping by reducing the drawbacks of employing the various sensors separately, such as the sensitiveness of optical information to cloud cover and the sensitivity of radar information to specific varieties of land cover and roughness of the surfaces. Although this fusion strategy has a lot of potential, there are still difficulties in processing the volume of information, integrating multiple information modalities, and efficiently optimizing the deep learning algorithm's parameters [15].

A significant body of research highlighting the essential function of these technologies in evaluating environmental modifications in this important ecosystem has been revealed by the observation of land cover and land use modifications employing GIS and remote sensing methods in human-induced mangrove forest regions in Bangladesh. Investigations in previous years have demonstrated how well Geographic Information Systems technologies paired with remote sensing information, especially from satellites like Landsat and Sentinel, can capture and analyze alterations in mangrove forest cover, extent, and health. These methods have provided benefits including extensive coverage, recurrent gathering of information, and the capacity to distinguish between different land cover classes, that are crucial for tracking changes brought on by humans in mangrove ecosystems. Indicators like the Normalized Difference Vegetation Index and spectral characteristics have been used by researchers to recognize and categorize modifications, facilitating the discovery of elements like urbanization, aquaculture growth, and deforestation that have an impact on these ecosystems. However, issues with information quality, imagine interpretation, and the requirement for fine-scale observation to detect minor modifications still exist. Even yet, the combination of remote sensing and GIS offers a lot of potential for improving the comprehension of the dynamics and preservation of Bangladesh's human-induced mangrove forests [16].

Understanding the link between land cover and urban heat dynamics via remote sensing technologies is important, as demonstrated by the land-cover categorization and its effects on Peshawar's land surface temperature [17]. Previous studies have emphasized the benefits of using satellite imagery, especially Landsat and MODIS information, to map different types of land cover and measure how much that effect affects LST. Studies have shown how important land cover is in controlling urban microclimates, with permeable surfaces like buildings and roads causing higher LSTs that are frequently linked to the urban heat island effects. The influence of modifications to land cover on LST variations in Peshawar has been examined using a variety of categorization approaches, including supervised and unsupervised techniques, together with GIS tools. However, issues with information quality, geographical resolution, and the requirement for highly temporal-resolved statistics to record cyclical temperature fluctuations still exist. However, these studies contribute to the region's initiatives at development strategy and climate adaptation by offering significant understanding into the effects of urbanization-related modifications to land cover and their consequences for Peshawar's thermal environment.

A variety of research has been done on employing remote sensing technologies in order to track and understand the dynamic character of urban settings. This is evident in the study of urban land cover and land use changes employing

Random Forest categorization of Landsat time series information. With its constant and wide-ranging coverage, Landsat satellite information has proven to be a useful tool for tracking modifications to urban land cover over time. Random Forest, a machine learning algorithm, has been used in several research due to its efficacy in categorizing different types of land cover in metropolitan settings. The benefits of Random Forest, including its capacity to manage complicated spectral and temporal structures, account for noisy input, and produce reliable and accurate categorization outcomes, have been demonstrated by these researches. The investigations covered a variety of urban applications, such as detecting land use changes, assessing urban expansion, and characterizing urban heat islands, demonstrating the adaptability of this method. Due to their considerable effects on urban sustainability, management of resources, and quality of the environment, it also emphasizes the rising significance of monitoring modifications to urban land cover and land use. Urbanization, a global trend, has caused fast and occasionally uncontrolled expansion, changing the number of impermeable surfaces, deforestation, and urbanization, among other aspects of the land cover. Wide-ranging effects of these changes include higher energy use, changing microclimates, and ecological disturbances. In order to give statistical knowledge into these urban transitions, academics have increasingly resorted to remote sensing and machine learning approaches like Random Forest. Although the approach has many benefits, there are still some problems, such as the necessity for strong validation techniques, complicated information pre-processing, and modifying variables in the model. The substantial body of research in this area however emphasizes the crucial role that remote sensing and Random Forest categorization serve in dealing with the changing dynamics of urban land cover and land use transformations [8].

The evaluation of deep learning approaches to solve the challenges of satellite imagery evaluation highlights the rising interest in utilizing techniques based on deep learning for land use and land cover categorization in Southern New Caledonia [18]. Convolutional neural networks, in particular, have shown potential in automating LULC categorization activities. They have the capacity to gather features from unprocessed information, adjust to heterogeneous landscapes, and scale to multi-spectral and hyperspectral datasets, among other benefits. The complex and changing landscapes of Southern New Caledonia require effective methods for identifying spatial interdependence within images. The necessity for significant training data that is labeled, problems with the algorithm's interpretability, vulnerability to overfitting, computing resource requirements, and the requirement for balancing the collection of local and contextual data are still problematic. However, deep learning constitutes a substantial development in LULC categorization and has the possibility to enhance the knowledge of and ability to control the dynamics of land cover and land use in Southern New Caledonia.

Machine learning approaches were used to forecast land cover and land use from satellite photos, underscoring the increasing interest in utilizing cutting-edge technology for precise and effective land categorization. For tracking and comprehending modifications to land cover, imagery from satellites has evolved into a vital resource, and machine learning techniques have proven effective instruments in this field [19]. Several machine learning methods have been utilized in multiple studies to estimate the types of land cover and land use from satellite imagery. These methods have a number of benefits, including the capacity to handle big datasets, record complicated spatial patterns, and respond to various topographies. The breadth of research on machine learning-based land utilization forecasts has been demonstrated across a variety of applications, from urban planning and monitoring the environment to agricultural and disaster management. It also emphasizes the significance of precise land-use land-cover forecasts in tackling current issues like urbanization, deforestation, and environmental degradation. The capacity to observe and simulate modifications to land cover is essential for informed decision-making and effective utilization of resources as the global population keeps on growing urbanize and landscapes transform. By simplifying the categorization procedure and supplying accurate and fast data, machine learning approaches have been important in expanding our knowledge of these shifts. The necessity for high-quality information with labels, modelling generalization across diverse locations, and the understanding of complicated machine learning systems remained obstacles regardless their benefits. Nevertheless, the collection of research in this area highlights the possibilities of machine learning approaches in improving the ability to anticipate and efficiently react to modifications in land cover and land use.

An increasing number of researchers are interested in using advanced neural network topologies to improve the precision and effectiveness of land cover categorization operations, as shown by the examination of the deep learning framework for patch-based land cover categorization. Due to their ability to gather pertinent characteristics from image updates deep learning architectures, in particular Convolutional Neural Networks, have become increasingly popularity in recent years. It renders them ideal for classifying land cover from satellite or aerial images [20]. The benefits of CNNs have been demonstrated in research, particularly the capacity to deal with complicated land cover patterns, the capacity to concurrently record spatial and spectral data, and their adaptation to multi-spectral and high-resolution images. This research examined at a variety of deep learning architectures, including model topologies, hyperparameter tuning, and transfer learning, and have shown the way they may be used to achieve the highest possible accuracy in classifying land cover. The also highlights how important precise land cover categorization is for purposes in environmental evaluation, urban development, agriculture, and disaster prevention. For making decisions and policy creation, the capacity to autonomously and accurately classify different kinds of land cover at the patch levels is crucial. In this environment, deep learning architectures, which can handle massive datasets and provide real-time data, are emerging as an innovative technology. The necessity for large amounts of labeled information for training, the ability to interpret of models, and the computing resources necessary for deep network training are still issues. However, the large amount of research in this area shows the enormous potential of deep

learning architectures in enhancing the ability to categorizing land cover and addressing important problems with land cover and land use assessment.

The comprehensive extraction of multiscale timing dependency used in the land-cover categorization with time-series data remote sensing images emphasizes the growing significance of using temporal data in land cover assessment. Conventional land-cover categorization frequently employs images from a single date, which could not accurately represent how quickly land cover varies. On the other hand, time-series imagery from satellites, which are often collected over a long period of time, provide an extensive amount of information for comprehending land cover dynamics. Multiscale timing dependency, which takes into account not only the spectrum data but also time-dependent trends and relationships among observations, has been identified by investigators as having potential. Recurrent neural networks and machine learning methods have both been investigated to obtain thorough temporal information that will increase the reliability of land cover categorization. The results of these investigations show the benefits of using time-series information in land cover research, allowing for more accurate monitoring of modifications to land cover, urbanization, agricultural methods, and environmental alterations. The requirement for more study in this field is highlighted by the fact that there are still issues with data pre-processing, handling cloud cover, and dealing with the computing needs of analyzing large time-series datasets [21].

An integrated strategy that combines nature-inspired optimization approaches with modern deep learning methodologies to improve the accuracy of land cover categorization is reflected in the most effective orientation whale optimization algorithm and hybrid deep learning systems for land cover and use categorization. Due to the increasing accessibility of remote sensing information as well as computer resources, conventional land use land cover categorization methods have experienced substantial improvements. A potential optimization method for adjusting the hyperparameters of deep learning systems is the optimum guiding whale optimization algorithm, an extension of the whale optimization algorithm. This algorithm demonstrates increased convergence and optimization characteristics and is motivated by the social behaviour of humpback whales. This method uses spatial as well as temporal data from satellite images in combination with deep learning networks to accomplish accurate LULC categorization. The also highlights the significance of precise LULC categorization in several applications, such as urban planning, environmental surveillance, disaster preparation, and agriculture. Deep learning networks and the best guide whale optimization technique are used to solve the problem of improving complicated models while taking into account the special properties of remote sensing information. There continue to be issues with modelling interpretability, algorithm integration, and the requirement for a large amount of labeled training information. However, this method offers an innovative research area with the possibility of substantially enhance the accuracy and effectiveness of LULC categorization, which would be advantageous to many fields that depend on land cover data for making decisions and policy development [22].

The previously mentioned investigations pertaining to the classification of land use and cover highlight the growing need of using a wide range of remote sensing technologies and sophisticated methodologies to improve the ability to precisely and effectively evaluate the dynamics of land cover. These approaches have demonstrated an enormous amount of potential in terms of their ability to offer comprehensive data on alterations in land use and cover. The effectiveness of these methodologies in documenting changes in land cover and land use through time, for example, has been shown by studies undertaken in places like Peshawar, Southern New Caledonia, and human-induced mangrove forest areas in Bangladesh. Through the utilization of remote sensing technology, researchers have been able to get broad coverage, track alterations in land cover classes, and evaluate the effects of deforestation, urbanization, and aquaculture growth on diverse ecosystems. These methods have great potential to handle modern issues like catastrophe preparedness, urban planning, environmental protection, and agricultural management, where precise land cover data is essential. However, there are several difficulties and disadvantages with these intriguing approaches. The significant need for labeled training information which can be labour-intensive and time-consuming to obtain, particularly for extensive land cover mapping projects is one of the main obstacles. Furthermore, because they might affect the precision and dependability of classification results, the reliability and interpretability of information from remote sensing remain to be a cause for concern. Large time-series dataset management and analysis can provide logistical and technological difficulties. In order to fully realize the potential of these techniques and assure their successful implementation in real-world scenarios where accurate and timely land cover information is critical for well-informed decision-making and efficient resource management it must be essential that these obstacles be addressed.

III. PROBLEM STATEMENT

From the above literature review it is observed that the most important tasks in environmental monitoring, urban planning, and natural resource management is classifying land cover and land use utilizing satellite information. A critical component of decision-making processes is the correct categorization of land cover kinds, such as forests, urban areas, agricultural fields, and water bodies. The complexity and size of current satellite imaging information is frequently excessive for conventional strategies for land cover and land use categorization to manage [23]. This study suggests a unique method for addressing the issue by fusing the strength of hybrid HPSO and ACO with CNN for land use and land cover categorization. The goal is to create a categorization system that is accurate and effective, capable of autonomously analyzing satellite images and categorizing different types of land cover. By combining PSO and ACO with human input, the optimization procedure is regulated by human knowledge. By combining domain-specific expertise, this "human in the loop" method can produce superior outcomes.

IV. PROPOSED HPSO-ACO-CNN

Data gathering, pre-processing, feature selection utilizing a human group based PSO algorithm, and CNN hyper parameter optimization employing ACO constituted the approach used in this study. The process of gathering data includes building the EuroSAT dataset, which consists of 27,000 annotated Sentinel-2 satellite photos representing ten distinct land use and land cover classifications over 13 spectral bands. After data collection, normalization and histogram equalization were used in image pre-processing to improve the quality of the images. PSO was used in feature selection to intelligently identify pertinent spectral bands, while ACO was used to optimize CNN hyperparameters which includes batch size and learning rate. The CNN model, created for LULC categorization, completed training, and optimization was made possible by ACO-DL, a modification of ACO that allows for simultaneous optimization of several parameters. This hybrid approach provides a comprehensive approach integrating optimization approaches with deep learning for efficient satellite image processing. Its goal is to increase LULC categorization accuracy. Fig. 1 shows the overall structure of the proposed framework.

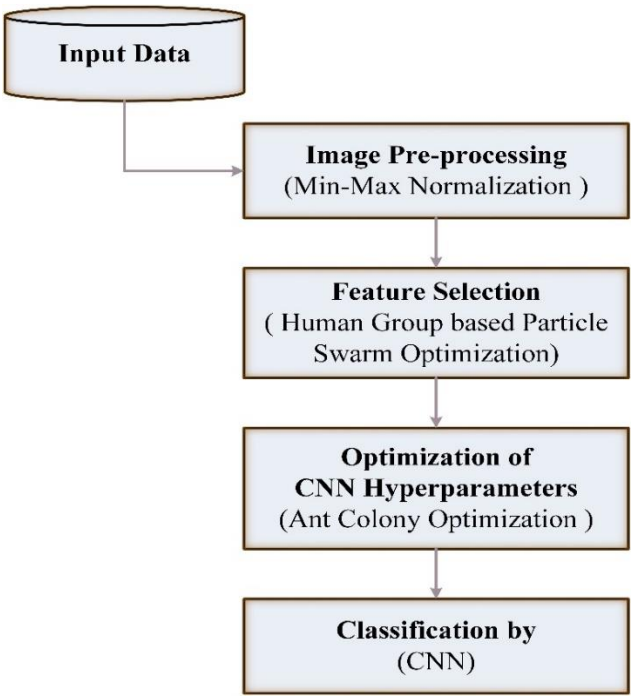


Fig. 1. Overall structure of the proposed framework.

A. Data Collection

The study presents an innovative set of satellite images for the categorization of land cover and land use. The Sentinel-2 satellite imagery that constitute up the 27,000 annotated images that constitute up the provided EuroSAT dataset depend on an overall of 10 distinct classifications. The patches are 64 by 64 pixels in size. The European Urban Atlas cities were chosen for the study's satellite images1. The given satellite image dataset, which includes thirteen spectral bands and has a significant amount of two thousand to three thousand image patches per class, differs significantly from earlier datasets in that it enables the investigation of multimodal fusion strategies in the overall setting of these bands. If deep neural networks need to be used for categorization, this is a particularly challenging problem. The offered dataset additionally depends on publicly available Earth observation information, opening up a variety of novel real-world applications. In accordance with the coverage in the European Urban Atlas, the areas included in the dataset were collected from cities distributed over thirty different European nations. Additionally, each individual picture patch's geoformation is made accessible to the public together with the labeled dataset EuroSAT. In order to obtain as much variation from the covered land cover and land use classifications as feasible, the study also extracted images taken throughout the year [24][1].

B. Image Pre-processing using Min-Max Normalization

To improve the quality of the satellite images, normalization and histogram equalization techniques are used after data collection. By altering the range of pixel values, a process known as image normalization, or contrast stretching, one may enhance the visually appealing qualities of satellite-image collection. (1) is a well-known simple formula that expresses the typical scenario of a min-max normalization to generate an additional image spanning from 0 to 1.

$$H_{out} = (H_{in} - Min)\frac{newMax - newMin}{Max - Min} + newMin \qquad (1)$$

Where the original satellite image is denoted as H_{in}, the minimum and maximum intensity values, which range from 0 to 255, are represented as Min and Max, respectively, the image after min-max normalization is denoted as H_{out}, and the new minimum and maximum values are denoted as $NewMin$ and $newMax$. The histogram equalization approach is then applied to enhance the image quality without eliminating any of the image's borders, patches, or points. The histogram equalization approach adjusts the normalized images' mean brightness to the allowable range's midpoint, while maintaining the original brightness prevents intrusive artifacts from appearing in the images.

C. Feature Selection using Human Group-based Particle Swarm Optimization

In this work, feature selection is done using Human Group-based Particle Swarm Optimization, which has the distinct benefit of simulating human cognitive capacities in optimization problems. By adding a human-guided component, HPSO improves upon the communal intelligence of particle swarm optimization, in which particles stand in for potential subsets of characteristics. The feature selection process is guided by heuristics and important domain experience provided by this human-in-the-loop technique, which increases its efficiency and context awareness. HPSO assures the selection of the most important characteristics while minimizing computational overhead by fusing human understanding with the computational power of PSO. This method is especially well-suited for difficult tasks like satellite image processing where domain knowledge is essential for precise feature selection. It improves the quality of selected

[1]https://ieeexplore.ieee.org/abstract/document/8736785/

characteristics, which in turn improves the efficiency of deep learning models like Convolutional Neural Networks. After generating the characteristic vectors, characteristics are chosen employing the human group-based PSO method. PSO is a population-based searching algorithm that usually simulates bird behaviour. In Eq. (2), is employed to modify the particle's position p_j and velocity v_j in order to produce new locations for each particle.

$$v_j(m+1) = w \times v_j(m) + h_1 \times d_1 \times \left(la_j(m) - p_j(m)\right) + h_2 \times d_2 \times \left(ga_j(m) - p_j(m)\right)$$

$$p_j(m+1) = p_j(m) + v_j(m+1) \quad (2)$$

where, m stands for the number of iterations, h_1 and h_2 are expressed as random real integers between [0, 1], w is a representation for the acceleration weight, a_j is a symbol for the best position, $la_j(m)$is a symbol for the local best position, and $ga_j(m)$ is a symbol for the global optimal position of the particle. In PSO, an adaptable uniform mutation is used to increase convergence and simplify implementation after the HGO method has been used to initially affect the particles.

A discrete multi-label is first converted into a continuous label using HGO. The employed approach locates the obtained feature vectors in accordance with decisionc_j, where the vectors of the particle's location are supplied as $p_j(m) = \left(p_j, 1, p_j, 2, p_j, C\right)$.

The feature selection algorithm's capacity for exploration is improved by the adaptive uniform mutation. The variety and choice of the mutation on each particle, p_j in this operator are controlled by a nonlinear functionp_n. Eq. (3) is used to update $p_n$at each cycle.

$$\mathrm{p_n} = 0.5 \times \mathrm{e}^{\left(-10 \times \frac{\mathrm{m}}{\mathrm{M}}\right)} + 0 \quad (3)$$

Where, m represents the number of iterations, M is designated as the maximum iteration, and the $p_n$value tends to fall as the number of iterations rises. If the $p_n$value is greater than the random number between [0, 1], the mutation selects the s elements at random from the particle. The mutation value of the items contained in the search space is then reset, with s serving as an integer value that limits the mutation range. Eq. (4) mathematically denotes the value of s as:

$$s = max\{1, |C \times p_n|\} \quad (4)$$

The following describes the human group-based PSO algorithm's step-by-step procedure.

Step 1: Establish the particle swarm's initial parameters, including (a) the number of iterations M, the swarm size T_k , and the archive size T_b. A non-dominated solution is saved into the archive after steps (b) initialize the particle locations, (c) estimate the aim of each particle, and do so.

Step 2: The particular best position of the particles is updated using the Pareto dominance relationship. The particular best position of the particles continues to remain unaltered if the new position $p_j(m+1)$ is superior to the previous personal best position$la_j(m)$, set$la_j(m+1) = p_j(m+1)$, where a_j is shown as the best position and $la_j(m)$ is shown as the local best position.

Step 3: Choose the global finest position from the archives according to the variety of solutions. To choose the particle's global optimal position$ga_j(m)$, a binary tournament is employed after initially calculating the crowding distance value.

Step 4: The decision value c_j is then initialized depending on$ga_j(m)$. The feature vector c's decision c_j is each a binary value$c_j = \pm 1, j = 1,2, \dots, T$. Each characteristic vector c is associated to the fitness value V(c), which is thought of as the weighted average of T stochastic contributions$W_i\left(c_i, c_1^i, \dots, c_k^i\right)$. However, the significance of decisions $c_j^i, j = 1,2, \dots, K$ and other K selections affects their contributions.

Eq. (5) mathematically illustrates the fitness function.

$$V(c) = \frac{1}{T}\sum_{i=1}^{T} W_i\left(c_i, c_1^i, c_2^i, \dots, c_k^i\right) \quad (5)$$

The total quantity of variables that interact decision values is denoted by the integer index$K = 0,1,2, \dots T-1$. The parameter$P \in [0,1]$, which represents the probability that each member has been informed of their contribution to the decision, determines the knowledge level of the $n^{th} n^{th}$ member. Each member n determines their individual estimated fitness utilizing (6) depending on their degree of knowledge.

$$V_n(c) = \frac{\sum_{i=1}^{T} \breve{c_{ni}} W_i\left(c_i, c_1^i, c_2^i, \dots, c_k^i\right)}{\sum_{i=1}^{T} \breve{c_{ni}}} \quad (6)$$

where, $\breve{c}$ is referred to as the matrix, whose generic member c_{ni} examines the numerical value one with probabilities (P and 0) with probabilities (1-P).

Step 5: Eq. (7) is employed to modify the particle's location $p_j$and velocity c_j in accordance with the decision valuec_j.

$$v_j(m+1) = w \times v_j(m) + h_1 \times d_1 \times \left(la_j(m) - p_j(m)\right) + h_2 \times d_2 \times \left(ga_j(m) - p_j(m)\right) \quad (7)$$

$$p_j(m+1) = p_j(m) + v_j(m+1) \quad (8)$$

Step 6: Apply Eq. (7) and Eq. (8) to uniform mutation.

Step 7: Utilizing the crowding distance approach, upgrade the external archives.

Step 8: Examine the termination circumstance: if the proposed algorithm completes the maximum number of iterations, the process should be terminated; otherwise, move back to phase 2. The HGO algorithm's fitness function $V_n(c)$ is used to remove the most deficient particles.

D. Optimizing CNN Hyperparameters using Ant Colony Optimization Strategy

The study utilizes Ant Colony Optimization to optimize numerous parameters simultaneously, which makes it a suitable method for fine-tuning Convolutional Neural Network hyperparameters. ACO is an optimization method that draws

inspiration from nature and is particularly efficient at navigating intricate search spaces. As such, it may be used to determine the best possible combination of hyperparameters for deep learning models. Its benefit is that it can investigate a large variety of hyperparameter values and adaptively modify them while optimizing. When working with hyperparameters, ACO's probabilistic method is extremely beneficial since it resembles how actual ants choose the shortest path in their natural environment. This method facilitates effective parameter space exploration and exploitation. This work uses the effectiveness of ACO to achieve optimal generalization, decreased overfitting risk, and enhanced CNN performance. As a result, it provides a reliable method for improving the model's accuracy in applications like satellite image categorization. ACO, developed by Marco Dorigo in 1992, is a standard heuristic swarm intelligence program that uses probabilistic calculations to identify the best planning pathway. ACO is a system based on positive feedback in which the ant finally chooses the path with the highest pheromone concentrations in order to obtain the best outcome possible in terms of the regulatory mechanisms.

The Convolutional Neural Network utilized in this study's Deep Learning framework was optimized using ACO. The study also altered the conventional ACO by using multitype ants to simultaneously improve different variables. The total quantity of ant varieties in ACO-DL is equal to the number of characteristics that need to be optimized. As a result, ACO-DL was able to optimize simultaneously a number of parameters related to the model in order to produce the best possible answer to the function of objective. The number of batches (A) in the network and the starting learning rate (L) in Adam were optimized for the CNN framework using ACO. The objective function (F (A, L)) selected was an accurate rate of predictions. Additionally, a given interval's values for A and L were determined in Eq. (9) and Eq. (10).

$$F_{max}(A, L) \tag{9}$$

$$k.t.\begin{cases} A \in [A_{min}, A_{max}] \\ L \in [L_{min}, L_{max}] \end{cases} \tag{10}$$

The fundamental concept is to iteratively discover the shortest path to the best solution of the goal function. In the meantime, the study established the subsequent two termination standards in order to ensure the efficiency of the optimization algorithm: 1) No apparent increase in accuracy; 2) the maximum number of repeats.

E. Classification Using Convolutional Neural Network

The CNN is the most efficient and productive approach network among deep learning techniques. Because CNNs can categorize intricate contextual images, they are widely used to categorize remote sensing data. Usually, these methods are not required for completing an output image prediction. CNNs are feed-forward neural structures that employ substantially local correlations to produce judgements by imposing an immediate interaction arrangement between neurons in neighbouring segments of the system. A maximal layer of pooling, the network layer, numerous convolutional layers, and fully linked layers constitute their architecture. Every stage of convolution calculates the weighted average of the prior characteristic using a channel before sending its findings via a stimulation functions to obtain the outcome. Using this method, the kernel measurement is computed to find neighbourhood correlations while preserving consistency for each region throughout the data clusters. The final characteristic pattern is created using constants at the lowest attainable unit level. The many levels of convolutional or layers of pooling are finally interfaced into a coherent unit using a fully coupled network of neurons. Eq. (11) and Eq. (12) gives the convolution operation.

$$fm_n^q(u, v) = UU_{\substack{id(x,y)e_n^q(w,t) \\ d \quad x,y}} \tag{11}$$

$$F_n^q = \left[fm_n^q(1,1), \ldots, fm_n^q(u, v), \ldots, fm_n^q(U, V)\right] \tag{12}$$

Following the extraction of the characteristics, a down sampling or pooling process is employed to gather the intersection of characteristics that are resistant to moderate transverse changes and deformation. It is given in Eq. (13).

$$R_n^q = s_r\left(F_n^q\right) \tag{13}$$

Similar to this, R_n^q stands for the Qth input feature-map's pooling characteristics-map of the mth layer, and s_r stands for the pooling operation. The maximum, average, L2, overlapping, and spatial pyramid pooling formulae are used in CNN. An activation function is used to speed up learning and offer a method of decision-making for a complicated feature-map. Both the non-linearity of the characteristics and the accelerated learning rate are provided by these activation functions. ReLu, sigmoid, tanh, maxout, and SWISH activation functions all have the same capability for supplying nonlinearity and resolving the vanishing gradient issue.

$$t_n^q = S_x\left(F_n^q\right) \tag{14}$$

S_x stands for the activation function in Eq. (14), F_n^q for the convolution output, and t_n^q for the converted output.

The two main decisions regarding design for CNN that offer superior efficiency and eliminate the overfitting issue are training and optimization. The number of extra problems for training the information generally grows along with the volume of information. The framework has difficulties when a novel or unfamiliar dataset is presented. Overfitting is a result of this issue, which dropout and batch normalization can solve. The dropout mechanism is employed to disable a large number of nodes at the conclusion of each cycle of the training phase. To enhance entire accuracy, strengthen the system's resistance to overfitting, and quicken the gradient descent process' convergence, batch normalization aims to impose a zero mean and a one standard deviation across every activation function in the established layer and for every single inadequate batch. The fully connected layer, the last component of the CNN framework as depicted in Fig. 2, combines every component with an additional layer to categorize. It gathers data from the characteristic extraction phase and analyses the output from every step before it. As a consequence, data categorization is accomplished by nonlinearly linking a set of chosen characteristics.

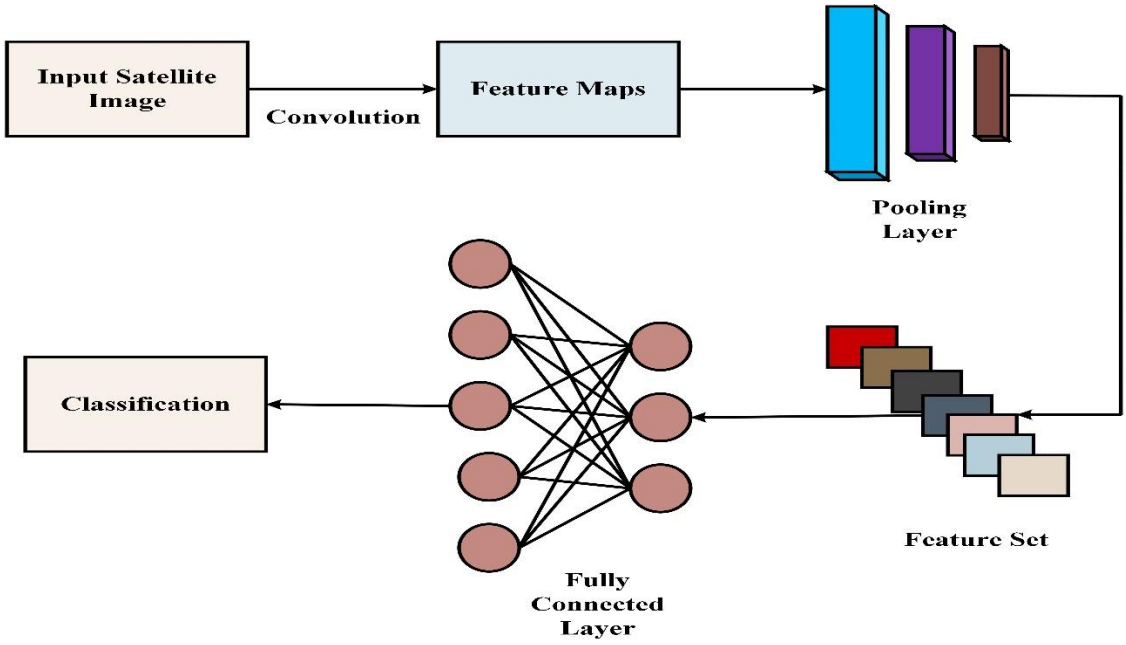


Fig. 2. CNN model.

V. Results and Discussion

The study first acquired a special EuroSAT dataset made up of twenty-seven thousand annotated satellite images from Sentinel-2 that included thirteen spectral bands and ten different land cover and land use classifications. To improve the quality and usability of these images for subsequent evaluation, necessary pre-processing techniques such as normalization and histogram equalization were used. The main contribution of this research is the combination of CNNs and PSO and ACO optimization approaches to enhance the accuracy of land cover and land use categorization. PSO was utilized to effectively choose the most pertinent spectral bands, decreasing data noise and redundancies. ACO significantly improved crucial CNN hyperparameters including batch size and learning rate, which improved the efficiency of the framework as a whole. On the EuroSAT dataset, the study assessed the hybrid PSO-ACO-CNN architecture and contrasted its performance with that of conventional categorization techniques and independent CNN models.

A. Performance Evaluation

To evaluate the success of categorization, assessment indicators are crucial. The method most frequently used for this objective is an estimation of precision. A classifier's accuracy for any particular set of data may be assessed by the proportion of test datasets that it properly classifies. Because making the optimal decisions will not be possible if the accuracy metric is used alone. To evaluate the performance of the classifier, researchers additionally employed other factors. Accuracy, recall, precision, and F1-score measures were used to evaluate the performance of the suggested technique. The following is a description of each measure's definitions:

T_{pos} (True Positive) refers to the amount of information that has been correctly categorized.

The term $Fpos$ (False Positive) represents the volume of reliable information that was incorrectly categorized.

False negatives (F_{neg}) are instances where incorrect information has been given an actual classification.

The categorization of incorrect information values is referred to as T_{neg} (True Negative).

The classifier's accuracy displays how frequently it makes the right assumption. The ratio of accurate forecasts to all other credible hypotheses is known as accuracy. It is demonstrated by Eq. (15).

$$Accuracy = \frac{T_{pos}+T_{neg}}{T_{pos}+T_{neg}+F_{pos}+F_{neg}} \quad (15)$$

The amount of correctly classified outcomes is determined by calculating the precision, or level of accuracy, of a classifier. Reduced false positives are the result of improved accuracy, whereas many more are the result of decreased precision. The percentage of instances that are correctly categorized compared to all occurrences is the definition of precision. It is defined by Eq. (16).

$$P = \frac{Tpos}{Tpos+Fpos} \quad (16)$$

The sensitivity of a categorization, or how much relevant information it produces, are determined by recall. The overall quantity of F_{neg} reduces with improved recall. Recall is the ratio of cases that have been correctly categorized to all of the predicted occurrences. This is demonstrable by Eq. (17).

$$R = \frac{Tpos}{Tpos+Fneg} \quad (17)$$

The combination of metrics known as F-measure, which reflects the weighted mean of recall and accuracy, are obtained by adding precision and recall. It is characterised by Eq. (18).

$$F1 - score = \frac{2 \times precision \times recall}{precision \times recall} \quad (18)$$

Area under the ROC Curve, or AUC, is a well-known evaluation metric for binary classification problems in deep learning and machine learning. The area under the receiver operating characteristic (ROC) curve, which is a graphic representation of the binary identification algorithm's efficacy, is evaluated by the area under the curve (AOC). The classifier in a binary classified problem tries to figure out whether the input information belongs to a positive or negative division. The $Tpos$ vs. the $F_{pos}$is shown on the ROC curve for various classification criteria. AOC values range from 0 to 1, with higher numbers denoting more efficiency. An absolutely randomized classifier has an AOC of 0.5, whereas an optimal classifier has an AOC of one. Because the method considers all possible degree of detection and provides a single number to compare the performance of different classifiers.

A deep learning model's training and testing accuracy score over a number of training epochs are summarized in Fig. 3. Every row displays the associated training accuracy and testing accuracy for an epoch number that ranges from 10 to 100. Testing accuracy assesses the model's effectiveness on new or validation information, whereas training accuracy shows how effectively the model is effective on the training information it was shown during training. Both training and testing accuracy often increase as the number of training epoch's rises, suggesting that the framework is learning from the information and getting more proficient in generating predictions. The model attains exceptionally accurate levels on both the training and testing datasets by the end of 100 epochs, indicating that it has acquired the ability to generalize to new, unanticipated information successfully. The graph shows the growth of the model's effectiveness as it goes through training.

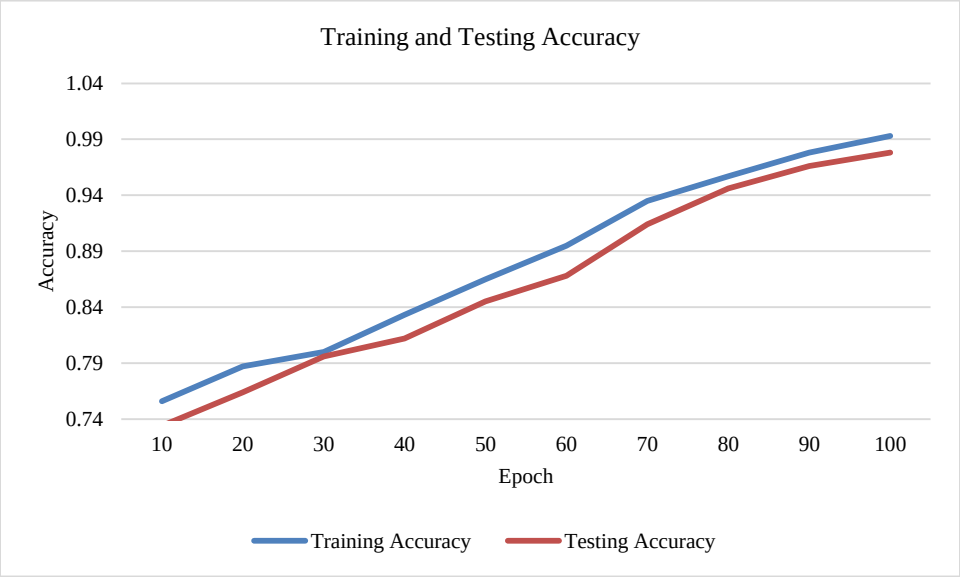


Fig. 3. Training and testing accuracy.

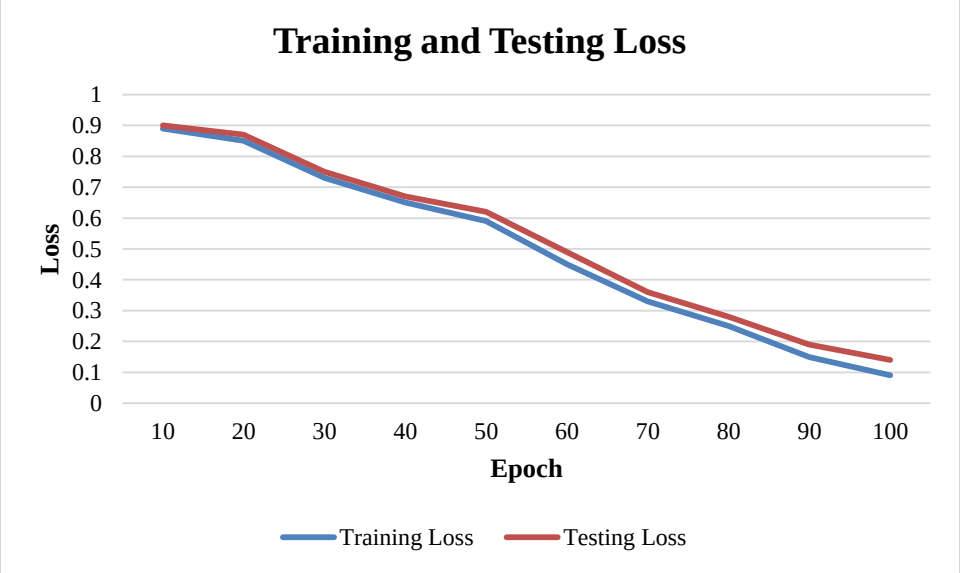


Fig. 4. Training and testing loss.

The testing and training loss values for a deep learning model throughout a variety of training epochs are shown in Fig. 4. The testing loss and the training loss are shown in each row, which is associated with an individual epoch number between 10 and 100. Testing loss evaluates the model's effectiveness on observed or validation information, where usually lower values indicate higher generalization. Training loss examines how well the model fits the training information, with lower values suggesting a better fit. This graph shows that both training and testing loss constantly reduce as the number of training epochs rises. The pattern indicates that if the model is trained, it becomes better at reducing errors and making predictions that are more accurate. The model's decreasing loss values show how it learns, and the lowest losses after 100 iterations show that the model has successfully converged and is capable of making accurate projections on both the training and testing datasets.

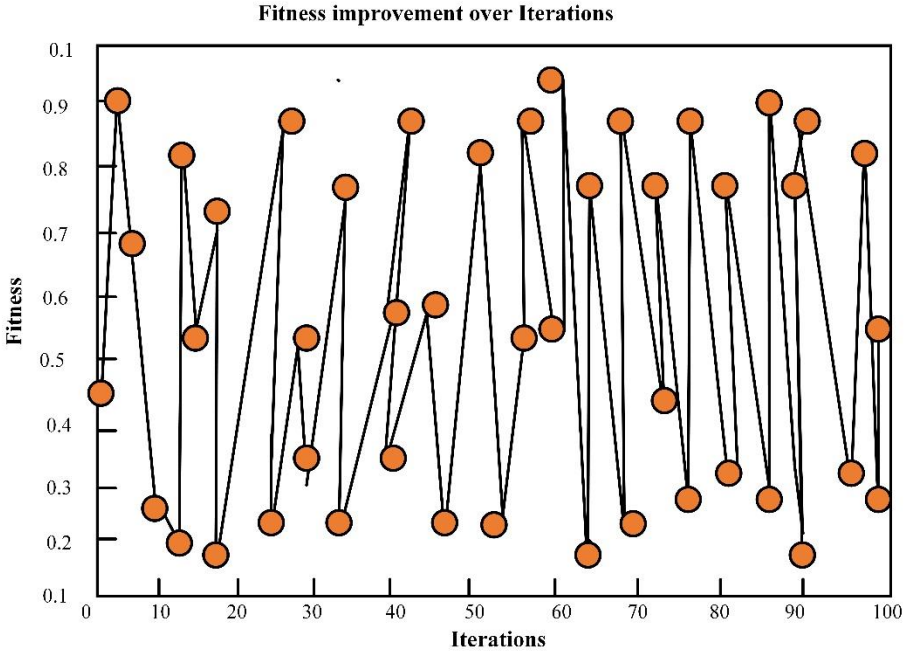


Fig. 5. Fitness improvement over iteration.

The development of an optimization algorithm—specifically, Ant Colony Optimization—over a number of iterations is seen in Fig. 5. The y-axis shows the fitness of the algorithm's created solutions, while the x-axis shows the quantity of iterations or generations. In ACO, fitness often refers to how well or effectively a solution addresses the issue at hand. The algorithm continually updates and improves its solutions to increase their fitness as it moves through iterations. As a result, the graph depicts how the solutions' fitness changes over time and, ideally, converges to an optimum or substantially optimal solution. Any levelling out or stability in the graph's later iterations denotes that the algorithm has probably achieved an optimal solution or a point of decreasing effectiveness. The sharp decrease or large loss in fitness towards the beginning of the graph's iterations signals rapid improvement. This illustration assists in evaluating the algorithm's rate of convergence and potency in locating superior solutions to the current optimization challenge.

The performance metrics of the HPSO-ACO-CNN hybrid deep learning model are summarized in Fig. 6. In order to evaluate the model's performance in a classification position, it offers important assessment metrics. The "Accuracy" statistic measures the model's overall accuracy in making predictions, and a high result of 99.3% shows that the model performs well in terms of categorization. "Recall" (98.7%) assesses the model's capability to properly recognize every single positive example, while "Precision" (99.2%) measures the model's capacity to correctly categorize positive cases. Precision and recall are combined into one score called the "F1-Score" (98.7%), which takes into account the trade-off between both. The HPSO-ACO-CNN model is very accurate and dependable in its categorization task, with an especially strong capacity to categorize positive situations properly while retaining a high overall accuracy level, according to these high values across all metrics.

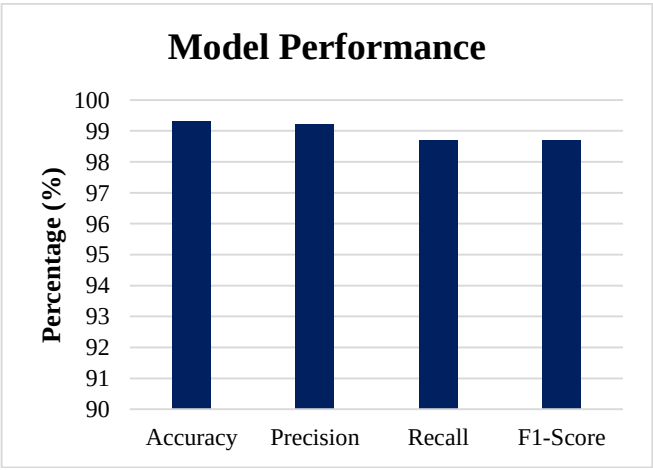


Fig. 6. Model performance.

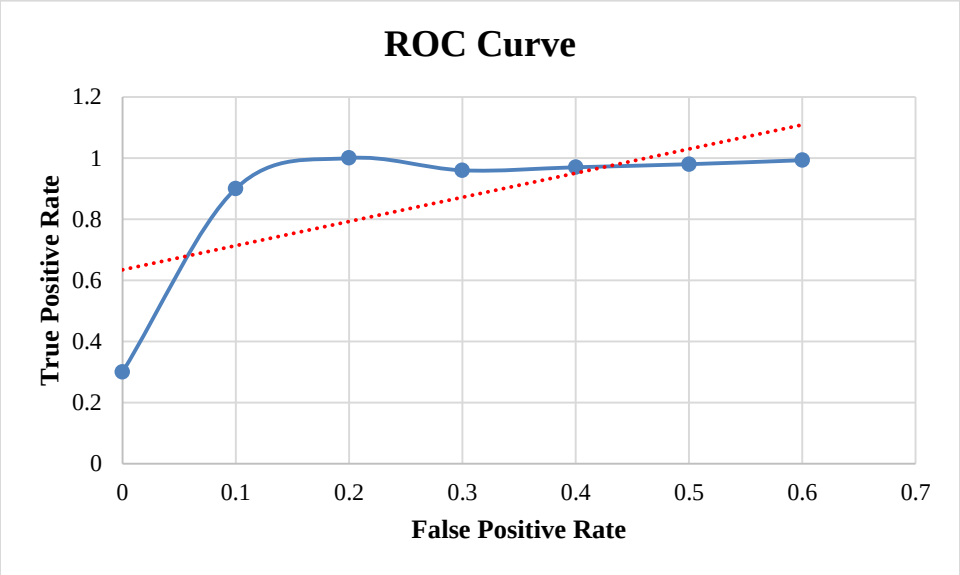


Fig. 7. ROC curve.

The True Positive Rate and False Positive Rate values for a binary categorization model at various threshold levels are shown in Fig. 7. These numbers are frequently employed to build a ROC curve. The fraction of real positive cases that the model properly classifies as positive is measured by TPR, sometimes referred to as sensitivity or recall.

TABLE I. COMPARISON OF PERFORMANCE METRICS OF PROPOSED METHOD WITH OTHER EXISTING APPROACHES

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DNN [25]	88.2	87.89	90	90
MSVM [26]	93.70	94.90	96.50	94.90
LSTM [27]	97.40	97.80	98.70	97.80
Proposed HPSO-ACO-CNN	99.3	99.2	98.7	98.7

On the other hand, FPR measures the percentage of real negative cases that the model misclassifies as positive. The graph displays how these rates alter when the threshold for categorization changes from 0 to 0.6. The TPR typically rises as the threshold rises, showing that the model gets better at properly recognizing positive situations but frequently at the expense of a larger FPR. The ROC curve, created from these results, graphically illustrates the trade-off between TPR and FPR at various threshold levels, assisting in evaluating the model's categorization effectiveness and determining the best threshold in accordance with the demands of the particular application.

The suggested HPSO-ACO-CNN is a hybrid of the Deep Neural Network (DNN), Multiclass Support Vector Machine (MSVM), Long Short-Term Memory (LSTM), and Deep Neural Network for a specific task. The Table I and Fig. 8 provides a number of significant efficiency measures for each model, one for each row: "Precision" measures the model's capacity to accurately classify positive cases, "Recall" measures the model's capacity for correctly recognizing all actual positive cases, and "F1-Score" is a balanced metric combining precision and recall. "Accuracy" denotes the overall proportion of correct predictions generated by the model.

The outcomes show that the suggested HPSO-ACO-CNN model exceeds the competition with the greatest values for accuracy (99.3%), precision (99.2%), and F1-Score (98.7%), demonstrating its better performance in the task at hand. Additionally, LSTM performs well, whereas DNN and MSVM score slightly more severe on these criteria. Together, these measures offer insightful comparisons of these models' success in the particular categorization task, with higher values representing better model effectiveness.

TABLE II. COMPARISON OF DATASETS OF PROPOSED METHOD WITH OTHER EXISTING APPROACHES

Datasets	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Landsat 8 Imagery 2014 [28]	89	87	87	88
Landsat 5 Thematic Mapper Imagery [29]	94.17	95	96	94
Proposed EuroSAT Dataset	99.3	99.2	98.7	98.7

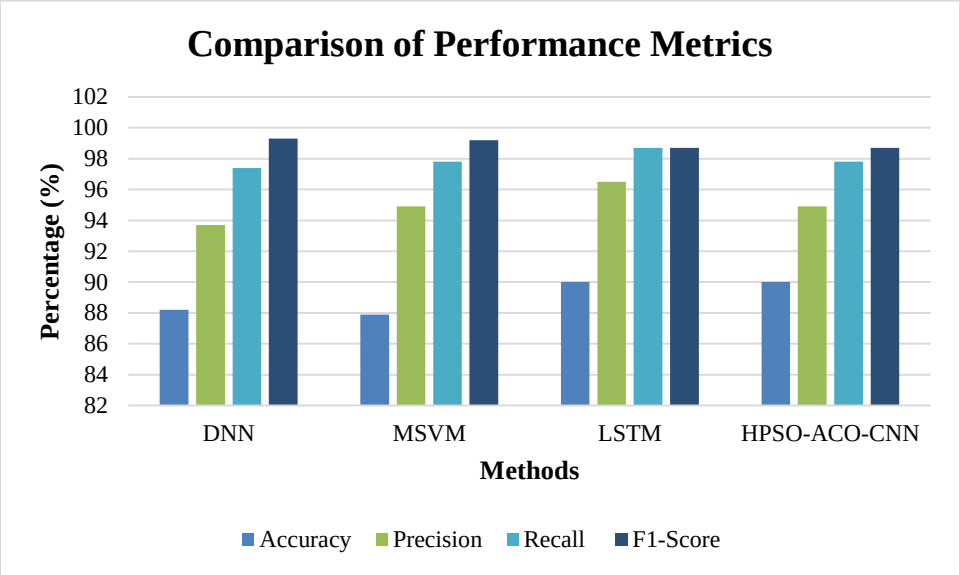


Fig. 8. Comparison of performance metrics of proposed method with other existing approaches.

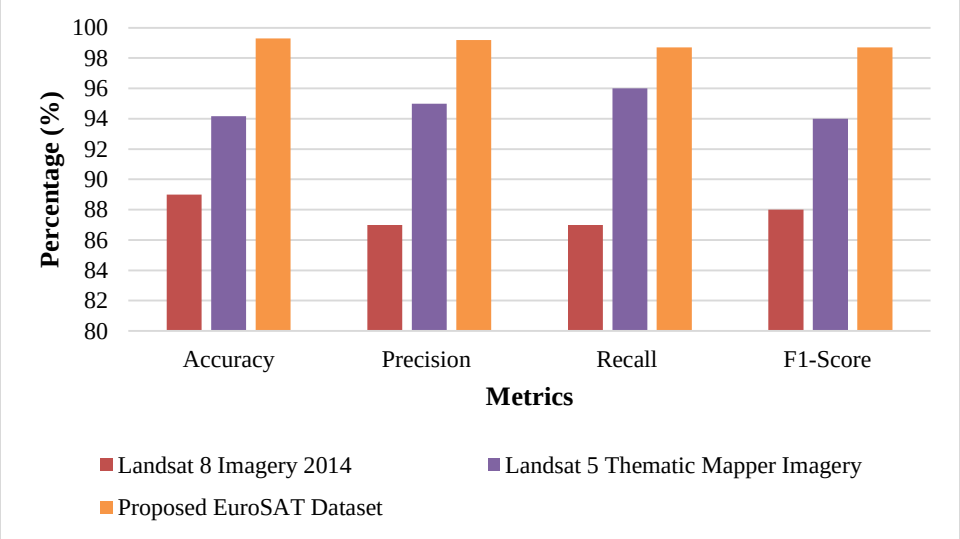


Fig. 9. Comparison of datasets of proposed method with other existing approaches.

A comparison of datasets utilizing the suggested approach in comparison to other current methodologies is shown in Table II and Fig. 9. Four important criteria are used to assess their effectiveness: F1-Score, Accuracy, Precision, and Recall. The initial dataset, the 2014 Landsat 8 Imagery, has an accuracy of 89%, 87% recall, 87% precision, and an 88% F1-Score. Landsat 5 Thematic Mapper Imagery, the second dataset, was particularly better than the initial one, with an F1-Score of 94%, accuracy of 94.17%, precision of 95%, and recall of 96%. The suggested EuroSAT Dataset performed outstandingly, achieving 98.7% recall, 99.2% precision, 99.3% accuracy, and a 98.7% F1-Score. These findings show that the suggested EuroSAT Dataset outperforms the other datasets in all four standards, indicating that it is the most effective alternative for the given objective, which is probably connected to the categorization or analysis of satellite images. Variability in data features, including resolution, spectral bands, and landscape variety, might be the cause of the variances in comparison results between datasets. The suggested methods could perform better on datasets whose properties are comparable to those that were encountered during the development of the EuroSAT dataset. These datasets might include high-resolution and diversified satellite images.

B. Discussion

The findings show that the proposed HPSO-ACO-CNN model has a number of benefits over other machine learning techniques already in utilization for the categorization of land use and land cover using satellite images. With an accuracy of 99.3%, precision of 99.2%, recall of 98.7%, and an F1-Score of 98.7%, the HPSO-ACO-CNN model outperformed in all assessment measures. These findings demonstrate that the hybrid technique, which combines a CNN with PSO and ACO, significantly improves the classification capabilities of the model. The model excels at accurately detecting positive

instances while reducing false positives and false negatives, as seen by its excellent accuracy and recall scores. A precise categorization of land cover and land use is essential in applications like environmental monitoring and catastrophe management. While DNN and MSVM are reasonable models in comparison, they fall short of HPSO-ACO-CNN's performance. Although the LSTM model also exhibits comparable performance, HPSO-ACO-CNN stands out because to its greater accuracy and precision. These results illustrate the effectiveness of combining deep learning methods with optimization algorithms, emphasizing the potential for more precise and reliable mapping of land use and land cover in the context of sustainable land management and protecting the environment.

VI. Conclusion and Future Works

The study concludes by presenting a novel technique that substantially enhances the precision of classifying land use and land cover using satellite images. The merging of ACO, CNN, and HPSO algorithms results in significant performance increases in the proposed HPSO-ACO-CNN model. Combining CNN hyperparameter optimization with spectral band selection yields remarkable accuracy, precision, recall, and F1-Score performance for this hybrid architecture. Results from experiments conducted on the EuroSAT dataset demonstrate how well the HPSO-ACO-CNN model performs when compared to other methods and standalone CNN models. In addition to addressing important problems with feature selection, parameter optimization, and model training, the work creates new opportunities for satellite image analysis. This novel method has great potential for a number of uses, such as sustainable land use, urban planning, environmental monitoring, and disaster management. It highlights how deep learning techniques and optimization strategies may be combined to improve remote sensing applications. Regarding potential avenues for future research, there are a number of intriguing options to consider. An intriguing line of investigation is the expansion of the HPSO-ACO-CNN architecture to handle larger and more complicated datasets of satellite images, potentially incorporating other spectral bands and land cover categories. Additionally, assessing the model's resilience and scalability in various environmental conditions and geographical areas may yield unexpected findings. Finally, there is potential to further the more general goals of environmental conservation and sustainable land management by investigating applications of transfer learning and customizing the model for additional Earth observation tasks, such change detection and crop monitoring.

References

[1] "Application of Google earth engine python API and NAIP imagery for land use and land cover classification: A case study in Florida, USA - ScienceDirect." Accessed: Sep. 18, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S157495412100265X

[2] "A deep learning framework for land-use/land-cover mapping and analysis using multispectral satellite imagery | SpringerLink." Accessed: Sep. 18, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s00521-019-04349-9

[3] "Scale Sequence Joint Deep Learning (SS-JDL) for land use and land cover classification - ScienceDirect." Accessed: Sep. 18, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0034425719306133

[4] "Remote Sensing | Free Full-Text | Mapping Land Use from High Resolution Satellite Images by Exploiting the Spatial Arrangement of Land Cover Objects." Accessed: Sep. 18, 2023. [Online]. Available: https://www.mdpi.com/2072-4292/12/24/4158

[5] "Applied Sciences | Free Full-Text | Integrating Convolutional Neural Network and Multiresolution Segmentation for Land Cover and Land Use Mapping Using Satellite Imagery." Accessed: Sep. 18, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/11/12/5551

[6] "Full article: Land use/land cover and land surface temperature analysis in Wayanad district, India, using satellite imagery." Accessed: Sep. 18, 2023. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/19475683.2020.1733662

[7] "Full article: Methodological evaluation of vegetation indexes in land use and land cover (LULC) classification." Accessed: Sep. 18, 2023. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/24749508.2019.1608409

[8] S. Amini, M. Saber, H. Rabiei-Dastjerdi, and S. Homayouni, "Urban Land Use and Land Cover Change Analysis Using Random Forest Classification of Landsat Time Series," Remote Sensing, vol. 14, no. 11, p. 2654, Jun. 2022, doi: 10.3390/rs14112654.

[9] "Applied Sciences | Free Full-Text | Sentinel-2 Satellite Imagery for Urban Land Cover Classification by Optimized Random Forest Classifier." Accessed: Sep. 18, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/11/2/543

[10] "Statistical features for land use and land cover classification in Google Earth Engine - ScienceDirect." Accessed: Sep. 18, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2352938520306340

[11] "A machine learning-based classification of LANDSAT images to map land use and land cover of India - ScienceDirect." Accessed: Sep. 18, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2352938521001609

[12] "Assessment of land-use and land-cover changes in Pangari watershed area (MS), India, based on the remote sensing and GIS techniques | Applied Water Science." Accessed: Sep. 18, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s13201-021-01425-1

[13] "Using Landsat satellite data for assessing the land use and land cover change in Kashmir valley | GeoJournal." Accessed: Sep. 18, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s10708-019-10037-x

[14] "Remote Sensing | Free Full-Text | Improvement in Satellite Image-Based Land Cover Classification with Landscape Metrics." Accessed: Sep. 18, 2023. [Online]. Available: https://www.mdpi.com/2072-4292/12/21/3580

[15] J. V. Solórzano, J. F. Mas, Y. Gao, and J. A. Gallardo-Cruz, "Land Use Land Cover Classification with U-Net: Advantages of Combining Sentinel-1 and Sentinel-2 Imagery," Remote Sensing, vol. 13, no. 18, p. 3600, Sep. 2021, doi: 10.3390/rs13183600.

[16] Md. J. Faruque et al., "Monitoring of land use and land cover changes by using remote sensing and GIS techniques at human-induced mangrove forests areas in Bangladesh," Remote Sensing Applications: Society and Environment, vol. 25, p. 100699, Jan. 2022, doi: 10.1016/j.rsase.2022.100699.

[17] G. Mehmood, A. Waheed, and M. Zareei, "Land-Cover Classification and its Impact on Peshawar's Land Surface Temperature Using Remote Sensing," Sep. 2021.

[18] G. Rousset, M. Despinoy, K. Schindler, and M. Mangeas, "Assessment of Deep Learning Techniques for Land Use Land Cover Classification in Southern New Caledonia," Remote Sensing, vol. 13, no. 12, p. 2257, Jun. 2021, doi: 10.3390/rs13122257.

[19] T. K. Das, D. K. Barik, and K. V. G. R. Kumar, "Land-Use Land-Cover Prediction from Satellite Images using Machine Learning Techniques," in 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India: IEEE, May 2022, pp. 338–343. doi: 10.1109/COM-IT-CON54601.2022.9850602.

[20] H. Fahmi and W. P. Sari, "Analysis of deep learning architecture for patch-based land cover classification," in 2022 6th International

Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia: IEEE, Dec. 2022, pp. 1–5. doi: 10.1109/ICITISEE57756.2022.10057895.

[21] J. Yan et al., "Land-Cover Classification With Time-Series Remote Sensing Images by Complete Extraction of Multiscale Timing Dependence," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 1953–1967, 2022, doi: 10.1109/JSTARS.2022.3150430.

[22] V. N. Vinaykumar, J. A. Babu, and J. Frnda, "Optimal guidance whale optimization algorithm and hybrid deep learning networks for land use land cover classification," EURASIP J. Adv. Signal Process., vol. 2023, no. 1, p. 13, Jan. 2023, doi: 10.1186/s13634-023-00980-w.

[23] B. Van Leeuwen, Z. Tobak, and F. Kovács, "Machine Learning Techniques for Land Use/Land Cover Classification of Medium Resolution Optical Satellite Imagery Focusing on Temporary Inundated Areas," Journal of Environmental Geography, vol. 13, no. 1–2, pp. 43–52, Apr. 2020, doi: 10.2478/jengeo-2020-0005.

[24] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Introducing Eurosat: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," in IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia: IEEE, Jul. 2018, pp. 204–207. doi: 10.1109/IGARSS.2018.8519248.

[25] "Full article: Deep neural network ensembles for remote sensing land cover and land use classification." Accessed: Oct. 21, 2023. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/17538947.2021.1980125

[26] "Remote Sensing | Free Full-Text | Satellite Image Classification Using a Hierarchical Ensemble Learning and Correlation Coefficient-Based Gravitational Search Algorithm." Accessed: Oct. 21, 2023. [Online]. Available: https://www.mdpi.com/2072-4292/13/21/4351

[27] "Multitemporal Relearning With Convolutional LSTM Models for Land Use Classification | IEEE Journals & Magazine | IEEE Xplore." Accessed: Oct. 21, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9343734/

[28] "Using Landsat satellite data for assessing the land use and land cover change in Kashmir valley | GeoJournal." Accessed: Oct. 21, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s10708-019-10037-x

[29] "Sustainability | Free Full-Text | Detecting and Analyzing Land Use and Land Cover Changes in the Region of Al-Jabal Al-Akhdar, Libya Using Time-Series Landsat Data from 1985 to 2017." Accessed: Oct. 21, 2023. [Online]. Available: https://www.mdpi.com/2071-1050/12/11/4490

IAM-TSP: Iterative Approximate Methods for Solving the Travelling Salesman Problem

Esra'a Alkafaween[1], Samir Elmougy[2], Ehab Essa[3], Sami Mnasri[4], Ahmad S.Tarawneh[5], Ahmad Hassanat[6]
E-Learning and Education Resources Center, Mutah University, Karak, Jordan[1]
Department of Computer Science, Mansoura University, Mansoura, Egypt[2, 3]
Department of Computer Science, University of Tabuk, Tabuk, Saudi Arabia[4]
Computer Science Department, Mutah University, Karak, Jordan[5, 6]

***Abstract*—TSP is a well-known combinatorial optimization problem with several practical applications. It is an NP-hard problem, which means that the optimal solution for huge numbers of examples is computationally impractical. As a result, researchers have focused their efforts on devising efficient algorithms for obtaining approximate solutions to the TSP. This paper proposes Iterative Approximate Methods for Solving TSP (IAM-TSP), as a new method that provides an approximate solution to TSP in polynomial time. This proposed method begins by adding four extreme cities to the route, a loop, and then adds each city to the route using a greedy technique that evaluates the cost of adding each city to different positions along the route. This method determines the best position to add the city and the also the best city to be added. The resultant route is further improved by employing local constant permutations. When compared to existing state-of-the-art methods, our experimental results show that the proposed method is more capable of producing high-quality solutions. The proposed approach, with an average approximation of 1.09, can be recommended for practical usage in its current form or as a pre-processing step for another optimizer.**

Keywords*—Greedy algorithms; TSP; NP-Hard problems; polynomial time algorithms; combinatorial problems, optimization methods

I. INTRODUCTION

At the beginning of the seventeenth century, Thomas Penynington and William Hamilton modeled the first mathematical problem corresponding to the Traveling Salesman Problem (TSP). This problem represents a game in which the winner connects twenty points by moving from one point to another, following some precise paths. This game/problem is called Hamilton Circuit Theory [1]. Afterward, in graph theory, TSP becomes a problem of identifying the optimal (shortest) Hamiltonian cycle visiting a set of cities (points) using a matrix of distances between the cities. TSP is one of the classic problems of combinatorial optimization, and it is widely investigated and considered a standard for assessing the performance of computational methodologies and algorithms.

The principal objective of TSP is that the salesman visits all the cities in the minimal tour and then returns to the starting point with the assumption that the distances between the cities are known and the necessity of visiting each city is only once. Despite high-size instances of TSP being resolved in the literature, TSP is proven to be NP-hard even for small-size instances [2].

The theoretical and practical relevance of the TSP problem comes from the fact that numerous applicable real-world and engineering problems can be solved by adapting the TSP solutions, such as circuit design [3], scheduling [4], and DNA [5, 6], in addition to logistics and transportation and even space exploration missions [7]. Therefore, the study of TSP has significant theoretical and practical value, leading to cost savings and other benefits.

Moreover, the application fields of TSP include any problem involving the search for the shortest paths. These fields vary in domains from big data classification [8, 9, 10], deployment and routing of IoT networks [11, 12, 13, 14, 15, 16], financial prediction [17, 18], image processing [19], and [20], computer vision [21], [22], and [23]. TSP can be defined by finding a Hamiltonian cycle that visits each vertex precisely once, with the least amount of total weight feasible, given a full undirected weighted graph G = (V, E) with vertex set V and edge set E. The total weight of a cycle is equal to the weights of each of its individual edges. The book in [24] presents a comprehensive mathematical and theoretical analysis of the calculability and complexity of the TSP. One of the critical issues of methods of resolution of TSP is the enhancement of the accuracy of the algorithm to rapidly find the optimal or near-optimal solutions.

Indeed, heuristics and meta-heuristics are the most successful methods used to resolve the TSP [25, 26, 27]. In this regard, Genetic Algorithms (GA), Ant Colony Optimization (ACO), Simulated Annealing (SA), and Particle Swarm Optimization (PSO) are among the most commonly used algorithms to resolve TSP. GA was successfully used for problems involving global search due to its rapid convergence and fast search process. However, it has some issues and has poor performance when achieving the local search. ACO is another robust optimizer with a high-resolution capacity. However, its convergence performance highly depends on the initial parameters, mainly the appropriate initial quality of the pheromone. PSO is characterized by its ease of use since it has a few numbers of initial parameters to set. However, other optimizers give better results for many TSP test problems. SA is another optimizer known for its capacity to avoid local optima and is often hybridized with other algorithms for this capacity.

However, like with ACO, the initial parameter values have a significant impact on the search process. Despite substantial research on TSP, its combinatorial nature suggests that there is still opportunity for development in this field, motivating the work of this paper. TSP is inextricably tied to handling other combinatorial problems with comparable characteristics, such as the Traveling Salesman Problem with Time Windows (TSPTW), Vehicle Routing Problem (VRP), and Capacitated Vehicle Routing Problem (CVRP). TSP approaches and algorithms are frequently used as a foundation for addressing these related optimization challenges. As a result, progress in TSP solving may provide useful insights and tactics relevant to a broader class of combinatorial problems in logistics, transportation, and network optimization. Because these challenges are interdisciplinary, knowledge and approaches can be transferred across fields, creating a more thorough grasp of combinatorial optimization.

The major contribution of this paper lies in presenting an Iterative Approximate Method for Solving TSP (IAM-TSP), which provides a polynomial-time approximate solution to the TSP. The proposed method begins by including four extreme cities along the path. Following that, a loop successively adds each city to the route by calculating the cost at various points along the route. The approach considers each city to add and finds the most suitable location to put it, the one that minimizes the total cost. This method is improved further by using local constant permutations on the output of IAM-TSP, which considerably improves the final route, referred to as IAM-TSP+.

We evaluated the proposed algorithms against standard TSP datasets to determine how they performed in terms of key performance measures. The proposed IAM-TSP performed nearly identically to some of the typical TSP approximation algorithms given in the results section, but the proposed IAM-TSP+ surpassed all approximation methods compared on the majority of TSP instances.

The rest of the paper is organized as follows. Section II illustrates and discusses the advantages and drawbacks of the recent most relevant studies resolving the TSP. Section III identifies and investigates the proposed methodology for resolving the TSP. Section IV presents the experimental setup and the results, Section V concludes the study.

II. Related Work

Given TSP's extensive history, its cutting-edge landscape is extremely diverse. To solve the TSP, several techniques, paradigms, and approximations algorithms have been used. Heuristics, for example, are a type of approximation method aimed to finding near-optimal solutions to NP-hard problems in polynomials time. One popular and simple method for building a TSP tour involves starting the tour at any node, traversing minimum-cost arcs to each successive node until all nodes are visited, and then returning to the starting node to finish the tour. This method is known as the Nearest Neighbor heuristic [28].

There are also many popular methods, such as: the 2-Opt heuristic [29], Farthest Insertion algorithm [30], Nearest Insertion algorithm [30], Cheapest Insertion algorithm [30], Arbitrary Insertion algorithm [31], Repetitive Nearest Neighbor algorithm [32], Concave hull with heuristics, and Concave Hull No Heuristic [33]. In what follows, the main recent studies proposing TSP resolution methodologies are investigated: The study in [33] introduces a concave hull-based algorithm to resolve the Euclidean symmetric TSP by establishing concentric hulls and then merging them in one tour. Two novel metrics are suggested: the “Average Waiting Distance” and the “min AWD” of a tour. The results show that an optimal AWD does not guarantee the optimal tour.

In study [34], the authors enhanced the accuracy of TSP solutions for numerous sizes of the problem. The used algorithm is a modified ACO with a better convergence during the TSP search process. To prevent trapping into local optima, this algorithm decreases the high concentration of pheromone during the route selection. The diversity in the algorithm is ensured using entropy weighted learning. According to the results, this work improved ACO and solved TSP better than the standard ACO.

A new variant of TSP, called TSPJ, which includes the schedule of jobs in a set of positions, was introduced in study [35]. Since in TSP, the transport time is longer than the operation time, the considered objective in TSPJ is to minimize the make-span, equal to the needed time to achieve the longest job. The resolution of the TSPJ involves four local search methods. Using the CPLEX system, the results indicated that the solutions given by the four used heuristics are too close to the optimal (a gap less than 6%). In the same regard, the study in [36] aims to resolve the TSPJ. Applicative and practical contexts of TSPJ had been discussed, as well as the parameters specific to TSPJ such as the completion time, configuration time, and resource variation.

In study [37], the authors used the aim was to resolve a TSP variant called the multiple TSP (mTSP) problems using a hybrid algorithm. The latter relies on an EAX heuristic to optimize the intra-tour and a tabu search neighborhood search to optimize the inter-tour. The objectives of the problem were the minimization of both the longest path and the total traveled distance. To reduce the neighborhood search time, a reduction approach is proposed to avoid computing the nonpromising possible solutions. The experiments involve a comparison with five approaches tested on 41 known TSP test problems and 36 new large-size ones. However, the other variants of TSP were not assessed, and comparisons with other classes of heuristics were not achieved.

The study in [38] suggests a new seriation strategy named “tree-penalized Path Length” (tpPL). Data seriation is a famous problem in data analysis. Itis the process of sequencing and ordering data according to their similarity. TSP in this study is considered a seriation method. The goal is to linearly order the data using simultaneously the TSP, tpPL, and optimal leaf order (OLO) methods. Optimal paths are transferred from TSP to OLO. In terms of computational complexity, TSP and tpPL have the same order of complexity. Hence, TSP heuristics may be used to resolve the tpPL. Tested on more than forty datasets, the tpPL has a better performance than TSP and OLO with the same computational complexity. The study in [39] introduces a compression-based TSP heuristic for data micro-aggregation.

Micro-aggregation is a method for disrupting and aggregating personal data using the concept of k-anonymity. By simultaneously considering the respect for privacy and the usefulness of data, the introduced TSP heuristic was the most efficient in solving the problem of micro-aggregation. In contrast, heuristics relying on TSP encrust scalability issues. Unlike other heuristics, the algorithm proposed in this study can reduce the execution time of the TSP. The tests carried out on small and medium-sized data affirm the trade-off between computation time and the rate of loss of micro-aggregation information.

In study [40], a new problem, named the traveling thief problem (TTP), is proposed by combining the knapsack problem (KP) and the TSP. The PTT resolution method relies on sequences selection heuristics. A set of selection operators/thieves are involved to select the cities/objects to progressively create the tour.

The study in [41] introduces two new TSP versions. Named pollution TSP(PTSP) and energy minimization TSP (EMTSP), the new versions add environmental constraints to the TSP. The aim of PTSP is to reduce fuel consumption, and carbon emissions. The aim of EMTSP is to reduce the cost of a trip according to the distance and the carried load. The method of resolution of the two TSP variants, MILP-GA-LS, relies on a mixed integer linear programming model to identify initial possible solutions, then multi-operator GA to enhance the found solutions. Afterward, an iterative local search algorithm was used to enhance the solutions. However, numerous constraints were not considered such as the time-window of customers. Moreover, an issue arises concerning the complexity of the introduced approach, since it was proven that, for small-size instances of PTSP and EMTSP, exact methods give better results than the MILP-GA-LS.

In study [42], a new ACO variant is proposed to solve the TSP. Named DAACO; this algorithm dynamically changes the number of ants to avoid falling into local optima and to prevent long time of convergence. Besides, DAACO uses a local selection process to enhance the quality of ants and the needed time for the search process. Among the used twenty TSPLIB test problems, all the DAACO solutions were optimal except one. These results confirm the advantageous quality of solutions and time of convergence of DAACO compared to other optimizers.

In study [43], a hybrid Ant Colony (AC)-Tabu Search (TS)-Firefly Algorithm (FA) called ACTS-FATS is proposed. The hybridization avoids the probability of premature convergence, then, reduces the chance of trapped in local optima. Tested with the TSPLIB95, the hybridization does not generate additional execution time compared to AC, TS and FA.

In study [44], Dhouib-Matrix, a column-row method, is proposed to resolve polynomial time TSP. The process of this method is as follows: after defining the distance matrix, a start position is selected to choose rows. Then, columns are discarded, and the route is transformed to a tour. The advantage of the introduced Dhouib Matrix is that it needs only n iterations to find the route between n cities.

In study [45], a comparative study is proposed to discuss the recent algorithms and methodologies used to resolve the TSP using metaheuristics. The focus is set on the numerous versions of the BA, FA and PSO optimizers.

In study [46], a new version of the TSP, called TSP-D is proposed. The latter is a classic TSP that involves a truck and a drone (Unmanned Aerial Vehicles (UAV)). In TSP-D, it is assumed that UAV has low battery capacities and can transport cargo per flight. The issue is to determine which UAV and which truck should serve which client. The tests demonstrate that the time of service can be reduced by using the UAVs with distinct speeds according to the cargo weight. However, the used samples (number of clients) are between 30 and 60, which makes the method valid only for small-size TDP-D instances.

The authors in [47] propose an algorithm to minimize travel costs and maximize the overall profit. Simulated annealing (SA) and genetic algorithm (GA) with dedicated mutation operators were used. A concept of tour plots is used to generate the final solutions. In terms of computation time, the SA is better than the GA.

In study [48], Qi-ACO, an ACO based on quantum computing, is developed to resolve the four-dimensional TSP (4DTSP). Fuzzy type-2 variables are used due to the uncertain aspect of the investigated problem of travel emissions and costs. The 4DTSP is characterized by the existence of numerous paths and conveyances between the cities. A process is implemented in Qi-ACO for generating qubits considering the constraints of carbon emission, cost, and time. The initialization and update of pheromone is based on qubit. A faster computation is achieved in Qi-ACO due to the quantum calculations. Performed statistical tests to confirm the performance of the introduced approach. However, numerous constraints are not considered such as vehicle speed and route selection. The readers are referred to [49] for more comprehensive methodologies on modeling and designing greedy heuristic methods to resolve the TSP.

III. The Proposed Iterative Approximate Methods for Solving TSP (IAM-TSP)

The IAM-TSP is proposed in this paper as an approximation to the TSP, which has a greedy algorithmic nature. It begins by selecting four cities from the input set to represent the east, north, west, and south most positions. These cities are added to the route list in the following order: east, north, west and south, followed by a return to the beginning point (east). The algorithm then enters a loop, adding each remaining city to the path one by one. The method calculates the cost of adding a city to various points along the route throughout each iteration. It carefully investigates all feasible positions for each city, picking the best position and city at the lowest cost. Finally, this approach computes the route's total cost and outputs the final result.

The cost is calculated using the Euclidean distance between the consecutive cities on the route. Algorithm 1 shows the pseudocode processes of the proposed IAM-TSP. Fig. 1 illustrates the progress of the IAM-TSP solution of the Rat195 LIBTSP real-world problem [50]. Because the TSP is a

minimization problem, the term "best" refers to the smallest value throughout this paper, i.e., minimum route.

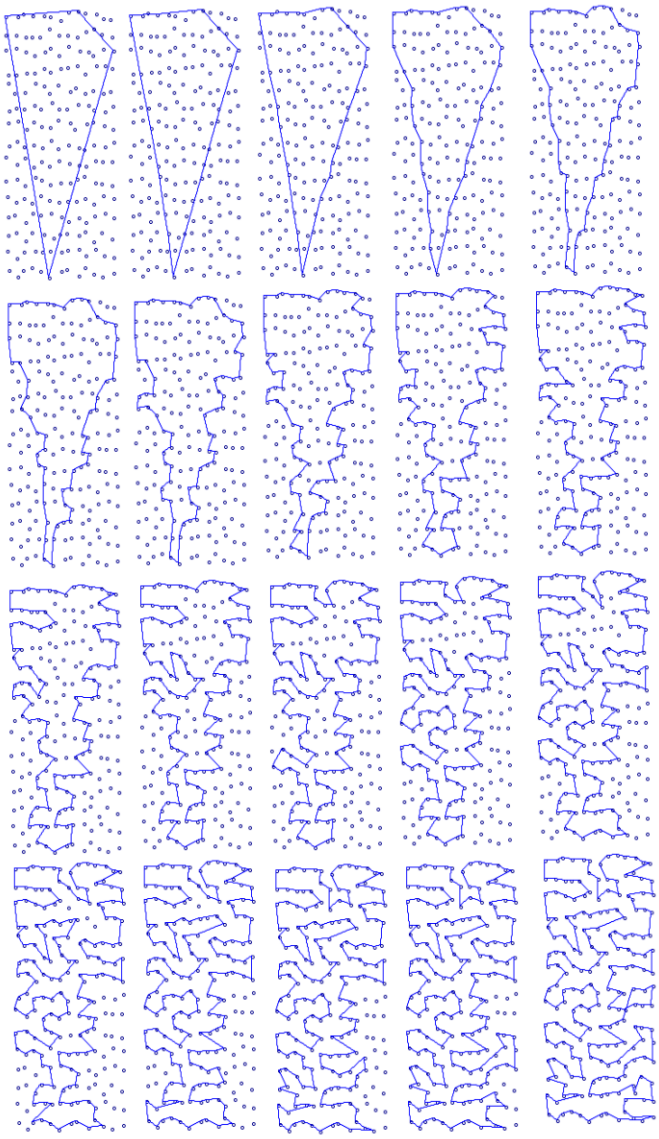


Fig. 1. The progress of the IAM-TSP solution of the Rat195 LIBTSP real-world problem, the progress is made from left to right and from top to bottom, and the results are displayed after ten iterations from the previous state.

According to Algorithm 1, the proposed IAM-TSP has a time complexity of $O(n^2)$, where n is the number of cities in the Cities x y list. This is because, in the while loop, the algorithm finds the best position to put each city from the Open list into rout and checks every feasible solution, which takes O(n) time. This operation is repeated until Open is empty, which means that in the worst scenario, the while loop runs n times, giving a total time complexity of $O(n^2)$.

The space complexity of the solution is $O(2n + n^2)$, where n is the number of cities in the Cities xy list. This is because we use two lists rout and open of size n, which take O(2n) space, and the quadratic space comes from the space reserved by the Possible Solutions matrix, whose size is n rows, each hosting n cities. The rest of the data structures used in the solution take O(1) or O(n) space, so the overall space complexity can be asymptotically approximated to $O(n^2)$. However, if we establish a square matrix storing the distances between each city and the others, space complexity stays asymptotically quadratic too; this is not done in this study, but it is a typical approach to removing the burden of distance computation. It should be emphasized that IAM-TSP provides an approximation solution for the TSP, and it does not ensure finding the optimal solution, because finding the best position for each city alone does not guarantee finding the optimal solution, which necessitates the involvement of all cities at the same time.

In order to improve the IAM-TSP performance, and because involving all cities makes the problem NP-hard, we opt for involving a constant number of local cities (k), and calculate all the permutation sequences starting from the first city in the output Route until the k city, finding the best solution and updating the Rout during this process, after which the algorithm goes into a loop moving by one city to find the next k permutations until $n - k$, this enhanced version is called IAM-TSP+. Algorithm 2 shows the pseudocode processes of the proposed IAM-TSP+, and Fig. 2 depicts the IAM-TSP+ resultant Route of the ATT48 TSPLIB real-world problem [50] in comparison to its optimal route.

Algorithm 1: The proposed IAM-TSP algorithm

Require: Cities xy (List of cities with x and y coordinates) of size n (number of cities)
Ensure: Cost (Total cost of the route using Euclidean distance), and Route (the best possible sequence of cities)
1: Create an empty list Route to store the order of visited cities.
2: Get the East, North, West, and South cities and add them to the Route list. This is done using the minimum/maximum of x and y coordinates.
3: Initialize a Boolean list Visited of size n (number of cities) and set all elements to false.
4: Set the Visited status of the cities in Route to true (East, North, West, and South).
5: Create an empty list Open to store the cities that have not been visited.
6: Add all cities in Cities xy to Open if their Visited status is false.
7: while Open is not empty do
8: Create an empty list of list Possible Solutions to store the possible solutions (routs).
9: for each city i in Open do
10: Find the best location to insert it into Route and add the new route to possible Solutions.
11: end for
12: Find the best solution Best in Possible Solutions and keep track of the city ID.
13: Update Route by Best.
14: Remove the city ID from Open that satisfies Best.
15: end while
16: Calculate the cost of the Route using the Euclidean distance and store it in
Cost.
17: return Cost, Route

Algorithm 2: The proposed IAM-TSP+ algorithm

Require: Cities xy of size n and k=5 (local permutations)
Ensure: Cost, Route
1: Route=IAM-TSP(Cities xy)
2: for i = 1 to n − k do
3: Create an empty list of list Possible Solutions
4: Possible Solutions= all k local permutations of Route(from i to i + k)
5: Update Cost(Possible Solutions)
6: index = *argmin*(Cost)
7: Route = Possible Solutions[index]
8: end for
9: return Cost(Route), Route

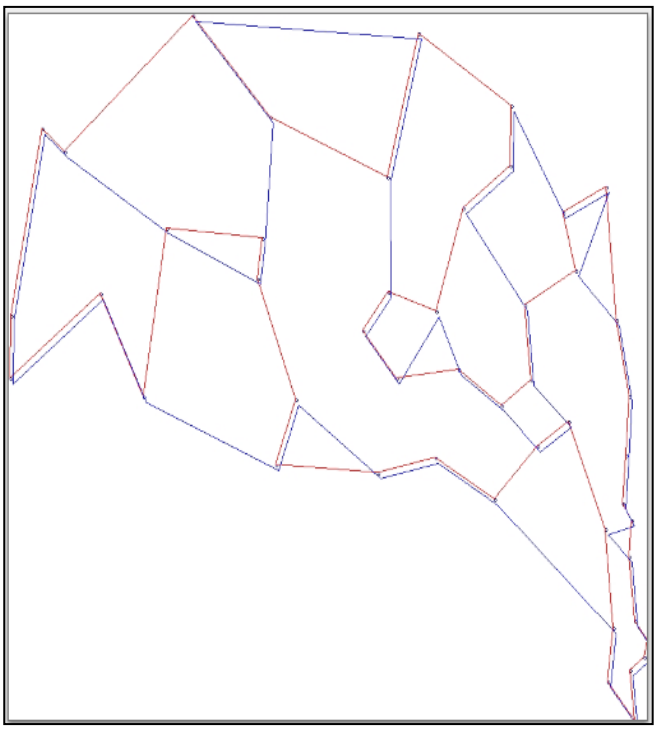


Fig. 2. The IAM-TSP+ resultant route of the ATT48 TSPLIB real-world problem (in blue) n comparison to its optimal route (in red), k = 5, IAM-TSP+ cost = 34877, and the optimal cost = 33523. The optimal route is a little shifted right and down to avoid edge overdrawing.

According to Algorithm 2, the proposed IAM-TSP+ has a time complexity of $O(n^2 + n.k!)$, where n is the number of cities in the Cities xy list. This is because the improved version uses the proposed IAM-TSP as an initial solution, this time is added to the time consumed by sliding the k permutations along the Route, which consumes O(n.k!) in the worst scenario, giving a total time complexity of $O(n^2 + n.k!)$. Since k is a constant number, the time complexity of O(n.k!) is often greater than that of $O(n^2)$. O(n.k!) may still be more efficient than $O(n^2)$ for small values of k, but as (k) grows higher, the growth rate of O(n.k!) quickly exceeds that of $O(n^2)$. Therefore, the time complexity of IAM-TSP+ can be asymptotically approximated to O(n.k!). The space complexity of the IAM-TSP+ is similar to that of the proposed IAM-TSP, with the exception of the space required by the possible Solutions matrix, which contains k! rows of routes, each of which hosts n cities, resulting in a total space complexity of $O(n^2 + n.k!)$, which can be asymptotically approximated to O(n.k!). This is a problem for machines that have limited memory resources, particularly when k is large, and therefore, k needs to be decided based on the available memory resources.

IV. Experimental Setup and Results

To verify the quality and effectiveness of the proposed methods for solving TSP, IAM-TSP and IAM-TSP+ were applied to nine TSP instances, each with a known optimal solution. Those TSPs are from the TSPLIB [50], which has vertices between 40 and 500, namely: a280, att48, berlin52, KroA100, ch150, ch130, pr76, lin105, and pcb442. We chose these specific instances to facilitate comparison to other methods that have been repeatedly used in many studies.

We compare the performance of the proposed methods to other related methods that proposed in the recent years; these include:

- NN: Nearest Neighbor algorithm
- FI: Farthest Insertion algorithm
- CH: Concave hull with Heuristic.
- CNH: Concave hull No Heuristic
- NI: Nearest Insertion algorithm.
- CI: Cheapest Insertion algorithm
- AI: Arbitrary Insertion algorithm
- RNN: Repetitive Nearest Neighbor algorithm
- 2-Opt: 2-Opt algorithm

It deserves to be noted that the aforementioned methods were not developed for this work; rather, we directly reference their results on each standard TSP as stated by [33], where all parameters used for each method can be obtained. It is also worth noting that the majority of these methods' results were obtained by repeating each method a number of times and reporting the average performance. However, we do not need to do the same for the proposed methods because each produces the same result on a specific TSP regardless of how many times the method is run.

We performed simulation experiments using C# of Microsoft visual studio 2022. The hardware and software specifications of the system are as follows:

11th Gen Intel(R) Core (TM) i7-1165G7 @ 2.80GHz, 8.00 GB RAM, and Windows 11 Pro, 64-bit operating system.

The results of the proposed methods to the other compared methods are shown in Table I, in which the optimal tour length (cost) is recorded as reported in the TSPLIB standard library. As it can be seen from Table I, The proposed IAM-TSP+ outperforms the IAM-TSP on all TSPs, which is to be expected given that the latter is an input to the former and the former employs nine local permutations along the resultant route, giving it more chances to identify better solutions.

TABLE I. PERFORMANCE COMPARISON OF THE PROPOSED METHODS TO NINE RELATED METHODS ON DIFFERENT TSP INSTANCES

Instance	Optimal	IAM-TSP	IAM-TSP+	CH	CNH	FI	NI	CI	AI	NN	RNN	2OPT
a280	2579	3051.0804	2978.079155	3335.9	3448	2953	3072.7	3110.7	2995.7	3369.7	3037.7	3018.1
att48	33523	35595.13404	34877.093	35618	35811.2	35267	37893.3	36391.9	35723.1	42112.7	39237	37458.4
berlin52	7542	8497.33404	8031.31761	9013.7	9013.7	8175	9097.7	9007.3	8346.4	9265.4	8182.2	8368.6
ch150	6528	7367.615815	7167.696739	7176.5	7309.6	7148	8066.9	7988.7	7228.8	7734.2	7078.4	7379.4
ch130	6110	6664.37646	6584.106521	7038	7038.8	6855	7381.6	7164.9	6625.8	7747.4	7198.7	6755.5
kroA100	21282	23375.94197	22212.77837	22899.1	23310.1	22874	25957	25073.5	23270.7	27084.2	24699	24401.9
pr76	108159	115547.4825	113155.4313	115790	118877	117173	130029	126837	116098	145227	130921	118838.3
lin105	14379	16285.06763	15948.15549	15596.1	15730.7	15331	18287.9	17327.6	15802.6	18646	16939	16322.2
pcb442	50778	57860.0934	57458.28	66961.6	72425	57537.9	60667.9	59493.1	58001.5	64819.2	59975	57354.4

Furthermore, the proposed IAM-TSP+ outperforms not only IAM-TSP, but also, all methods compared on 5 TSPs, namely att48, berlin52, ch130, KroA100, pr76, followed by the FI method, which also performs well.

Aside from the FI, the proposed IAM-TSP outperforms many other methods without the need for additional improvement and achieves solutions closer to optimal in two instances: att48 and pr76. It is interesting to note that both of the proposed methods achieve performance that is very close to the optimal solution in some instances. In order to illustrate the different performances of both of the best performers, IAM-TSP+ and FI, Fig. 3 compares the resultant routes' costs of both methods on several TSPs while Fig. 4, 5, and 6 provide comparisons of the proposed methods to the other nine related methods on small-scale TSP instances. As it can be seen from these figures, the proposed methods achieve better or comparable performance when compared to the other approximation methods that attempt to find the best possible TSP solution.

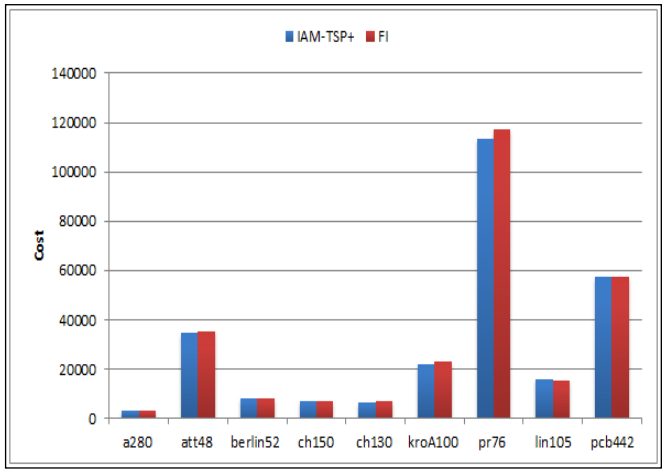


Fig. 3. The routes' cost results of the proposed IAM-TSP+ compared to that of the FI method.

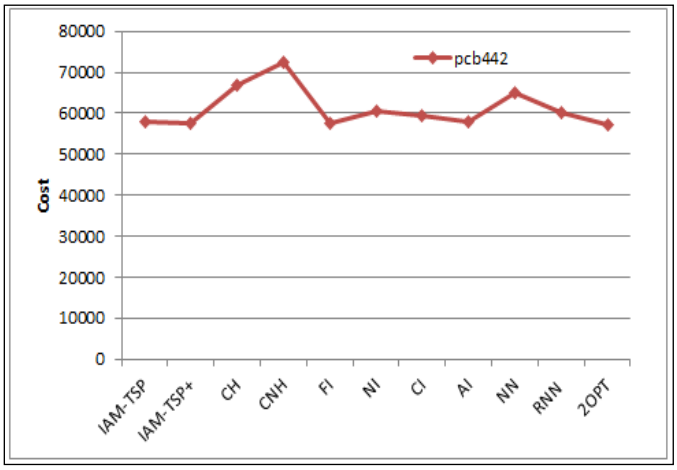


Fig. 4. The routes' cost results of the proposed methods compared to that of the other methods on Att48 TSP.

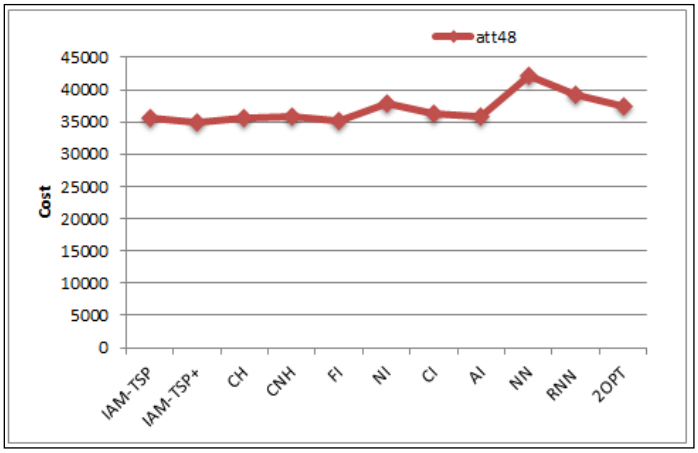


Fig. 5. The routes' cost results of the proposed methods compared to that of the other methods on pcb442 TSP.

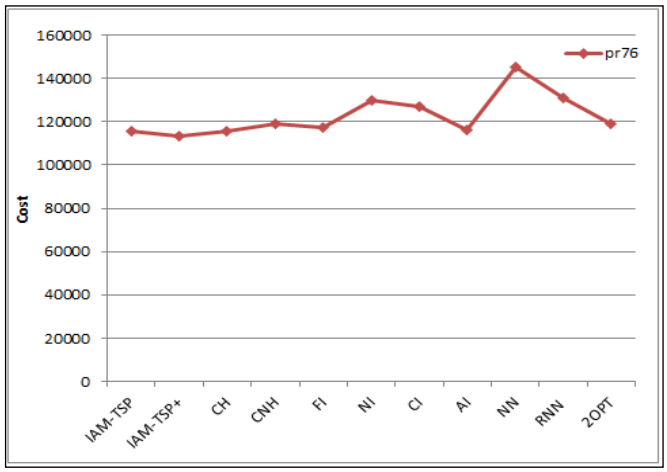


Fig. 6. The routes' cost results of the proposed methods compared to that of the other methods on pr76 TSP.

The approximation ratio, which is the ratio of the method's output to the optimal tour cost, quantifies how much the approximate solution differs from the optimal solution. The approximation ratios of all methods compared are presented in Table II. As it can be seen in the table, the proposed IAM-TSP+ delivers the best overall approximation, with an average of 1.09 overall TSPs, compared to the next competitor (FI), which has an average approximation of 1.10.

The proposed AIM-TSP also performs well, with an average approximation of 1.11, surpassing seven comparable methods. In addition to the approximation ratio, the error rate represents the percentage difference between the solution's fitness value and the known optimal solution, and it is determined as follows [51]:

$$\text{Error Rate} = \frac{solution - optimal\ solution}{optimal\ solution} * 100\% \quad (1)$$

The comparison results for the error rates are presented in Table III. As it can be seen in Table III, the proposed IAM-TSP+ delivers the minimum overall error rate, with an average of 8.51% overall TSPs, compared to the next competitor (FI), which has an average error rate of 9.50%. The proposed AIM-TSP also performs well, with an average error rate of 11.44%, surpassing seven comparable methods. Considering the costs, approximation ratios, and error rates of the resulting Routes, the proposed methods, specifically the IAM-TSP+, outperform all related methods aimed at solving the TSP in general.

TABLE II. COMPARISONS OF THE APPROXIMATION RATIOS OF THE PROPOSED METHODS TO THE OTHER RELATED METHODS

Instance	Optimal	IAM-TSP	IAM-TSP+	CH	CNH	FI	NI	CI	AI	NN	RNN	2OPT
a280	2579	1.183047848	1.15474182	1.2934858	1.33695	1.14502	1.19143	1.20617	1.16157	1.30659	1.1779	1.17026
att48	33523	1.061812309	1.040392954	1.0624944	1.06826	1.05202	1.13037	1.08558	1.06563	1.25623	1.1704	1.117394
berlin52	7542	1.126668528	1.064879025	1.1951339	1.19513	1.08393	1.20627	1.19429	1.10666	1.22851	1.0849	1.1096
ch150	6528	1.128617619	1.09799276	1.0993413	1.11973	1.09498	1.23574	1.22376	1.10735	1.18477	1.0843	1.130423
ch130	6110	1.090732645	1.077595175	1.1518822	1.15201	1.12193	1.20812	1.17265	1.08442	1.26799	1.1782	1.105646
kroA100	21282	1.098390281	1.043735475	1.0759844	1.0953	1.0748	1.21967	1.17816	1.09345	1.27263	1.1605	1.146598
pr76	108159	1.068311306	1.046195243	1.0705535	1.09909	1.08334	1.2022	1.17269	1.0734	1.34271	1.2104	1.098737
lin105	14379	1.132559123	1.109128277	1.0846443	1.09401	1.06621	1.27185	1.20506	1.09901	1.29675	1.1781	1.135142
pcb442	50778	1.139471689	1.131558549	1.3187128	1.42631	1.13313	1.19477	1.17163	1.14226	1.27652	1.1811	1.129513

TABLE III. ERROR RATE COMPARISONS

Instance	Optimal	IAM-TSP	IAM-TSP+	CH	CNH	FI	NI	CI	AI	NN	RNN	2OPT
a280	2579	18.3047848	15.47418203	29.348585	33.6952	14.5017	19.1431	20.6165	16.1574	30.6592	17.786	17.02598
att48	33523	6.181230913	4.039295424	6.2494407	6.82576	5.2024	13.0367	8.558	6.56296	25.6233	17.045	11.7394
berlin52	7542	12.66685282	6.487902542	19.513392	19.5134	8.393	20.6272	19.4285	10.6656	22.8507	8.4885	10.95996
ch150	6528	12.86176188	9.799276026	9.9341299	11.973	9.49755	23.5738	22.3759	10.7353	18.4773	8.4314	13.04228
ch130	6110	9.073264485	7.759517522	15.188216	15.2013	12.1931	20.8118	17.2651	8.4419	26.7987	17.818	10.56465
kroA100	21282	9.839028145	4.37354747	7.59844	9.52965	7.4805	21.9669	17.8155	9.34452	27.2634	16.053	14.65981
pr76	108159	6.831130592	4.619524286	7.0553537	9.90921	8.33403	20.2198	17.2686	7.3404	34.2715	21.045	9.873704
lin105	14379	13.2559123	10.9128277	8.4644273	9.40051	6.62077	27.1848	20.5063	9.90055	29.6752	17.807	13.51415
pcb442	50778	13.94716885	13.1558549	31.871283	42.6307	13.3127	19.4767	17.1631	14.2256	27.6521	18.111	12.95128

V. CONCLUSION

In this paper, we presented two geometric-based approximation greedy algorithms for solving the classical TSP, one of which we call IAM-TSP which is proposed based on locating four extreme nodes/cities and then iterating to determine the best potential positions of each node/city. The other method is known as IAM-TSP+, and it is simply an improved version of the first one, with employing local constant permutations to improve the first method's result.

The experimental results of the proposed methods on nine TSP instances show that, when compared to nine recent related methods, the proposed methods (particularly the IAM-TSP+) provide promising solutions for the classical TSP. This is seen in the resulting routes' cost effectiveness, approximation ratios, and error rates. The enhanced performance (of the IAM-TSP+) is due to the selection of the best local solution, which was accomplished by exploring all k-permutations and considering the best local solution after obtaining the initial solution from the pure IAM-TSP, where k = 5 in all experiments.

The proposed method's limitations include the time and space complexity, which is rather high for the proposed IAM-TSP+, making the evaluation of the proposed methods on large TSP instances particularly difficult on restricted resource machines. As a result, parallel or distributed computation may facilitate the evaluation of proposed methods. In addition to employing the output route as an initial seed for the genetic algorithm [52, 53]. Our future research will focus on such issues.

REFERENCES

[1] M. D. A. C. Hasibuan, et al., Pencarian rute terbaik pada travelling salesman problem (tsp) menggunakan algoritma genetika pada dinas kebersihandan pertamanan kota pekanbaru, SATIN-Sains dan Teknologi Informasi1 (1) (2015) 35–46.

[2] S. Arora, The approximability of np-hard problems, in: Proceedings of the thirtieth annual ACM symposium on Theory of computing, 1998, pp.337–348.

[3] E. Duman, I. Or, Precedence constrained tsp arising in printed circuit board assembly, International Journal of Production Research 42 (1) (2004) 67–78.

[4] F. Su, L. Kong, H. Wang, Z. Wen, Modeling and application for rolling scheduling problem based on tsp, Applied Mathematics and Computation 407 (2021) 126333.

[5] S. Hannenhalli, E. Hubbell, R. Lipshutz, P. A. Pevzner, Combinatorial383 algorithms for design of dna arrays, Chip Technology (2002) 1–19.

[6] S.-Y. Shin, I.-H. Lee, D. Kim, B.-T. Zhang, Multiobjective evolutionary optimization of dna sequences for reliable dna computing, IEEE transactions on evolutionary computation 9 (2) (2005) 143–158.

[7] J. Ahn, E. Choi, D. Lee, Application of routing problems to space exploration missions, in: AIAA SCITECH 2023 Forum, 2023, p. 1966.

[8] D. Ying, Competition decision for bottleneck traveling salesman problem based on big data mining algorithm with multi-segment support, in: 2018 3rd International Conference on Smart City and Systems Engineering (IC-SCSE), IEEE, 2018, pp. 725–729.

[9] B. Jose, T. R. Ramanan, S. M. Kumar, Big data provenance and analytics in telecom contact centers, in: TENCON 2017-2017 IEEE Region 10 Conference, IEEE, 2017, pp. 1573–1578.

[10] A. B. Hassanat, Furthest-pair-based decision trees: Experimental results on big data classification, Information 9 (11) (2018) 284.

[11] S. Mnasri, N. Nasri, T. Val, An overview of the deployment paradigms in the wireless sensor networks, Performance Evaluation and Modeling in Wireless Networks (PEMWN 2014)

[12] A. Abadleh, E. Al-Hawari, E. Alkafaween, H. Al-Sawalqah, Step detection algorithm for accurate distance estimation using dynamic step length, in:2017 18th IEEE International Conference on Mobile Data Management (MDM), IEEE, 2017, pp. 324–327.

[13] S. Tlili, S. Mnasri, T. Val, A multi-objective gray wolf algorithm for routing in iot collection networks with real experiments, in: 2021 National Com-407 puting Colleges Conference (NCCC), IEEE, 2021, pp. 1–5.

[14] A. Mars, A. Abadleh, W. Adi, Operator and manufacturer independent d2d private link for future 5g networks, in: IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), IEEE, 2019, pp. 1–6.

[15] A. Aljaafreh, K. Alawasa, S. Alja'afreh, A. Abadleh, Fuzzy inference system for speed bumps detection using smart phone accelerometer sensor, Journal of Telecommunication, Electronic and Computer Engineering 9.

[16] A. Abadleh, B. M. Al-Mahadeen, R. M. AlNaimat, O. Lasassmeh, Noise segmentation for step detection and distance estimation using smartphone sensor data, Wireless Networks 27. doi:10.1007/s11276-021-02588-0.

[17] N. Ghatasheh, H. Faris, R. Abukhurma, P. A. Castillo, N. Al-Madi, A. M. Mora, A. M. Al-Zoubi, A. Hassanat, Cost-sensitive ensemble methods for bankruptcy prediction in a highly imbalanced data distribution: A real case from the spanish market, Progress in Artificial Intelligence 9 (2020) 361–375.

[18] G. A. Altarawneh, A. B. Hassanat, A. S. Tarawneh, A. Abadleh, M. Alrashidi, M. Alghamdi, Stock price forecasting for jordan insurance companies amid the covid-19 pandemic utilizing off-the-shelf technical analysis methods, Economies 10 (2) (2022) 43.

[19] A. B. Hassanat, V. S. Prasath, M. Al-kasassbeh, A. S. Tarawneh, A. J.Al-shamailh, Magnetic energy-based feature extraction for low-quality fin-429 gerprint images, Signal, Image and Video Processing 12 (2018) 1471–1478.

[20] A. S. Tarawneh, C. Celik, A. B. Hassanat, D. Chetverikov, Detailed investigation of deep features with sparse representation and dimensionality reduction in cbir: A comparative study, Intelligent Data Analysis 24 (1) (2020) 47–68.

[21] E. Hamadaqa, A. Abadleh, A. Mars, W. Adi, Highly secured implantable medical devices, in: 2018 International Conference on Innovations in Information Technology (IIT), IEEE, 2018, pp. 7–12.

[22] A. S. Tarawneh, A. B. Hassanat, D. Chetverikov, I. Lendak, C. Verma, Invoice classification using deep features and machine learning techniques, in:2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), IEEE, 2019, pp. 855–859.

[23] A. B. Hassanat, On identifying terrorists using their victory signs, Data Science Journal 17.

[24] W. J. Cook, In pursuit of the traveling salesman, in: In Pursuit of the Traveling Salesman, Princeton University Press, 2011.

[25] A. B. Hassanat, E. Alkafaween, On enhancing genetic algorithms using new446 crossovers, International Journal of Computer Applications in Technology 55 (3) (2017) 202–212.

[26] A. Hassanat, K. Almohammadi, E. Alkafaween, E. Abunawas, A. Hammouri, V. S. Prasath, Choosing mutation and crossover ratios for genetic algorithms—a review with a new dynamic approach, Information 10 (12) (2019) 390.

[27] E. Alkafaween, A. B. Hassanat, Improving tsp solutions using ga with a new hybrid mutation based on knowledge and randomness, Communications Scientific letters of the University of Zilina 22 (3) (2020) 128–139.

[28] P. Hart, The condensed nearest neighbor rule (corresp.), IEEE transactions on information theory 14 (3) (1968) 515–516. [29] G. A. Croes, A method for solving traveling-salesman problems, Operations research 6 (6) (1958) 791–812.

[29] G. A. Croes, A method for solving traveling-salesman problems, Operations research 6 (6) (1958) 791–812.

[30] D. J. Rosenkrantz, R. E. Stearns, P. M. Lewis, Approximate algorithms for the traveling salesperson problem, in: 15th Annual Symposium on Switching and Automata Theory (swat 1974), IEEE, 1974, pp. 33–42.

[31] J. Brest, J. Zerovnik, An approximation algorithm for the asymmetric traveling salesman problem, Ricerca operativa 28 (1999) 59–67.

[32] T. Wainwright, Solving problems involving hamilton circuits, Journal of Mathematics and Science: Collaborative Explorations 1 (1) (1997) 83–90.

[33] K. Ihsan Kilic, L. Mostarda, Novel concave hull-based heuristic algorithm for tsp, in: Operations Research Forum, Vol. 3, Springer, 2022, p. 25.

[34] K. Yang, X. You, S. Liu, H. Pan, A novel ant colony optimization based on game for traveling salesman problem, Applied Intelligence 50 (2020) 4529–4542.

[35] M. Mosayebi, M. Sodhi, T. A. Wettergren, The traveling salesman problem with job-times (tspj), Computers & Operations Research 129 (2021) 105226.

[36] M. Mosayebi, The variants of traveling salesman problem with job-times (tspj).

[37] P. He, J.-K. Hao, Hybrid search with neighborhood reduction for the multiple traveling salesman problem, Computers & Operations Research 142 (2022) 105726.

[38] D. A. Aliyev, C. L. Zirbel, Seriation using tree-penalized path length, European Journal of Operational Research 305 (2) (2023) 617–629.

[39] A. Maya-L´opez, A. Mart´ınez-Ballest´e, F. Casino, A compression strategy for an efficient tsp-based microaggregation, Expert Systems with Applications 213 (2023) 118980.

[40] D. Rodr´ıguez, J. M. Cruz-Duarte, J. C. Ortiz-Bayliss, I. Amaya, Asequence-based hyper-heuristic for traveling thieves, Applied Sciences 13 (1) (2023) 56.

[41] V. Cacchiani, C. Contreras-Bolton, L. M. Escobar-Falc´on, P. Toth, Amatheuristic algorithm for the pollution and energy minimization traveling salesman problems, International Transactions in Operational Research 30 (2) (2023) 655–687.

[42] H. Liu, A. Lee, W. Lee, P. Guo, Daaco: adaptive dynamic quantity of ant aco algorithm to solve the traveling salesman problem, Complex & Intelligent Systems (2023) 1–14.

[43] S. S. Harahap, P. Sihombing, M. Zarlis, Combination of ant colony tabu search algorithm with firefly tabu search algorithm (acts-fats) in solving the traveling salesman problem (tsp), Sinkron: jurnal dan penelitian Teknik informatika 8 (1) (2023) 212–221.

[44] S. Dhouib, A new column-row method for traveling salesman problem: the dhouib-matrix-tsp1, International Journal of Recent Engineering Science 8 (1) (2021) 6–10.

[45] E. Osaba, X.-S. Yang, J. Del Ser, Traveling salesman problem: a perspective review of recent research and new results with bio-inspired metaheuristics, Nature-Inspired Computation and Swarm Intelligence (2020) 135–164.

[46] Cengiz, C. Yilmaz, H. T. Kahraman, C¸. Sui¸cmez, Effects of variable uav speed on optimization of travelling salesman problem with drone (tsp-d), in: Smart Applications with Advanced Machine Learning and Human-Centred Problem Design, Springer, 2023, pp. 295–305.

[47] N. Garg, M. K. Kakkar, G. Gupta, J. Singla, A performance evaluation of genetic algorithm and simulated annealing for the solution of tsp with profit using python, in: Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2022, Volume 3, Springer, 2022, pp. 13–26.

[48] M. Das, A. Roy, S. Maity, S. Kar, A quantum-inspired ant colony optimization for solving a sustainable four-dimensional traveling salesman problem under type-2 fuzzy variable, Advanced Engineering Informatics 55 (2023) 101816.

[49] ´E. D. Taillard, Design of Heuristic Algorithms for Hard Optimization: With Python Codes for the Travelling Salesman Problem, Springer Nature, 2023.

[50] G. Reinelt, Tsplib—a traveling salesman problem library, ORSA journal on computing 3 (4) (1991) 376–384.

[51] S. S. Ray, S. Bandyopadhyay, S. K. Pal, Genetic operators for combinatorial optimization in tsp and microarray gene ordering, Applied intelligence 26 (2007) 183–195.

[52] E. Alkafaween, A. B. Hassanat, S. Tarawneh, Improving initial population for genetic algorithm using the multi linear regression based technique (mlrbt), Communications-Scientific letters of the University of Zilina 23 (1) (2021) E1–E10.

[53] A. B. Hassanat, V. S. Prasath, M. A. Abbadi, S. A. Abu-Qdari, H. Faris, An improved genetic algorithm with a new initialization mechanism based on regression techniques, Information 9 (7) (2018) 167.

An Overview of Different Deep Learning Techniques Used in Road Accident Detection

Vinu Sherimon[1], Sherimon P.C.[2], Alaa Ismaeel[3], Alex Babu[4], Sajina Rose Wilson[5], Sarin Abraham[6], Johnsymol Joy[7]
College of Computing and Information Sciences,
University of Technology and Applied Sciences, Muscat, Sultanate of Oman[1]
Faculty of Computer Studies, Arab Open University, Muscat, Sultanate of Oman[2, 3]
PG Department of Computer Applications and Artificial Intelligence,
Saintgits College of Applied Sciences, Kerala, India[4, 5, 6, 7]

***Abstract*—Every year, numerous lives are tragically lost because of traffic accidents. While many factors may lead to these accidents, one of the most serious issues is the emergency services' delayed response. Often, valuable time is lost due to a lack of information or difficulty determining the location and severity of an accident. To solve this issue, extensive research has been conducted on the creation of effective traffic accident detection and information communication systems. These systems use new technology, such as deep learning algorithms, to spot accidents quickly and correctly and communicate important information to emergency workers. This study provides an overview of current research in this field and identifies similarities among various systems. Based on the review findings, it was found that researchers utilised various techniques, including MLP (Multilayer Perceptron), CNN (Convolutional Neural Network), and models such as DenseNet, Inception V3, LSTM (Long short-term memory), YOLO (You Only Look Once), and RNN (Recurrent Neural Network), among others. Among these models, the MLP model demonstrated high accuracy. However, the Inception V3 model outperformed the others in terms of prediction time, making it particularly well-suited for real-time deployment at the edge and providing end-to-end functionality. The insights gained from this review will help enhance systems for detecting traffic accidents, which will lead to safer roads and fewer casualties. Future research must address several challenges, despite the promising results showcased by the proposed systems. These challenges include low visibility during nighttime conditions, occlusions that hinder accurate detection, variations in traffic patterns, and the absence of comprehensive annotated datasets.**

Keywords*—Deep learning; road traffic; road accident detection; MLP; CNN; LSTM; DenseNet; RNN; inception V3

I. INTRODUCTION

Road crashes are one of the most prominent reasons for death, disability, and hospitalization of people worldwide. 1.35 million people worldwide lose their lives due to road accidents every year [1]. Almost 3,700 people worldwide lose their lives in collisions with automobiles, buses, motorbikes, bicycles, lorries, or pedestrians every day [1]. Cyclists, motorcyclists, and pedestrians account for more than half of the fatalities. According to estimates, crashes rank as the eighth most common cause of mortality worldwide across all age categories and the most common cause of death for children and young adults (ages 5 to 29) [1]. The traffic authority has seen that many vehicles are travelling at high speeds and without fear in crowded areas. The results of traffic accidents are fairly evident. To address these issues, though, a reliable traffic monitoring and management system is needed [2].

Unfortunately, the age group most affected by road accidents is people aged between 18 and 45 years, which accounts for approximately 70% of accidental deaths [3]. Most of these fatalities occur due to a lack of medical care given at the appropriate time. The expenditures associated with medical care, property damage, legal actions, and reduced worker productivity are significant. In addition to the human cost, road accidents place significant economic pressure on countries, taxing healthcare systems and diverting revenue away from other vital sectors, which hinders economic expansion. Due to its vital impact on public safety, economic stability, and general societal well-being, road accident detection is of the utmost importance. To prevent these serious consequences, effective road accident detection is urgently required.

Current accident reporting systems that rely on reports from witnesses or have slow response times frequently cause delays in facilitating traffic control and providing emergency help. The lengthening of travel times for other road users as well as secondary accidents, congestion, and a worsening of injury severity can all result from this delay.

Deep learning [4] is a subset of machine learning that employs multilayer neural network models that are designed to function similarly to the human brain. Unlike other machine learning approaches, deep learning can automatically learn from data such as photos, videos, or text without the aid of any subject expertise. Deep learning employs a multi-layered neural network to extract features from the data and gets better and better at recognizing and classifying data on its own rather than depending on labels included in the unstructured data. In areas like machine vision, computer vision, speech recognition, audio recognition, natural language processing, social network filtering, pharmaceutical engineering, bioinformatics, medical image analysis, and gaming programs, deep learning architectures like deep neural networks, deep belief networks, recurrent neural networks, and convolutional neural networks have been used with related and, in some cases, improved outcomes.

A subfield of artificial intelligence (AI) called computer vision [5] allows computers and other systems to retrieve useful data from photos, videos, and other visual inputs, to analyze objects in photos and videos just like people do, and to take decisions or offer suggestions based on that data. Computer vision tasks include methods for acquiring, processing, and analyzing digital images to generate numerical or symbolic data. Using digital images and deep learning models, computer vision allows computers to accurately identify and classify objects and respond to them. A convolutional neural network understands single images, whereas recurrent neural networks work with video inputs, allowing computers to learn how a series of images is related to one another. Computer vision software can process images in real-time to detect road edges, read traffic signals, and identify other cars, objects, and pedestrians from the camera output.

Deep learning is revolutionizing road accident detection systems by using neural networks to improve responsiveness, accuracy, and adaptability. Deep learning allows these systems to use convolutional neural networks (CNNs), recurrent neural networks (RNNs), and other sophisticated designs to analyze complex visual and sensor data collected from various sources. Deep learning algorithms may develop robust and adaptable models that can cope with the intricacies of the real world by learning from a variety of datasets that cover a range of weather conditions, lighting conditions, and road types. Additionally, its capacity for real-time processing makes it easier to discover accidents quickly, enabling emergency services and traffic control authorities to act right away.

Several research gaps and unresolved issues emerged as the literature on deep learning-based road accident detection systems continued to develop. Even though deep learning techniques have been used in this field in important ways, a thorough synthesis of the literature in the field reveals some gaps that call for more research. This review article offers a current assessment of the state of the field, considering substantial advancements, new patterns, and cutting-edge technology in road accident detection systems. A current summary of the most recent developments in the field is covered. This research paper advances knowledge in this field by methodically evaluating and synthesizing the body of literature on deep learning-based road accident detection systems. This analysis can help researchers and practitioners and can act as a starting point for new research and technological breakthroughs. Policymakers and emergency service providers can prioritize investments in technology, infrastructure, and staff by comprehending the benefits and drawbacks of various deep learning-based solutions.

The rest of the paper is organized as follows: Section II describes the methodology employed to conduct the literature review and discusses the databases, search terms, and inclusion/exclusion criteria used to select relevant papers. Section III then includes a detailed analysis of each selected piece of literature, discussing the main findings, and contributions of each study. Also, a concise summary of the key findings and outcomes of the reviewed papers is included in this section. Section IV provides conclusion, summarizes the main findings and contributions of the literature review, and Section V provides recommendations for future research directions.

II. Materials and Methods

A. Literature Sources and Search Strategies

The information for this review was obtained by looking at publications that have been published between January 2019 and June 2023 using the well-known search engines Google Scholar, Scopus, and Web of Science. A combination of keywords like artificial intelligence, deep learning, machine learning, artificial neural networks, convolutional neural networks, and accident detection was used for data searching.

B. Identification and Selection of Relevant Studies

The articles from the preliminary search that had abstracts with full texts were retrieved. Along with computer searching, non-electronic sources were also used, such as manual searches for relevant journals and publications. The articles were chosen in accordance with the title and after reviewing the abstracts pertaining to our study subject. 202 publications that fit the review's objectives were primarily found through the initial search. Due to data duplication, nine articles were ignored, and 57 were removed due to topic irrelevance. Thus, 136 articles were included for the second stage of data selection. These articles were then subjected to the subsequent inclusion and exclusion criteria.

C. Eligibility Criteria of the Studies

According to the inclusion criteria, the articles must be specifically focused on deep learning-based road accident detection systems; they must also explicitly address the deep learning technology utilized in the study model; they must explicitly discuss the data sets that were used for the model's evaluation, validation, or training; and they must explicitly mention a quantitatively measurable predictive outcome. Articles written in languages other than English and those with topics other than deep learning technology were excluded. 118 articles were excluded after applying the criteria.

D. Data Extraction and Management

The number of articles was further lowered to 18 after applying these eligibility requirements.

Fig. 1 displays the flowchart of the screening and selection of literature.

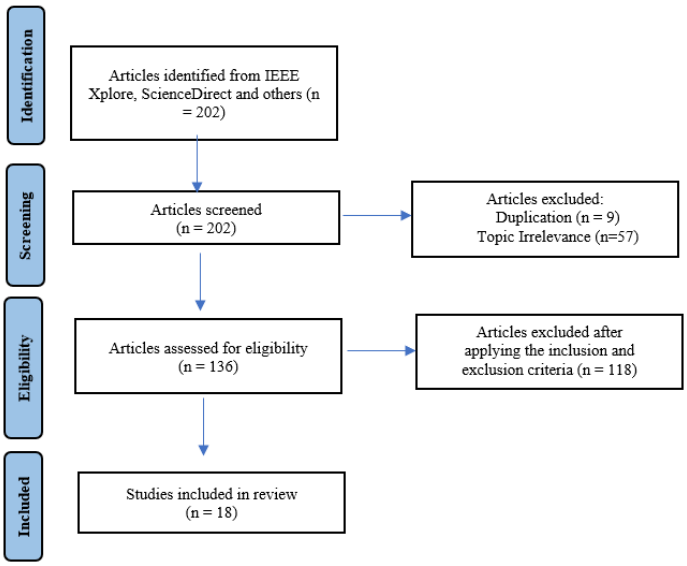


Fig. 1. Flowchart for screening and selection of literature.

III. Discussion

This study in [6] investigates relevant data variables, such as accident severity, the number of casualties, and the number of cars, to uncover trends in road automobile accident data and to suggest a predictive model. Consequently, a pre-processing model is created to transform raw data using interquartile outlier removal, generalization of data attributes, and the elimination of missing and nonsensical features. The performance of road accident prediction is studied using and evaluating four classification methods: naïve Bayes, random forest, decision trees, and multinomial logistic regression. Except for naïve Bayes, the results address acceptable levels of accuracy for car accident prediction. The results are examined using a data-driven methodology to identify the critical variables causing traffic accidents and to suggest ways to prevent them.

In this research [7], a deep learning accident prediction model is proposed by combining extended features such as sentiment analysis, emotions, weather, geocoded locations, and time information with data derived from tweet messages. According to the data, accident detection accuracy increased by 8%, bringing the test's accuracy up to 94%. The suggested methodology fared better than the state-of-the-art methods now in use, increasing accuracy by 2% and 3%, respectively, to reach 97.5% and 90%. The high-performance computing constraints brought forth by detector-based accident detection, which required massive data calculation, were likewise addressed by our method. The outcomes have increased confidence that utilizing cutting-edge features improves traffic accident identification and prediction.

The work done by [8] provides an overview of the technologies used in road accident detection and information transfer system. They also emphasized the importance of automatic road accident detection and information communication systems in all vehicles, whose role is to send a notification message to the nearest hospital and police station upon detecting an accident. As the performance of such systems can be greatly improved with the inclusion of technologies like social network data analysis, sensor data, machine learning, and deep learning, the authors have used artificial neural networks to build a new model. The entire system is structured as three modules, where the first module comprises the collection of image data. The second module is about pre-processing the data and the last module comprises accident detection and notification, which extracts visual and temporal features using the PYTORCH architecture and neural networks called a multi-layer perceptron (MLP). The proposed system proved to be very useful in detecting incidents with an F1 score of 0.98 and accuracy of 98 percent.

The high number of accident-related fatalities in India is a topic of this study [5], which highlights the importance of quick guidance. The paper suggests a method for identifying accidents using real-time CCTV data from roadways. It uses a Hierarchical Recurrent Neural Network (H-RNN) model that has been trained to distinguish between accident- and non-accident-related video frames. H-RNN is a potential option for quick accident detection because of its accuracy in image classification—over 95%—and applicability for video analysis—it can capture temporal dependencies. With the help of this study, accident victims will receive much better response times and outcomes.

The research described in [9] highlighted the need of minimizing the time required by an accident-detection system in sending notifications to the emergency services. To meet this objective, the authors proposed a software-oriented accident recognition system using convolutional neural networks (CNN) that could send real-time alerts to the nearest hospitals and police stations. The new system used only the available resources with the modifications brought through the deployment of AI techniques on edge devices attached to roadside CCTV cameras for classifying images. As the input to the convolutional neural network (CNN), live footage from CCTV cameras were used, from which the probability of the occurrence of an accident was generated. Accident Recognition and Alerting System (ARAS) is divided into three main components, based on their functionality. The components are the Edge Accident Recognition Module (EARM), Alert Database (AD) and Alert Web Interface (AWI). The EARM is responsible for recognizing accidents and sending alert messages. The alerts are received by the AWI, which is a website provided to the emergency services. The AD acted as a common point between the EARM and the AWI. Three kinds of CNN architectures are used which are Inception v3, DenseNet, and ResNet-50 respectively thus resulting in three new models. The performance of these models was compared, and it was found that Inception v3 detected the accident throughout as opposed to DenseNet and ResNet-50. Also, the Inception v3 model displayed the best prediction time, proving it to be an effective choice for real-time, end-to-end, edge deployment. These models

additionally help in avoiding the transmission of duplicate alerts.

A new support system for self-driving cars is proposed in [10] that could detect vehicular accidents through a dashboard camera and spontaneously report it to local authorities. The Mask R-CNN framework is used for object detection along with an object tracking algorithm to make the new system more efficient. The object detection component serves to identify the type of vehicle as either a car, bus, or truck. The output of this step includes the class IDs, generated masks, bounding boxes, and the detection scores of vehicles for a video frame. Whereas the part comprising a centroid tracking algorithm is used for vehicle tracking.

Additionally, trajectory intersection, velocity computation, and lane detection are also included to improve the performance of the vehicular accident detection framework. The proposed model is compared with existing models using two important performance criteria, i.e., Accident Detection Rate (ADR) and False Alarm Rate (FAR). As the number of parameters is more, the reliability of the new system in accurately recognizing accidents also got magnified with 79.05% Accident Detection Rate and 34.44% False Alarm Rate. Despite these advancements, the system is handicapped by the unavailability of large, annotated datasets that can be used to test and improve the system further.

The research study in [11] proposed Object Detection and Tracking System (ODTS) for automatically detecting and monitoring unexpected events on CCTVs in tunnels. The unexpected events are categorized as either Wrong-Way Driving (WWD), Stop, Person out of the vehicle in a tunnel, or fire. The system accepts selected frames of video at specified time intervals as input, which are then converted into Bounding Box (BBox) and is simultaneously classified into its corresponding type by the detection module which uses a faster RCNN learning algorithm for the same. The detected objects are then assigned a unique ID number by the object tracking module which uses the SORT framework. However, automatic object detection systems in tunnels suffer from the common problem of low illumination videos.

A deep learning model of Faster R-CNN was used for training dataset to address this issue. However, the model does have a limitation in that the ODTS employs an object tracking function with car objects only. The ODTS module applied the Car Accident Detection Algorithm (CADA) for classifying several events. The inclusion of these technologies resulted in the system being able to identify the incidents within 10 seconds after testing with the image that contained each incident but due to the lack of sufficient Fire objects in the training set, the likelihood of false detection in the case of Fire is high.

An automated system for detection of accidents at traffic intersections is proposed in [12]. The proposed framework is made of three main modules, which are object detection, object tracking and the accident detection module respectively. The object detection part is primarily based on the YOLOv4 approach, which makes it very efficient and accurate. The item monitoring module uses the Kalman filter and the Hungarian algorithm for affiliation whose performance is further enhanced with the use of a new cost function. It can accommodate several challenging situations, such as occlusion, overlapping objects, and shape changes that occur while tracking the objects.

Whereas the accident detection module was constructed with the aid of trajectory conflict analysis. The trajectories of the objects are analysed in phrases of velocity, angle, and distance so that extraordinary forms of trajectory conflicts consisting of vehicle-to-vehicle, vehicle-to-pedestrian, and vehicle-to-bicycle can be detected. These unique features make the model suitable for real-time traffic surveillance. The performance of the proposed framework is evaluated using the video sequences under numerous illumination conditions that were obtained from YouTube. The robustness of the system is proved through the lower fake alarm rates and an excessive detection rate.

A traffic accident prediction system that predicts traffic accidents on highways is proposed in study [13]. To accurately predict accidents, the model uses Java's high-level neural network API framework. Here are the steps the model goes through: A convolutional layer is first given traffic accident records as input and is then filtered to remove duplicate fields. After removing the unwanted records, the adjusted dataset is fed to the new layer. Weights are then assigned to the retrieved features based on random number generation, and finally network backpropagation is performed until the error reaches a threshold. The performance of the proposed convolutional neural network (CNN) is evaluated and compared with the traditional neural network-based backpropagation prediction methods. The results show that the new algorithm has much lower loss and higher prediction accuracy compared to the conventional devices. The incident prediction model proposed in this paper has only a small impact on the prediction rate, but it can be improved by choosing other features to input to the CNN or applying a more efficient model instead of the CNN.

A framework for automatic detection of road accidents in surveillance videos is proposed in [14]. A convolutional auto-encoder was incorporated in the model for extraction of deep representation, in the form of appearance and motion features as well as their correlations, from the spatiotemporal video volumes (STVVs) of raw pixel intensity and then detection for any abnormalities in surveillance videos. The accident detection ability of the model is further increased by combining the complementary appearance and motion information. Additionally, a collision is also sensed using the information about the joints of the trajectories of the vehicles over space-time dimensions. Once the joint trajectories are made, a one-class SVM with an RBF kernel is used to generate the outlier score from his intermediate representation to identify unusual/unseen/abnormal/outlier events. The performance measure of this model is evaluated using real accident videos collected from the

CCTV surveillance network of Hyderabad City in India. The result witnessed a reduction in the false alarm rate and an increase in the reliability of the detection system by using the proposed approach. But the system still faced the challenges of low visibility at night, occlusions, and large variations in the normal traffic pattern, which need to be addressed in the future.

To ensure that victims receive timely aid, the study given in [15] created a system to identify traffic incidents in real time and send an immediate alarm message to the nearest control unit. As the existing structures that use features such as IR sensors, IMU sensors, ARDUINO UNO etc., lack accuracy in detecting accidents, the authors attempted to build in more accuracy and effectiveness in their system through the usage of deep learning techniques such as convolutional neural networks. As no dataset regarding accidents occurring in India was available dataset was created to train the CNN model. The trained system is then included with the cameras, positioned at accident-susceptible areas, to seize the video of the vehicles on the road. Then for video classification, deque was used for performing the rolling prediction averaging. Once an accident is detected, officials are informed about the same by sending them an alert message using the GSM module that's included in conjunction with the camera. The proposed version proved to be more dependable and economical compared to the current structures. It could detect accidents with a high level of accuracy with almost eighty-five percent of successful image prediction. The author similarly proposed incorporating speed-detection cameras as future work, which could make the system more efficient.

Using both video and audio data generated from dashboard cameras, the research in [16] proposed a car crash detection system. Multimodal data is composed of three different types of data, namely, video data, audio features of audio data, and spectrogram images of audio data which were collected from a video sharing platform like YouTube. The data were then standardized and normalized to occupy a standard frame. Deep learning techniques like gated recurrent unit (GRU) and convolutional neural network (CNN) were then used to classify the data as either positive or negative clips. The CNN module consisted of five conv2D layers with ReLu activation function, four batch normalization layers with 0.9 momentum, three max pooling layers, and a global max pooling layer. A weighted average ensemble is used as an ensemble technique for combining the three different classifiers. Finally, the classification performance of the proposed car crash detection system was compared with the existing system, and it was found that the proposed system performed better than its base counterpart, which relied on single modal data (i.e., video or audio data only) to detect possible car-crashes. The proposed model could bring in a drastic improvement in the accuracy of the system. The authors further proposed an enhancement to their model using the inclusion of three-dimensional filters, which may further improve the efficiency more.

An automated system capable of detecting traffic accidents from video is proposed in [17]. A deep learning (DL)-based approach is used, which shows a high performance in accomplishing computer vision tasks that involves complex features relationship. The work assumes that accident events can be represented using data that have visual features occurring in a temporal way. The model hence comprises a feature extraction phase that uses the convolutional method, followed by temporary pattern identification accomplished using recurrent neural networks. For accident detection, a dense ANN approach is used.

Therefore, the model can be structured into three portions where each portion focuses on spatial feature extraction, temporal feature extraction and binary classification, respectively. The model was trained using the imageNet dataset. As the precision of the accident detection system was found to be under satisfactory, the model was modified using a transfer learning process and the weights for the model were adjusted using a new dataset. The modification included a ConvLSTM layer-based neural network being added up to the existing architecture for enhanced feature extraction and usage of a dense artificial neural network block using regularization methods for better accident detection.

Under the ConvLSTM layer-based neural network model, the feature vector computed by the adjusted InceptionV4 architecture was taken as the input. For this, two sets of data are used, where the first one comprised image for the visual extractor and the other had videos for training the temporal extractor. Both data sets contain positive and negative accidents. The modification also included the application of four techniques in video segmentation, in which the first one consisted of video segmentation without frame discrimination., the second one was used to skip the frame intruder to reduce the redundancy, the third technique helped in calculating a pixel-to-pixel comparison of two consecutive images and finally the fourth technique is substantially similar to the third except that the threshold was set to 0.98 against the 0.9 mark in the third. Through the modification the accuracy improved to 0.98. But the proposed model still had a few limitations, which is, the model required large datasets with clear data to get trained well.

A feature fusion-based deep learning model for video-based accident detection is suggested in study [18], that could achieve better detection speed and accuracy with restricted computing resources. The system comprises an interest module for capturing the appearance features of the crash images. The module was blended with ResNet to enhance the speed of the traditional convolution neural network. A 1 x 1 convolutional layer is employed to reduce the size of the output feature map, which is then entered into the Conv-LSTM network to accurately extract the motion features of crashes. This Conv-LSTM network has a bonus over its traditional counterpart in terms of being lighter and preserving spatial information. Finally, a global pooling layer is used to locate a crash. This proposed system won over the current video-based crash detection

systems as they suffered from low detection accuracy and excessive computational costs. However, the proposed version has a few limitations like falsely detecting traffic scenes as crashes, showcasing few misdetections in complicated, rare, and ambiguous environment, etc. The model may also be, in addition, stepped forward to become aware of one-of-a-type types/severity tiers of crashes.

An accident detection system that may track accidents at the generation of occurrence is given in [19]. The proposed model is a fusion of CNN and LSTM layers for the classification of nonstop video captured by the camera. In the CNN-LSTM network, CNNs are used for image feature extraction and exceed LSTMs for sequence prediction. When compared with other existing models composed of high-priced sensors and pointless hardware, the proposed model showed improved performance in terms of cost, reliability, and accuracy. In the future, the model's performance may be enhanced using some supervised and unsupervised techniques. Particularly, supervised learning can be used to discover the accidents from the frames which can be flagged anomalously via way of means of unsupervised models. More state-of-the-art techniques for monitoring accident-prone areas may be used.

A system to detect accidents from video footage provided to it using a camera is proposed in [20]. The objective of this model, which is built using advanced Deep Learning Algorithms, i.e., convolutional neural networks (CNN) and the LSTM network is to detect accidents within seconds of its occurrence. In the CNN-LSTM network, CNN is used for modeling spatial data like images and feature extraction which is passed on to the LSTM for sequence prediction. The pictures captured through the Pi-cameras are broken down into frames and fed to the automated system and while an accident is detected, an alert message is dispatched using the GSM module. Along with the message, the frame at which an accident is detected and the percentage of chances of accident are also conveyed. Upon assessing the model's performance, it yields an average accuracy of 92.83%.

An automatic car accident detection method based on Cooperative Vehicle Infrastructure Systems (CVIS) and machine vision such that roadside intelligent devices recognize and locate crashes efficiently is given in [21]. For this purpose, a new image dataset (CAD-CVIS) is created based on video sharing websites consisting of various kinds of accident types, weather conditions, and accident locations. The data includes video features of the car's motion parameters, and the concept of accident detection is since the motion parameters will change dramatically when an accident occurs. The CAD-CVIS dataset used in the study was divided into three parts. A training set (80%) used to train the parameter weights of the network, a validation set (5%) used to fir and test hyperparameters such as learning rate and dropout rate and a test set (15%) used to evaluate the performance of various algorithms for detecting car accidents. Here, each part of the dataset contains all types of accident.

A deep neural network algorithm, YOLO-CA, was used to detect accidents and their location. The performance of the model was enhanced by including Multi-Scale Feature Fusion (MSFF) and a dynamically weighted loss function to help detect small objects better. Finally, the performance of the proposed model was evaluated and compared with other object detection models, which showed significant improvements in terms of response time by rescue agencies and rescue efficiency.

A prototype that uses a machine learning approach to reduce road accidents caused by reasons such as drowsiness, fatigue, and inattention is developed in [22]. The CNN algorithms help avoid accidents, notify drivers when they detect drowsiness, and are very efficient at classifying datasets. This approach combines machine learning techniques with other concepts such as avoidance of drinking, direction control, speed control, and distance maintenance. The new system is superior to many existing systems as it uses more sophisticated and robust sensors than the others. Additionally, the proposed system uses IoT components cheaper than available mechanisms such as Wireless Sensor Networks (WSN), Cloud computing, and Industry 4.0, etc., so it offers a high-tech solution to avoid accidents at low cost. In conclusion, it can be stated that the proposed system offers a great opportunity to avoid accidents at an affordable price.

Table I presents an overview of the reviewed papers. A comprehensive overview of the reviewed papers, highlighting the dataset used, technique employed, observations made, and performance, is given in the table. The dataset used column provides insight into the specific data sources utilised in the studies. This information is essential for understanding the scope and applicability of the research findings. The methodology column outlines the methodologies, algorithms, or experimental approaches employed by the authors. This section sheds light on the scientific tools and methods utilised in each study, allowing readers to assess the validity of the research. The observation column summarises the key findings or outcomes of each paper. The performance column summarises the accuracy of the models.

TABLE I. SUMMARY

Author & Year	Dataset used	Methodology	Observations	Performance
(Pourroostaei Ardakani et al., 2022)	Online traffic accident dataset based on UK	Decision trees, Random Forest, Multinomial logistic regression, and Naïve Bayes algorithm	This paper presents a predictive model for road car accidents, focusing on significant data features like accident severity, casualties, and vehicles. Four classification methods are used, with acceptable accuracy levels. Among the four, random forest algorithm can predict car accidents with more accuracy.	85% accuracy
(Azhar et al., 2022)	Twitter messages	Deep learning techniques	The paper presents a deep learning accident prediction model that uses tweet messages and additional features like sentiment	94% accuracy

			analysis, emotions, weather, and geo-coded locations.	
(L et al., 2022)	The web scraping method was used to construct the image dataset from scratch.	Feed forward Neural networks (MLP)	Visual and temporal feature extractors are used to distinguish between traffic collisions. The training process for fine-tuning the weights used the picture dataset. Because there aren't many instances given, the solution only applies to car accidents—motorcycles, bicycles, and pedestrians aren't included.	92% accuracy
(Parthiban et al., 2021)	CADP dataset, which includes recordings with accidents, and the DETRAC dataset, originally designed for vehicle object location.	Hierarchical Recurrent Neural Network	Compared to traditional RNNs, the H-RNN is more appropriate for video extraction. H-RNN based picture classifiers have provided precisions of over 95% for comparatively smaller datasets and require less preparation.	92.38% accuracy
(Chitale et al., 2020)	The TrafficNet dataset from DeepQuest AI	CNN, InceptionV3, Densenet and Resnet-50	A software-oriented approach employing CNN to detect road accidents. This produces the likelihood that an accident may occur or not. In contrast to DenseNet and ResNet-50, Inception v3 identified the accident continuously. The best prediction time was also achieved by the Intermediate Representation of Inception v3, indicating that this model is a good option for real-time, end-to-end edge deployment.	93.75% accuracy
(Chand et al., 2020)	The Microsoft Common Objects in Context dataset and dash-cam footage of accidents from YouTube	Mask R-CNN	The proposed model compared Accident Detection Rate (ADR) and False Alarm Rate (FAR). The suggested approach has an accurate accident detection rate of 79.05% for cars and a false alarm rate of 34.44% for cars.	79.05% accuracy
(Lee & Shin, 2019)	CCTV footage	Faster RCNN	Proposed Object detection and tracking system (ODTS) by fusing an object tracking method with a deep learning-based object detection procedure. The model achieved average precision values of 0.8479, 0.7161, and 0.9085 for cars, people, and fires, enabling the system to detect all accidents in less than 10 seconds.	84.79% precision for detecting cars
(Ghahremannezhad et al., 2022)	Customized dataset	YOLOv4 method with a pre-trained CNN model CSPDarknet53.	The Euclidean distances between all object pairs are calculated to identify the objects. The angle of collision trajectory conflicts involves near-accident and accident occurrences. With a false alarm rate of 6.89% and a detection rate of 93.10%, the proposed structure performed well.	93.10% detection rate
(Thaduri et al., 2021)	Traffic accident dataset	CNN, backpropagation	Polling method removes unnecessary data. The proposed CNN is compared with the backpropagation prediction method. Compared to the conventional BP algorithm, the suggested algorithm has a substantially smaller loss and a higher prediction accuracy.	98% accuracy
(Singh & Mohan, 2019)	CCTV Surveillance footage	Autoencoder, one class SVM	The proposed framework uses denoising autoencoders trained on traffic videos to automatically learn feature representation from spatiotemporal volumes of raw pixel intensity, replacing traditional hand-crafted features.	77.5% detection rate
(Rajesh et al., 2020)	Customized dataset	CNN	Dropout is used to prevent overfitting. The model was created using sequential API GSM module for alert. The proposed system uses a GSM module to detect road accidents, sending alert messages to nearby control rooms. It is reliable, economical, and highly accurate.	85% accuracy
(Choi et al., 2021)	Images from dashboard cameras	GRU, CNN	Utilizing video data, a CNN-and-GRU-based classifier, a GRU-based classifier utilizing audio characteristics from audio data, and a CNN-based classifier using spectrogram images from audio data make up the suggested model.	98.60% accuracy for Case Study 1 and 89.86% for Case Study 2
(Robles-Serrano et al., 2021)	Customized dataset	ConvLSTM	The suggested approach assumes that visual characteristics that appear over time describe traffic accident incidents. Thus, the model architecture consists of an extraction phase for visual features and a transient pattern identification phase.	98% accuracy
(Lu et al., 2020)	Traffic image datasets	ConvLSTM, ResNet, Vgg	The framework proposes a residual neural network and attention modules for extracting crash-related appearance features from urban traffic videos, which are then combined with Conv-LSTM for simultaneous capture.	87.78% accuracy
(T.S. et al., 2021)	CCTV image dataset	CNN, LSTM	The proposed model combines CNN and LSTM layers for continuous video classification, incorporating ResNet-50 and LSTM layers for spatial and temporal characteristics, with adjustments made for training images.	99.9% accuracy
(Ghosh et al., 2019)	CCTV video dataset	CNN, RNN	The proposed model combines CNN and LSTM layers for continuous video classification, inspired by Inception v3, with temporal and spatial features added to the existing Convolution	92.38% accuracy

			Network, divided into convolution and recurrent parts.	
(Tian et al., 2019)	The novel image dataset CAD-CVIS.	YOLO-CA	The paper presents an automatic car accident detection method using Cooperative Vehicle Infrastructure Systems (CVIS) and machine vision, utilizing a novel image dataset and a deep neural network model YOLO-CA, which enhances performance in detecting small objects.	90.02% average precision
(Razeeth et al., 2021)	Customized image datasets	CNN	The study uses machine learning to detect drowsiness, improve drunk and speed control, and prevent collisions using IoT. It uses convolutional neural network with Keras and MQ13, MAX30105, and L298N sensors.	Drowsiness is detected successfully, and collisions are prevented

IV. Conclusion

The analysis of related work highlights several studies and their methodologies. For instance, researchers have developed models using artificial neural networks, convolutional neural networks (CNN), and object detection algorithms to improve road accident detection. These models utilize image data, pre-processing techniques, and accident detection and notification modules to achieve high accuracy and efficiency. Techniques like social network data analysis, sensor data, and trajectory conflict analysis have also been employed to enhance detection capabilities. Researchers employed a variety of techniques like MLP, CNN, and CNN models, including DenseNet, Inception V3, LSTM, YOLO, RNN, and others. Although the MLP model showed high accuracy, the Inception v3 model provided the best prediction time, proving its suitability for real-time, end-to-end edge deployment.

The proposed systems demonstrate promising results, including high accuracy rates, low false alarm rates, and efficient real-time detection. However, challenges such as low visibility at night, occlusions, variations in traffic patterns, and the lack of annotated datasets still need to be addressed in future research. In most of the papers, the delays in giving timely medical attention to the injured are quoted as the major reason for the increase in traffic fatalities. The models used both video and audio data from the dashboard camera for detection. The major problem with these types of datasets is their size. To develop effective deep learning models, we must train our models on a large dataset.

In conclusion, this review reveals significant advancements in road accident detection using machine learning and deep learning approaches. The studies highlight the importance of accurate and efficient detection systems that can improve emergency response and reduce the impact of accidents.

V. Future Work

To overcome the problems related to data scarcity, we are planning to use different strategies, like different pre-trained convolutional neural networks or Spinal net, for feature extraction. Also, further research is needed to overcome the identified challenges in the existing literature and enhance the reliability and performance of such systems.

Acknowledgment

The research leading to these results has received funding from the Research Council (TRC) of the Sultanate of Oman under the Block Funding Program BFP/RGP/ICT/22/143.

References

[1] CDC, “Road Traffic Injuries and Deaths—A Global Problem,” Centers for Disease Control and Prevention. Accessed: Nov. 20, 2023. [Online]. Available: https://www.cdc.gov/injury/features/global-road-safety/index.html

[2] Y. Golhar and M. Kshirsagar, “Emerging Technologies for Driving Road Safety and Traffic Management for Urban Area,” J. Comput. Sci., vol. 17, no. 11, pp. 1104–1115, Nov. 2021, doi: 10.3844/jcssp.2021.1104.1115.

[3] “Road Accidents in India 2021.” Accessed: Nov. 20, 2023. [Online]. Available: https://morth.nic.in/sites/default/files/RA_2021_Compressed.pdf

[4] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, Art. no. 7553, May 2015, doi: 10.1038/nature14539.

[5] D. Chand, S. Gupta, and I. Kavati, “Computer Vision based Accident Detection for Autonomous Vehicles.” arXiv, Dec. 20, 2020. doi: 10.48550/arXiv.2012.10870.

[6] S. Pourroostaei Ardakani et al., “Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis,” Sustainability, vol. 15, no. 7, Art. no. 7, Jan. 2023, doi: 10.3390/su15075939.

[7] A. Azhar et al., “Detection and prediction of traffic accidents using deep learning techniques,” Clust. Comput., vol. 26, no. 1, pp. 477–493, Feb. 2023, doi: 10.1007/s10586-021-03502-1.

[8] L. Sandeep et al., “A Review on Accident Detection Using Deep Learning to Reduce the Response Time for Medical Help", Int. Res. J. Mod. Eng. Technol. Sci., vol. 04, no. 05, pp. 1907 - 1912, May. 2022

[9] P. Chitale, T. Dhope, A. Akhelikar, and S. Dholay, “Smart Accident Recognition and Alerting System for Edge Devices,” in 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Oct. 2020, pp. 486–491. doi: 10.1109/ICCCA49541.2020.9250902.

[10] D. Chand, S. Gupta, and I. Kavati, “Computer Vision based Accident Detection for Autonomous Vehicles,” in 2020 IEEE 17th India Council International Conference (INDICON), Dec. 2020, pp. 1–6. doi: 10.1109/INDICON49873.2020.9342226.

[11] K. B. Lee and H. S. Shin, “An Application of a Deep Learning Algorithm for Automatic Detection of Unexpected Accidents Under Bad CCTV Monitoring Conditions in Tunnels,” in 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), Istanbul, Turkey: IEEE, Aug. 2019, pp. 7–11. doi: 10.1109/Deep-ML.2019.00010.

[12] H. Ghahremannezhad, H. Shi, and C. Liu, “Real-Time Accident Detection in Traffic Surveillance Using Deep Learning,” in 2022 IEEE International Conference on Imaging Systems and Techniques (IST), Jun. 2022, pp. 1–6. doi: 10.1109/IST55454.2022.9827736.

[13] A. Thaduri, V. Polepally, and S. Vodithala, “Traffic Accident Prediction based on CNN Model,” in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), May 2021, pp. 1590–1594. doi: 10.1109/ICICCS51141.2021.9432224.

[14] D. Singh and C. K. Mohan, "Deep Spatio-Temporal Representation for Detection of Road Accidents Using Stacked Autoencoder," IEEE Trans. Intell. Transp. Syst., vol. 20, no. 3, pp. 879–887, Mar. 2019, doi: 10.1109/TITS.2018.2835308.

[15] G. Rajesh, A. R. Benny, A. Harikrishnan, J. Jacob Abraham, and N. P. John, "A Deep Learning based Accident Detection System," in 2020 International Conference on Communication and Signal Processing (ICCSP), Jul. 2020, pp. 1322–1325. doi: 10.1109/ICCSP48568.2020.9182224.

[16] J. G. Choi, C. W. Kong, G. Kim, and S. Lim, "Car crash detection using ensemble deep learning and multimodal data from dashboard cameras," Expert Syst. Appl., vol. 183, p. 115400, Nov. 2021, doi: 10.1016/j.eswa.2021.115400.

[17] S. Robles-Serrano, G. Sanchez-Torres, and J. Branch-Bedoya, "Automatic Detection of Traffic Accidents from Video Using Deep Learning Techniques," Computers, vol. 10, no. 11, Art. no. 11, Nov. 2021, doi: 10.3390/computers10110148.

[18] "A New Video-Based Crash Detection Method: Balancing Speed and Accuracy Using a Feature Fusion Deep Learning Framework." Accessed: Nov. 08, 2023. [Online]. Available: https://www.hindawi.com/journals/jat/2020/8848874/

[19] "IJISRT21JUN878.pdf." Accessed: Nov. 08, 2023. [Online]. Available: https://ijisrt.com/assets/upload/files/IJISRT21JUN878.pdf

[20] S. Ghosh, S. J. Sunny, and R. Roney, "Accident Detection Using Convolutional Neural Networks," in 2019 International Conference on Data Science and Communication (IconDSC), Mar. 2019, pp. 1–6. doi: 10.1109/IconDSC.2019.8816881.

[21] "Tian et al. - 2019 - An Automatic Car Accident Detection Method Based o.pdf." Accessed: Nov. 08, 2023. [Online]. Available: https://ieeexplore.ieee.org/ielx7/6287639/8600701/08832160.pdf?tp=&arnumber=8832160&isnumber=8600701&ref=aHR0cHM6Ly9zY2hvbGFyLmdvb2dsZS5jb20v

[22] "Accident Mitigation System with Drowsiness Detection: A Machine Learning and Iot with Hybrid Approach | IEEE Conference Publication | IEEE Xplore." Accessed: Nov. 08, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9491646

IoT-based Autonomous Search and Rescue Drone for Precision Firefighting and Disaster Management

Shubeeksh Kumaran, V Aditya Raj, Dr. Sangeetha J, V R Monish Raman
Department of CSE, M. S. Ramaiah Institute of Technology, Bengaluru, India

***Abstract*—Disaster management is a line of work that deals with the lives of people, such work requires utmost precision, accuracy, and tough decision-making under critical situations. Our research aims to utilize Internet of Things (IoT)-based autonomous drones to provide detailed situational awareness and assessment of these dangerous areas to rescue personnel, firefighters, and police officers. The research involves the integration of four systems with our drone, each capable of tackling situations the drone can be in. As the recognition of civilians and protecting them is a key aspect of disaster management, our first system (i.e., Enhanced Human Identification System) to detect trapped victims and provide rescue personnel the identity of the human located. Moreover, it also leverages an Enhanced Deep Super-Resolution Network (EDSR) x4-based Upscaling technology to improve the image of human located. The second system is the Fire Extinguishing System which is equipped with an inbuilt fire extinguisher and a webcam to detect and put off fire at disaster sites to ensure the safety of both trapped civilians and rescue personnels. The third system (i.e., Active Obstacle Avoidance system) ensures the safety of the drone as well as any civilians the drone encounters by detecting any obstacle surrounding its pre-defined path and preventing the drone from any collision with an obstacle. The final system (i.e., Air Quality and Temperature Monitoring system) provides situational awareness to the rescue personnel. To accurately analyze the area and its safety levels, inform the rescue force on whether to take precautions such as wearing a fire proximity suit in case of high temperature or trying a different approach to manage the disaster. With these integrated systems, Autonomous surveillance drones with such capabilities will improve the equation of autonomous Search and Rescue (SAR) operations to a great extent as every aspect of our approach considers both the rescuer and victims in a region of disaster.**

Keywords*—Search and rescue; firefighting; internet of things; disaster management

I. INTRODUCTION

India has a history of being particularly susceptible to natural disasters due to its geo-climatic conditions. Landslides, cyclones, earthquakes, floods, and droughts have been frequent occurrences. Every single year from 1990 to 2000 [1], there were approximately 4344 calamities and over 30 million people were greatly affected by disasters followed by a significant increase in frequency and extremity of natural disasters between 2000 and 2020. The government has initiated steps to improve its rescue support and approach to responding to disasters, but there is lots of room for improvement to protect the vulnerable and significantly reduce the impact of natural disasters in this densely populated country. Over the past few decades, Unmanned Aerial Vehicles (UAV) have increased in popularity rapidly. Drones are UAVs that are controlled using remotes or can fly autonomously [2]. They can be made to perform Search and Rescue (SAR) operations, provide early disaster warnings to humans, analyze the damage, deliver important supplies to isolated or inaccessible locations, and aid in communication with blacked-out areas. This paper deals with providing a sophisticated drone for autonomous search and rescue operations during disasters with the motive to rescue the lives of both the victims and the rescue personnel with the added capabilities of firefighting. This is achieved by the integration of multiple precision IoT sensors and cameras for completely utilizing the drone's position in a zone of disaster that is unsafe for any human. Apart from transmitting real-time sensor data to the ground rescue team, it is also integrated with a sophisticated Enhanced Human Identification System to keep track of all humans encountered during the drone's SAR operation. The paper also explains how inbuilt fire extinguishers installed in drones can be utilized to extinguish dangerous fires located at disaster sites, thereby providing a safe entry/exit path for rescue personnel, and hence improving the rescue effectiveness and efficiency.

In the past several researchers have worked on the various potential applications of drones in disaster management, such as Search and Rescue (SAR), damage assessment and communication. As discussed in study [3] the benefits to explore a range of applications of drones in disaster management and how they can change the entire aspect to solve key challenges involved. Similarly, in [4] presents the design and creation of a low-cost, autonomous drone that is capable of communication, surveillance, and transportation of medical supplies utilizing an integrated GPS and real-time video streaming. In [5], the author conducts a brief survey of video surveillance using drones and highlights the challenges that come along with the technology, such as battery life, data transmission, image processing, and privacy concerns. The high-rise buildings have less accessibility, but when sensors are set up and in case any fire is detected, it is meant to send a message to the control room and give the GPS location to the cloud-enabled drone can be piloted to the affected spots and give situational awareness to firefighters, was the idea given by the author of [6]. Similarly, drones can be used for assisting in rescues during times of disasters [7]. The authors of [8] present the idea of integrating fire-fighting drones that will act as first respondents and will ensure enhanced response time and send necessary details to the control center. [9] presents a similar approach for the early detection of forest fires using two sets of drones, i.e., Fixed wing Unmanned

Aerial Vehicles (UAV) and Rotary-wing UAV with the assistance of artificial intelligence. In study [10] the author gives a brief insight into the concept of using drone as a technology-driven solution that highlights the potential to revolutionize this domain by providing improvised access to healthcare services during times of disaster. In research [11], the authors provide a conceptual framework of drone swarms for fire suppression activities. The framework consists of various stages, such as fire detection, decision-making, and intervention. The results of study [12] indicate the approach of utilizing UAVs as aerial base stations to help in providing reliable communication coverage and connectivity in disaster scenarios. Low-cost-drone-borne synthetic-aperture radar imaging system that can generate high-quality images in various scenarios, including complex terrain and urban environments. The proposed system has the potential range of applications, disaster management, agriculture, and urban planning are a few examples [13]. The study in [14] presents a similar hybrid drone-based radar system for vital sign imaging in disaster management scenarios. Lastly, the development of an Enhanced Deep Super-Resolution Network (EDSR) [15] that can upscale images to make their resolution better can be used to identify humans with utmost precision after locating them.

Overall, the previous and related works show that drones have significant potential for firefighting and disaster-management applications. However, many challenges must be addressed to fully realize the potential of drones in disaster management. Most of the papers have been seen to implement a single task with the help of multiple drones which is not cost-effective and brings in the complication of maintaining them and having to always keep track of their coordination. Similarly, another challenge that was noticed in a few papers was that the overall features for disaster management were not implemented. Keeping these ideas and the demerits of the previous works in mind, this paper implements the integration of a single drone with the concepts of deep learning and the Internet of Things, the drone is infused with multiple systems such as the Air quality and temperature monitoring system, Active Obstacle Avoidance system and most importantly the Fire Extinguishing system. This brings novelty into the research in the implementation approach that has not been explored yet until now.

The rest of this paper is organized as follows. Section II explains the methodology of the proposed work where we introduce a system that has four systems integrated with our drone which consist of an Enhanced Human Identification system, Fire Extinguishing system, Active Obstacle Avoidance system and Air Quality and Temperature Monitoring system. In Section III, we will delve into Hardware setup, Section IV provides the results. The paper ends with Section V and Section VI respectively where we have written concluding remarks on the potential capabilities of the research and future scope.

II. METHODOLOGY

This research work depends on the strong autonomous capabilities of the drone. Being autonomous is the foundation and is achieved by the integration of a companion computer Raspberry Pi 3B which is connected to the flight controller, the Pixhawk 2.4.8 as shown in Fig. 1. The connection is made via the telemetry port on the Pixhawk. All the data is communicated using appropriate MAVLINK protocols. With this setup, we have achieved Raspberry Pi-controlled drone movements.

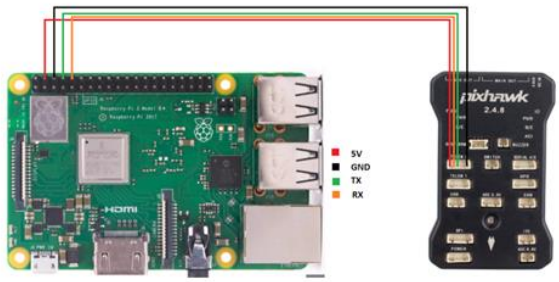


Fig. 1. Connections from raspberry Pi to Pixhawk 2.4.8.

Now the drone is capable of autonomous takeoff, landing, and maneuvering based on both a predefined path, as well as the surrounding conditions perceived by the sensors and cameras used. These main functionalities have been explained in detail in the following system:

1) *Enhanced* human identification system
2) *Fire* extinguishing system
3) *Active* obstacle avoidance system
4) *Air* quality and temperature monitoring system

A. Enhanced Human Identification System

The drone is equipped with an inbuilt webcam to provide rescue personnel with live footage to keep them regularly updated on the surrounding environment of the affected disaster site. However, the main purpose of this system is to provide a human detection system to identify trapped victims in disaster sites. This enables us to keep track of all the humans encountered in the drone's path during its flight. Unfortunately, the drone is expected to be recording subpar images in harsh situations where surroundings are filled with smoke, dust, fire, and relative movement of the drone and human. Hence, we are leveraging a EDSR based image resolution enhancement architecture to be able to identify the human with increased ease after the face detection system stores the faces in the database. Therefore, we are able to achieve a resulting output image with a four times Scaling Factor.

1) Advanced face detection: We have implemented the human identification by deploying a face detection feature that uses OpenCV and Deep-Learning technique called Single Shot Detector (SSD), where once a face is identified, the image of that face is clicked and saved in a database which is accessible only to rescue personnels. Once the image is stored, an alert message is sent to the rescue personnel with a prompt, informing them that a victim has been identified. This feature uses the "Twilio" API to send messages which uses a secure authentication token and a specific SSID which makes this a highly secure communication line as seen in Fig. 2. Using this

information, rescue personnel can keep track of the victim's location and carry out their rescue operation more efficiently.

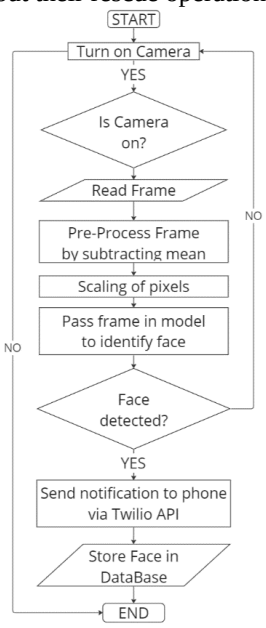


Fig. 2. Workflow of advanced face detection.

In the Algorithm 1, we have explained the Single Shot Detector which is used to identify the human face from an input frame.

Algorithm 1: Single Shot Detector

a) Load the pre-trained model and define the input size.

b) For each image in the input:

- Preprocess the image.
- Feed the image through the network.
- Obtain the predicted class scores and bounding box coordinates.
- Apply non-maximum suppression to remove overlapping boxes.
- Draw the remaining boxes on the image and output the result.

2) EDSR x4 image enhancement: The Enhanced Deep Super Resolution (EDSR) x4 architecture is based on a sophisticated CNN specifically designed to upscale low-resolution images to specifically bring out the hidden details. It is inspired by the ResNet architecture that is characterized by a deeply stacked layers which leverages residual connections. The architecture initially employed a scaling factor of 2, subsequently 3 and then finally 4 in order to reaching a x4 scaling factor. This accelerates both the training and the overall performance of the model. The images form the database are retrieved to process through EDSR and the Layers and Components for EDSR x4 as seen in Fig. 3 is explained as follows:

a) Input Layer:

- Accepts the low-resolution image as input.

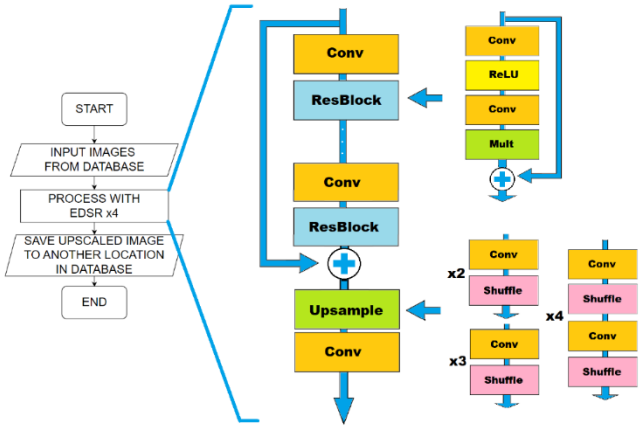


Fig. 3. Workflow of EDSR x4 image enhancement.

b) Feature Extraction Layers:

- Utilizes convolutional layers without batch normalization and ReLU activation to extract relevant features from the input image.
- ResNet-style architecture is employed for feature extraction.

c) Residual Blocks:

- These are the core building blocks of EDSR x4, following the ResNet style.
- Each residual block consists of:
  - Convolutional layers without batch normalization and ReLU activation.
  - Skip connections that bypass one or more convolutional layers and are added to the output.
  - Constant scaling layers placed after the last convolutional layers in each residual block with a factor of 0.1.
- The absence of batch normalization allows for greater range flexibility in the features' networks.

d) Skip Connections:

- These connections enable the network to learn residual details, facilitating the learning of high-frequency details.

e) Upsampling Layers:

- Utilizes techniques like nearest-neighbor interpolation or transposed convolution to upscale the feature maps.
- These layers increase the spatial resolution while preserving the learned features.

f) Output Layer:

- Produces the high-resolution image as the final output.
- By processing the human face images from the database using this technique, we can overcome the drawback of blurred images taken by the drone and

hence identify the human with ease due to its higher resolution resulting image.

B. Fire Extinguishing System

This system mainly deals with the detection and extinguishing of fire at disaster sites to ensure safety of both trapped civilians and rescue personnel. These outcomes are achieved by utilizing Raspberry Pi for on-board computations required to detect fire through an inbuilt Raspberry Pi camera along with servo motor to activate the fire extinguisher at the bottom of the drone. First, the armed drone monitors a particular area in its pre-defined path to check for any fires. Once the fire is detected, the drone lowers its altitude to get an effective range to carry out extinguishing process by using the inbuilt fire extinguisher after which it returns to optimum altitude to resume its surveillance process. The entire workflow of the fire extinguishing system is shown in Fig. 4.

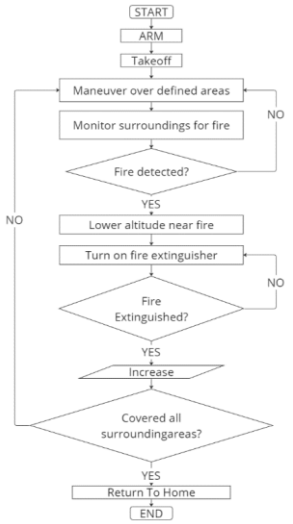


Fig. 4. Workflow of fire extinguishing system.

This classifier contains patterns of fire in images that have been previously trained, and it is used to identify potential fire regions in the video frames. Initialize Camera:

The cv2.VideoCapture() function is used to initialize the camera and establish a connection to capture live video feed. The camera is set to capture frames continuously in a loop.Process Video Frames: The code captures video frames from the camera in a loop using vid.read() function. For each frame, it performs the following operations:

1) Convert to grayscale: The frame is converted to grayscale using cv2.cvtColor() function. This step reduces the computational complexity and simplifies the fire detection process as the color information is not needed for fire detection.

2) Detect fire regions: The grayscale frame is passed to the loaded cascade classifier using fire_cascade.detectMultiScale() function. This function detects potential fire regions in the frame based on the patterns learned by the classifier.

3) Draw rectangles: For each detected fire region, a rectangle is drawn around it uses cv2.rectangle() function. This visually highlights the detected fire regions in the frame.

4) Display processed frame: The frame with highlighted fire regions is displayed using cv2.imshow() function. This allows visual inspection of the processed frame with detected fire regions.

5) Exit the loop: The code continues capturing video frames, applying fire detection, playing the alarm sound, and displaying processed frames until the 'q' key is pressed. Once the 'q' key is pressed, the loop is exited, and the code terminates.

Overall, the code continuously captures video frames, detects fire regions in the frames, plays an alarm sound, and displays processed frames with highlighted fire regions, providing a guide to the fire extinguishing system using a cascade classifier and a camera.

C. Active Obstacle Avoidance System

The active obstacle avoidance system is the main backbone of the autonomous flying aspect of the drone; it is to ensure the safety of the drone as well as any civilians the drone encounters. The system works primarily with the help of the four HC-SR04 sensors on each side of the drone as seen in Fig. 5, which are connected to the Arduino-Uno which is in turn connected to the Pixhawk flight controller. The drone's front and back movement is controlled by the Pitch. If the Pitch increases the drone moves backwards and if it decreases the drone moves forward. Similarly, the roll of a drone refers to the rotation of the drone around its longitudinal axis. When a drone rolls, it tilts to the left or right while keeping its heading unchanged. Roll controls the drone's left and right movements. If the roll increases the drone moves towards the right and if it decreases the drone moves towards the left.

Fig. 5. Representation of the front, back, right, and left sensors on the drone.

The Pixhawk flight controller controls the movement of the drone and receives commands from the remote control through its radio receiver. The Arduino-Uno is linked to the Pixhawk's RX and TX ports to integrate the active obstacle avoidance system with the flight controller. This is done to enable communication between the two devices. The Arduino-Uno sends obstacle detection information to the Pixhawk flight controller via the TX pin, which the Pixhawk interprets to adjust the drone's flight path to avoid the detected obstacles. The reason why digital pins 10 and 11 are used for this connection is that they are designated as hardware serial

communication pins on the Arduino-UNO board. By using hardware serial communication, the communication between the Arduino-Uno and the Pixhawk flight controller is fast, reliable, and minimizes the risk of communication errors

The first step in the methodology is to acquire the data from the ultrasonic sensors. The ultrasonic sensors are used to detect obstacles in the immediate vicinity of the drone and calculate the distance of the obstacle from the drone. The data acquired from the sensors is processed using algorithms that calculate the distance of the obstacles from the drone. Once the distance from the ultrasonic sensors has been acquired the sensors change the Pitch or roll value of the drone depending on which sensor has detected an obstacle.

The equation for the change in Pitch when an obstacle is detected by the FRONT SENSOR within its range is given by Eq. (1):

$$PITCH\_BACK = 1500 + 30 + ((100 - FRONT\_SENSOR) * 6) \quad (1)$$

where PITCH_BACK = the Pitch value to make sure the drone moves back,

1500 = Mid Term,

30 = initial increment when object detected at 99cm,

100 = Range of the sensors(cm),

FRONT_SENSOR = Front sensor distance (cm).

The relation between the Pitch and the distance is given by the following graph in Fig. 6.

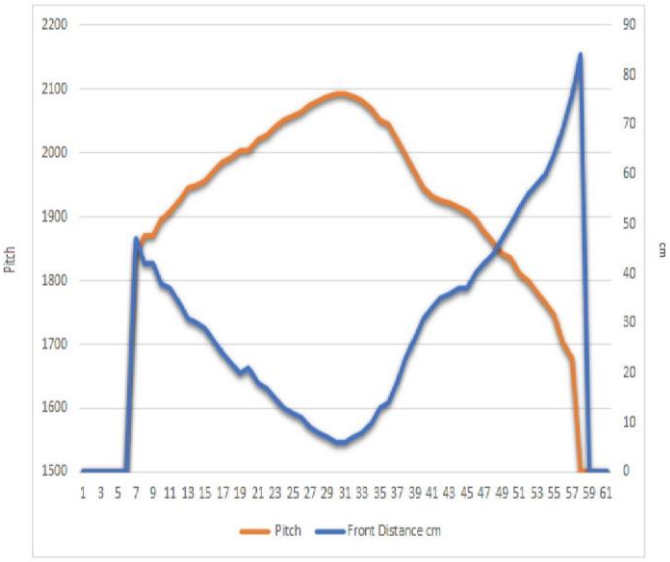


Fig. 6. Graph depicting the relation between pitch and the front distance.

Fig. 7 shows the working of the active obstacle avoidance system wherein we can see, once the ultrasonic sensors start operating, they start detecting any obstacle within the radius of 100 centimeters around it. Then through the digital pin number, they get to know which sensor has detected an obstacle and depending on that the roll or the Pitch must be changed, and then it either increases or decreases Pitch if the object is detected in the front or back respectively. The roll is changed by either decreasing or increasing it when the obstacle is detected on the right or left side respectively.

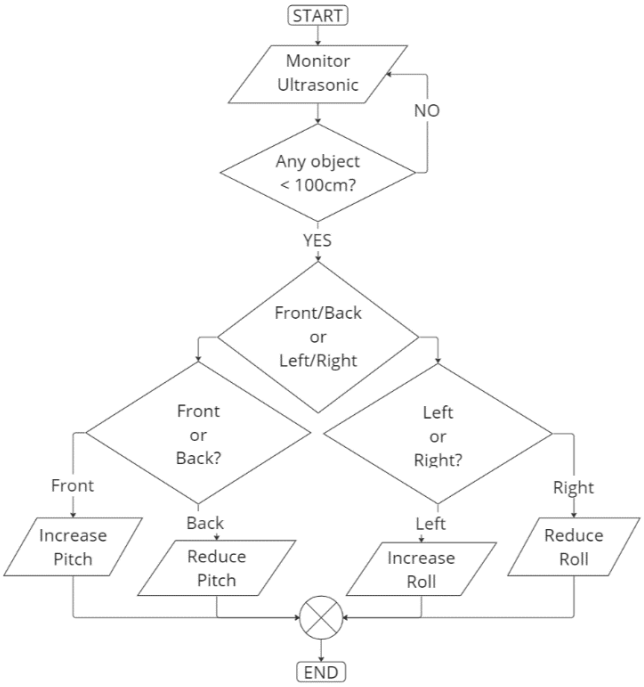


Fig. 7. Workflow of active obstacle-avoidance system.

D. Air Quality and Temperature Monitoring System

This system takes advantage of the drone's position in a region of disaster prior to the arrival of the rescue team. Alerting the rescue teams is done based on current environment conditions which are determined by several factors like temperature, humidity and air quality. To monitor these factors regularly our drone is equipped with a Temperature and smoke sensor which is seamlessly integrated into smartphones through an application called Blynk.App which obtains the data through the NodeMCU ESP8266 module and stores the data in Blynk.Cloud. Blynk.App is a mobile application to prototype, deploy and manage connected electronic devices at any scale. Real-time data are obtained by these connected devices and are stored in the Blynk.Cloud. Using an API, the devices easily connect to the platform and take advantage of all its advance features. The live values of the surrounding environment are being read by this system and can be viewed through the application for monitoring purposes.

1) Air quality monitoring: The sensor used to perceive data is the MQ2 sensor, which is sensitive to H_2, LPG, CH_4, CO, alcohol, smoke, propane, and air. Fig. 8 explains the step-by-step working of this monitoring system. The toxicity of the air and the presence of smoke is what we are interested in this Air Quality System. Once the sensor starts, the toxicity of the air is calculated by the concentrations of methane, ethane, butane, propane, and hydrogen present in the surrounding environment. The final toxicity value has to be perceived in terms of parts per million (ppm) for the convenience of easily

making decisions. The ppm calculation is based on the resistance ratio (RS/R0). RS is calculated using Eq. (2). Using the RS value calculated in Eq. (1) we find the R0 value using Eq. (3).

$$RS = [(VC - RL)/Vout] - RL \quad (2)$$

$$R0 = RS/\text{Fresh air ratio value from the datasheet} \quad (3)$$

Where RS = Change in resistance on detecting gas,

R0 = Stable sensor resistance in the fresh air,

V_C = Voltage Current,

V_{out} = Output voltage,

RL = Load Resistance.

Now that we have access to the live ppm values, we are able to conveniently perform analysis. In this research work, we have assumed three values of the threshold for monitoring purposes (i.e. 500 ppm, 700 ppm, and 800 ppm) for feasible testing. These values can be changed and set accurately for real-life scenarios. So, based on these three threshold values we generate appropriate response messages which will be uploaded to the Blynk.Cloud from where the message is sent to the user's phone via the Blynk.App. This way the rescue team can utilize appropriate safety gears and precautions based on the severity of the situation determined by this system.

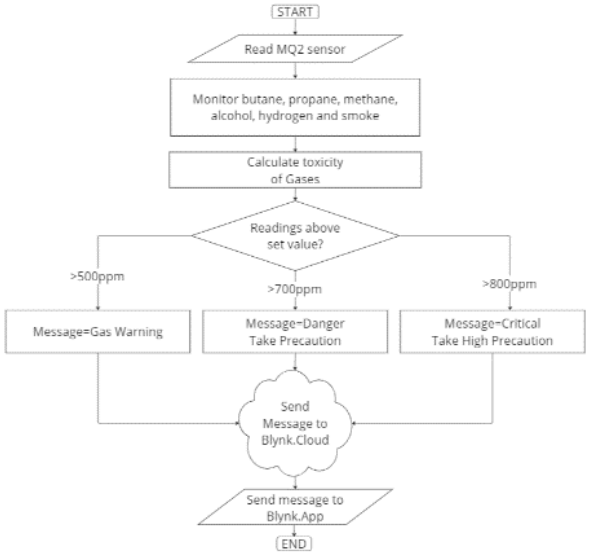


Fig. 8. Workflow of air quality monitoring system.

2) Temperature monitoring system: We are using the DHT11 sensor to monitor the temperature in the surrounding area. This sensor is connected to both the Raspberry Pi and the NodeMCU 8266 module. The workflow of this system is shown in detail in Fig. 9. Once the DHT11 Sensor starts, we obtain the temperature value in Celsius. Based on what temperatures the human body can bear or withstand, we have taken into consideration 35°C and 40°C as threshold values.

Based on the temperature threshold, the response message will help the rescue team to take appropriate safety measures. These appropriate response messages will be uploaded to the Blynk.Cloud from where the message is sent to the user's phone via the Blynk.App. These messages generated are sent to the rescue team and ground station to provide awareness about the situation in the disaster zone. Alerts are sent to the smartphone in several ways such as Blynk application as a popup alert, phone notification with the alert message and Sent as a mail to the recipient of choice with relevant information obtained from Blynk.Cloud.

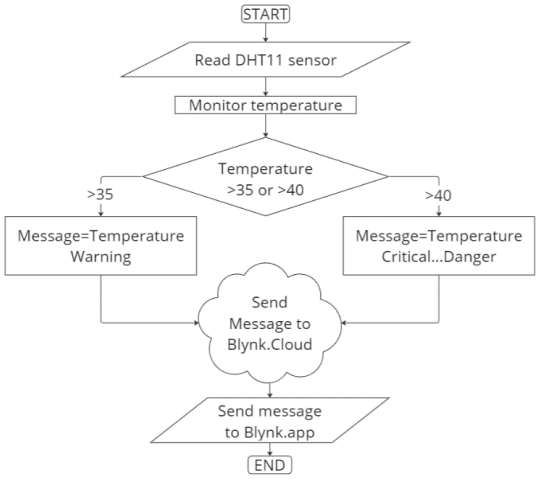


Fig. 9. Workflow of temperature monitoring system.

In real life scenarios, there are possibilities where there exists a high temperature environment where the content of smoke in the area might be less or even negligible, likewise, even the possibility of high smoke presence without the presence of fire as shown in Fig. 10.

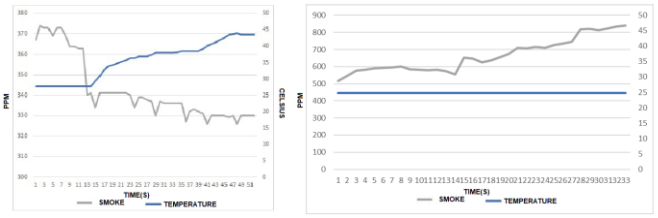


Fig. 10. Sensor readings for cases of high smoke / temperature.

Alerts are based on predefined safety values to ensure a safe and assessed environment as seen in Table I.

TABLE I. SPECIALIZED ALERT MESSAGES BASED ON ENVIRONMENTAL CONDITIONS

Parameter	Condition	Alert Message
Temperature	Temperature is greater than 30°C.	"Temp Over 30°C."
Temperature	If the temperature is greater than 45°C.	"Take Precaution!! Temp Over 45°C."
Gas Toxicity/Smoke	If greater than 500 ppm.	"Gas Conc Over 500 Ppm."
Gas Toxicity/Smoke	If greater than 700 ppm.	"Take Precaution!! Conc Over 700 Ppm."
Gas Toxicity/Smoke	If greater than 800 ppm.	"Condition Critical Gas Conc Over 800 Ppm."

III. HARDWARE SETUP

The various components required to establish the four systems have to be placed on the drone accordingly. The Fig. 11 shows how the Enhanced Human Identification system where the camera that is attached to the Raspberry Pi is placed in front of the drone such that we get the exact look at the Drone's view. The Active Obstacle Avoidance system consisting is placed on all four sides of the drone as seen in Fig. 11.

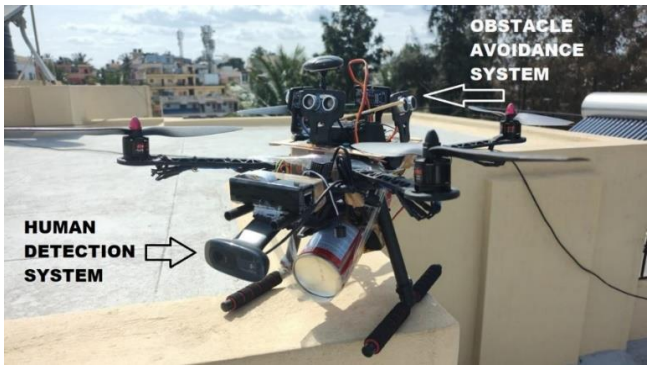


Fig. 11. Setup of the enhanced human identification system and the active obstacle avoidance system.

Fig. 12 shows the setup of the Fire Extinguishing system where the Fire extinguisher is placed under the drone and connected to a servo motor that presses on the spray when fire is detected.

Fig. 12. Setup of the fire extinguishing system.

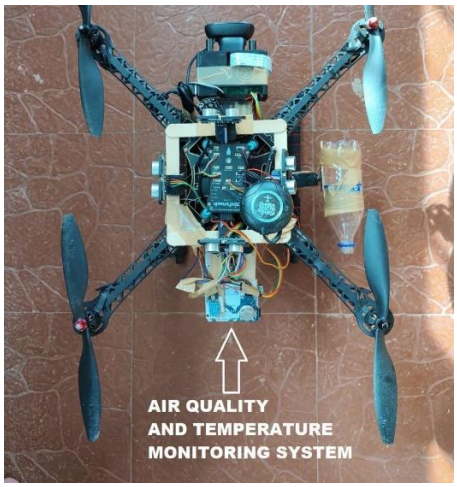


Fig. 13. Setup of the air quality and temperature monitoring system.

Lastly, Fig. 13 shows the setup of the MQ-2 sensor and DHT-11 sensors which are a part of the Air Quality and Temperature Monitoring system is attached over the drone's length such that the sensor readings are not affected by the heat generated by the components.

IV. RESULTS

The IoT-based autonomous drone has the true potential to revolutionize aid and relief in the field of disaster management. The drone's advanced capabilities to collect real-time data, analyze complex situations and facilitate well-informed decision-making can make a huge impact in the effectiveness of emergency response efforts, leading to an efficient method with increased safety measures to minimize the impact of disasters. To explain the working of our proposed solution, we have divided the following explanations into each of the dedicated systems of our proposed model which includes Enhanced Human Identification System, Fire Extinguishing System, Active Obstacle Avoidance System and Air-Quality and Temperature Monitoring System.

A. Enhanced Human Detection Identification System

1) Advanced face detection: The face detection system is used to keep track of the civilians encountered by the drone during its surveillance operation. This system works in real-time and continuously monitors the surroundings for humans. This happens as shown in Fig. 14. The Face Detection System uses Single Shot Detector Algorithm to detect trapped victims as seen in Fig. 14(a) and is stored into a dataset to keep track of the trapped victim as seen in Fig. 14(b). Once the victim is detected an alert message is sent to the rescue personnel to notify them of the trapped victim as seen in Fig. 14(c).

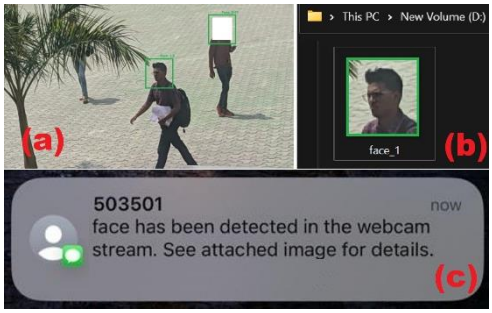


Fig. 14. (a): Detection of facial features in real time (b) Detected faces stored at dedicated location (c) Push notification alerts when face is detected.

2) EDSR x4 image enhancement: Once the image is stored in the database, by the Advanced Face Detection, it is enhanced to achieve a better resolution to identify the human easily. In Fig. 15(a), we can see the original image, and Fig. 15(b) shows the enhanced image using the EDSR Super-Resolution. It is evident that the enhanced image is more detailed and the facial features are brought up. Additionally, it has enhanced the 100 pixel x100 pixel image into a 400 pixel x 400 pixel image. This enhancement is done after its stored in the database because this process is resource intensive and therefore not viable to be a part before its stored in the database in the Advanced Face Detection System.

B. Fire Extinguishing System

This system utilizes a pre-trained cascade classifier-based fire detection system which can accurately detect the fire shown. Since we could not replicate large scale fires, we used images from online resources for testing purposes before deploying it on the drone. This image given as input can be seen in Fig. 16(a) and the camera's perspective of it is given in Fig. 16(b).

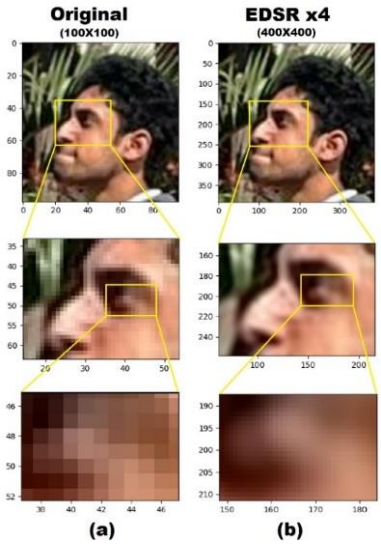


Fig. 15. (a): Original image from the database (b) EDSR x4 scaled image.

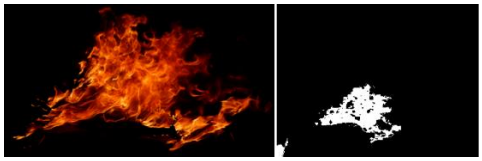


Fig. 16. (a) Sample fire used for testing fire detection (b) Camera's perspective of the fire.

The drone in on the ground as shown in Fig. 17(a), and it begins to take off and surveil in the defined areas. Once the drone encounters the fire region, it begins to extinguish the fire as shown in Fig. 17(b). We have been able to come up with an approach that effectively detects fire in the presence of high temperatures and smoke and tries to extinguish it. In our testing and evaluation phase, it was found that the system detected the fire with huge accuracy and precision.

Fig. 17. (a) Drone on ground with paper on fire in front (b) Drone extinguishing the fire.

C. Active Obstacle Avoidance System

Obstacle avoidance is an add-on feature to the autonomous maneuverings of the drone to ensure the drone does not collide with anything. The Arduino-Uno is used for this specific purpose, with it we have placed HC-SR04 ultrasonic sensors in all four directions of the drone to consider obstacles from any of the four sides. This system has multiple use cases, let us consider a civilian running away from disaster in a state of panic, and may not notice a drone in the field due to several reasons such as loud surroundings, too much smoke to look around, etc. Collision with the drone will cause injuries to the civilian which is against the objective of the drone's mission. Hence by using ultrasonic sensors, we overcome this problem. Another such use case can be the possibility of a building wall breaking down towards the drone where it must move away to protect itself. Hence, this system is programmed to consider all the obstacles at any given time and prioritizes the obstacle closer to the drone over the obstacle further away from it. Hence, we shall now consider these cases and understand the working of the active obstacle avoidance system. Hence, from the below cases considered it is observed that various combinations of sensor input are being given, the reaction of the drone is consistent and prioritizes the closer obstacle ensuring the safety of the objects surrounding it.

Case 1 - Obstacle in the Front and Back of the Drone.: There is an obstacle in the front and back of the drone. In this case, as we can see in Fig. 18(a), as the front distance increases the Pitch decreases and as the back distance increases the Pitch increases.

Case 2 - Obstacle in the Front and Right of the Drone: As we can clearly see from the graph below in Fig. 18(b), when there is an object in the front the pitch changes and the roll remain unchanged, while when there is an object detected in the left the pitch remains unchanged while the roll changes. Finally, when there is an obstacle detected both in the front and back the object changes both its pitch and roll which makes it move diagonally and the graph in Fig. 18(b) depicts the relationship between them.

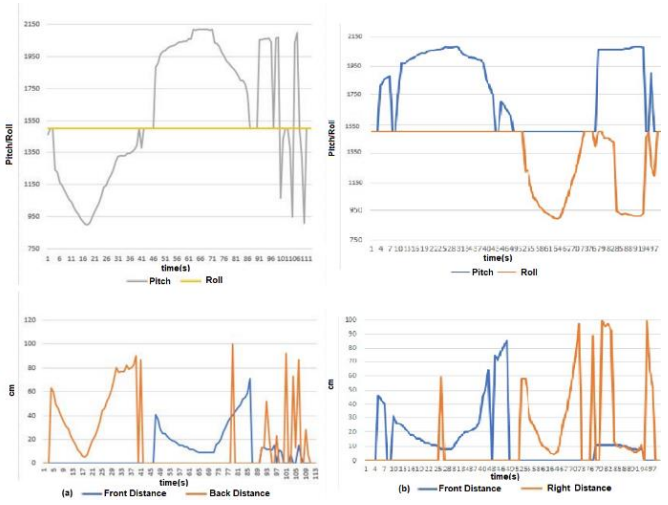


Fig. 18. Graph depicting the relation between Pitch, roll, and front and right distance for Case 1 (18 (a)), Case 2 (18 (b)).

Let us consider a situation as shown in Fig. 19(a) wherein the drone detects an obstacle with its FRONT SENSOR within the set range, now the drone has to move backwards and that can be achieved by increasing the Pitch of the drone as per the Eq. (3) as seen in Fig. 19(b).

D. Air Quality and Temperature Monitoring System

This system consists of sensors integrated with the drone to monitor the temperature, humidity and toxicity levels of the surrounding area to alert rescue personnel of the need to prepare themselves with high safety measures with the required precaution that can save their lives.

Fig. 19. (a): Obstacle introduced in front of the drone (b) Drone moving away from the introduced obstacle.

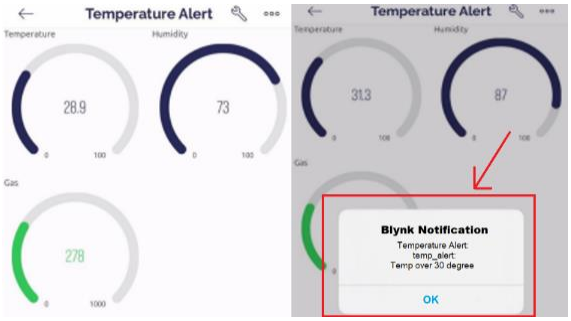


Fig. 20. (a): Live monitoring of sensor values on smartphone (b) Illustration of blynk.app alerting the user in case of high temperature.

The values read by the sensors are processed and shown in an easy minimalist way on the Blynk application for easy understanding by the user as seen in Fig. 20(a) which has gauges for temperature, humidity and gas. The depicted gas value is the combined toxicity of the composition of Butane, Methane, LPG, Smoke, Alcohol, Propane, and Hydrogen which is measured in ppm. If there is a spike in any of these levels, then the smartphone is alerted regarding the critical status based on the scenario in the following methods: Firstly, as seen in Fig. 20(b), the Blynk.App notifies the user with a popup alerting the user about the increased temperature conditions at the current location of the drone and similarly displays popups for unusually high gas toxicity. Secondly, As phone notification with the alert message. Fig. 21(a) shows the alerts for temperatures above 30 ℃. Similarly, notifications will be provided for higher temperatures as well. As seen in Fig. 21(b), the notification system ensures the user is alerted even when not using the Blynk application by providing a push notification feature which demonstrates the different notifications sent to the phone based on our tested gas toxic levels of 500ppm, 700ppm and 800ppm as per Table I and finally, Sent as a mail to the recipient of choice with relevant information.

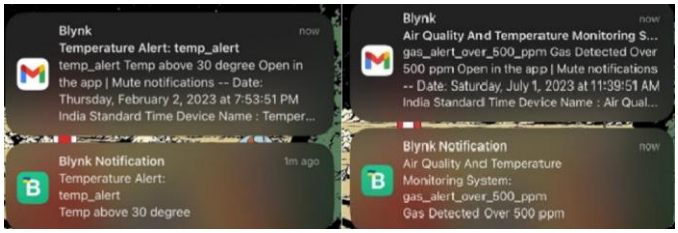


Fig. 21. (a): Blynk.app notification and mail sent to users for high-temperature.(b): Blynk.app notification and mail sent to users for high-toxic levels.

These alerts to the user are purely based on the readings from the sensors used. Hence, we are successfully able to accurately analyze the environmental situation using the onboard sensors and provide the rescue team with appropriate alert messages which will allow them to decide their approach to rescue. They can also opt to wear fire proximity suits in case of high temperatures, or chemical mask suits for protection against breathing toxic gases during the rescue. Due to a lack of testing with real toxic gases, we have set the temperature and gas threshold to ideal values to show output alerts. However, these values can be tuned as per real-life conditional levels easily.

V. Conclusion

In conclusion, the use of IoT-enabled drones for disaster management has enormous potential for enhancing the efficiency of emergency services during natural disasters and other emergencies. Rescue teams can quickly survey disaster-affected areas, find survivors, evaluate damage, and give first responders, on the ground real-time, situational awareness by utilizing these four systems namely Enhanced Human Identification System, Fire Extinguishing System, Active Obstacle Avoidance System and Air-Quality and Temperature Monitoring System. These systems provide added capabilities to these unmanned aerial vehicles solving major problems faced by rescue personnel. Emergency rescue personnel can use this knowledge to make wise decisions in saving the victims. With added Deep Learning model, the drone is more efficient in detecting human and fire, so that we can identify the trapped victims in disaster sites and extinguish dangerous fires, thereby creating an effective rescue process which ensures safety of victims and the rescue personnel involved.

VI. Future Work

Our current implementation of the method was designed with careful consideration of our economic constraints. The proposed solution adopts a cost-effective approach, employing a Raspberry Pi 3B as a companion computer and a basic Pixhawk 2.4.8 flight controller. However, the reliability of our obstacle avoidance system, utilizing the SR04 sensor, falls short of industrial standards due to its low resolution in distance input. Furthermore, the software approach employed for face detection, utilizing the Single Shot Detector, though faster, lacks accuracy. To enhance the overall efficiency of our methodologies, acquiring industry-standard components for drones is imperative. The following outlines the future scope of the project:

1) Incorporating a Nvidia Jetson Nano as a companion computer will enable us to integrate more sophisticated and resource-intensive algorithms, thereby enhancing computational capabilities.

2) The integration of thermal cameras is proposed as a more effective alternative to traditional cameras. Thermal cameras demonstrate improved capability in detecting humans through smoke, dust, and debris, making them invaluable in challenging environments.

3) Exploring more effective and lightweight alternatives to traditional fire extinguishing techniques is a potential avenue for improvement. This could involve adopting innovative approaches to firefighting that align with the latest industry standards.

By implementing these advancements, we anticipate a significant improvement in the overall efficiency and performance of our proposed methodologies for drone applications in various scenarios.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] UNICEF. "Disaster Risk Reduction." UNICEF India. [Accessed on June 17, 2023]. Available at: https://www.unicef.org/india/what-we-do/disaster-risk-reduction.

[2] Tanzi, Tullio Joseph, Madhu Chandra, Jean Isnard, Daniel Camara, Olivier Sebastien, and Fanilo Harivelo. "Towards" drone-borne" disaster management: future application scenarios." In XXIII ISPRS Congress, Commission VIII (Volume III-8), vol. 3, pp. 181-189. Copernicus GmbH, 2016.

[3] Daud, Sharifah Mastura Syed Mohd, Mohd Yusmiaidil Putera Mohd Yusof, Chong Chin Heo, Lay See Khoo, Mansharan Kaur Chainchel Singh, Mohd Shah Mahmood, and Hapizah Nawawi. "Applications of drone in disaster management: A scoping review." Science & Justice 62, no. 1 (2022): 30-42.

[4] Saha, HimadriNath, Srijita Basu, Supratim Auddy, Ratul Dey, Arnab Nandy, Debjit Pal, Nirjhar Roy et al. "A low cost fully autonomous GPS (Global Positioning System) based quadcopter for disaster management." In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), pp. 654-660. IEEE, 2018..

[5] Dilshad, Naqqash, JaeYoung Hwang, JaeSeung Song, and NakMyoung Sung. "Applications and challenges in video surveillance via drone: A brief survey." In 2020 International Conference on Information and Communication Technology Convergence (ICTC), pp. 728-732. IEEE, 2020.

[6] Jayapandian, N. "Cloud enabled smart firefighting drone using internet of things." In 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 1079-1083. IEEE, 2019..

[7] Apvrille, Ludovic, Tullio Tanzi, and Jean-Luc Dugelay. "Autonomous drones for assisting rescue services within the context of natural disasters." In 2014 XXXIth URSI General Assembly and Scientific Symposium (URSI GASS), pp. 1-4. IEEE, 2014.

[8] Mnaouer, Hajer Ben, Mohammad Faieq, Adel Yousefi, and Sarra Ben Mnaouer. "FireFly Autonomous Drone Project." arXiv preprint arXiv:2104.07758 (2021)..

[9] Kinaneva, Diyana, Georgi Hristov, Jordan Raychev, and Plamen Zahariev. "Early forest fire detection using drones and artificial intelligence." In 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1060-1065. IEEE, 2019.

[10] Pathak, Pankaj, Madhavi Damle, Parashu Ram Pal, and Vikash Yadav. "Humanitarian impact of drones in healthcare and disaster management." Int. J. Recent Technol. Eng 7, no. 5 (2019): 201-205.

[11] TAusonio, Elena, Patrizia Bagnerini, and Marco Ghio. "Drone swarms in fire suppression activities: A conceptual framework." Drones 5, no. 1 (2021): 17.

[12] Panda, Kirtan Gopal, Shrayan Das, Debarati Sen, and Wasim Arif. "Design and deployment of UAV-aided post-disaster emergency network." IEEE Access 7 (2019): 102985-102999.

[13] Bekar, Ali, Michail Antoniou, and Christopher J. Baker. "Low-cost, high-resolution, drone-borne SAR imaging." IEEE Transactions on Geoscience and Remote Sensing 60 (2021): 1-11.

[14] Yan, Jiaming, Zhengyu Peng, Hong Hong, Hui Chu, Xiaohua Zhu, and Changzhi Li. "Vital-SAR-imaging with a drone-based hybrid radar system." IEEE Transactions on Microwave Theory and Techniques 66, no. 12 (2018): 5852-5862.

[15] Lim, Bee, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. "Enhanced deep residual networks for single image super-resolution." In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 136-144. 2017.

Contactless Palm Vein Recognition System with Integrated Learning Approach System

Ram Gopal Musunuru[1], Dr. T Sivaprakasam[2], Dr G Krishna Kishore[3]
Research Scholar, Department of Computer Science and Engineering,
Annamalai University, Annamalainagar, Tamilnadu, India[1]
Assistant Professor, Department of Computer Science and Engineering, Annamalai University,
Annamalainagar, Tamilnadu, India[2]
Professor, Department of CSE, Dhanekula Institute of Engineering and Technology, Vijayawada, Andhra Pradesh, India[3]

***Abstract*—Palm Vein Recognition (PVR) is a new biometric authentication technology that provides both security and convenience. This paper describes a contactless PVR system (CPVR) that uses an integrated learning approach (ILA) to recognise the palm veins from the given input images while ensuring user comfort and ease of use. Contactless palm vein scanning technology is used in the proposed system, eliminating the need for physical contact with the scanning device. The proposed method combines advanced feature extraction techniques with a light gradient boosting machine (LightGBM) and transfer learning. A pre-trained model, EfficientNetB1, is used to train the model to extract significant factors from the input PVR images. The proposed method improves user comfort and reduces the risk of cross-contamination in environments where hygiene is critical, such as hospitals, banking, and other secured places. The cutting-edge contactless palm vein scanner captures the unique vein patterns beneath the user's palm without requiring direct physical contact. The proposed ILA illuminates and captures vein patterns using near-infrared (NIR) light, ensuring high accuracy and robustness. The system employs advanced pre-processing techniques and enhanced image segmentation techniques to continuously improve recognition accuracy. It adjusts to changes in the user's vein patterns over time, considering factors like ageing and injuries. The ILA improves the system's ability to adjust palm positioning and lighting changes. The ILA is also a Contactless Palm Vein Recognition System with numerous applications, such as access control, secure authentication for financial transactions, healthcare record access, and more. The system is built to be scalable, allowing organisations to use it in various settings, ranging from small-scale installations to large enterprise-level deployments. Finally, the proposed approach ILA used to recognise accurate users increased the detection rate.**

Keywords*—Palm Vein Recognition (PVR); Light Gradient Boosting Machine (LightGBM); Transfer Learning; Integrated Learning Approach (ILA)

I. INTRODUCTION

A Contactless Palm Vein Recognition System (CPVR) is a form of biometric identification that identifies people based on the unique vein patterns in their palms. It is an intensely reliable and accurate authentication method that can be used for various applications such as access control, identity verification, and secure interactions. The system typically uses near-infrared (NIR) light to illuminate the palm's veins. Hemoglobin in the blood absorbs this light, creating a distinct pattern of dark lines (veins) against a brighter background. The reflected light is captured by a camera or sensor, which makes an image of the palm's vein pattern. Advanced algorithms are used to extract and encode the unique vein pattern as a biometric template from this image. Because it requires a living hand with blood flow to function, Palm Vein Recognition (PVR) is known for its accuracy and resistance to spoofing. There are several limitations over the 2D-PVR such as less authentication, mismatched results Etc. In CPVR the advanced technique that helps in providing the high security is full-view 3D finger vein verification technique [1]. This technique leverages the unique patterns of blood vessels within an individual's finger to verify their identity [2]. In biometrics and computer vision, learning significant and selective indicators for palm print feature extraction and identification is essential [3]. Palm print recognition is a biometric authentication technique that uses an individual's palm's unique patterns of lines, ridges, and wrinkles [4].

Collecting and organizing data from various sources, ensuring accuracy and diversity, and implementing a secure storage and access system are all part of creating an innovative multidimensional vein database with fingerprints from the palm dorsal and wrist [5]. CNNs are commonly used in computer vision tasks because they can learn hierarchical features from images automatically. Concatenation refers to the process of combining or stacking features learned at different layers of the network for further processing [6]. A semantic feature selector is a component that aids in the identification and selection of the most pertinent features from a concatenated set of features. This step is critical for reducing the dimensionality of the data and retaining only the most discriminative features for vein recognition [7]. The concatenated feature vector is subjected to a semantic feature selector or another machine learning algorithm. This selector aids in the identification of the most informative features while excluding less relevant ones [8]. It could employ techniques like feature importance scores or attention mechanisms. Based on the vein pattern, the system uses classification or similarity scores to authenticate or verify the individual's identity.

The auto-encoder is trained to reduce the reconstruction error between the input vein images and the decoder output. This procedure fine-tunes the encoder to extract relevant vein patterns while filtering out noise and irrelevant data [9]. To

learn a compact and meaningful representation of the vein images, a DenseNet-based auto-encoder is used. Because of their dense connections and feature reuse, DenseNet architectures have produced excellent results in image-processing tasks. The auto-encoder's encoder compresses the input images into a lower-dimensional latent space while preserving essential features. When an individual's identity is verified, their vein image is passed through the trained encoder, producing a feature vector. Then, between this feature vector and the feature vectors stored in the database, a similarity measure, such as cosine similarity or Euclidean distance, is computed [10]. The individual is verified as a legitimate user if the similarity score exceeds a predefined threshold.

II. Literature Survey

Cho et al. [11] introduced a novel approach that identifies the palm vein and the palm-print obtained from the given input image in NIR spectral bands. The proposed novel approach focused on extracting the features of palm-vein and palm-print that enhances the model using the LBP approach. The scores gained by identifying the proposed and noticed techniques have shown. Finally, the scores show that the proposed approach received better results based on the checking performance. Xi et al. [12] introduced the parallel NN approach integrated with the texture features applied to facial expressions—the proposed approach developed by using the CNN, RBN, and capsule network. The components were extracted using GLCM and combined with features belonging to actual images. The accuracy reaches 98.16%, which is more improved up to 3.71% compared with existing models. Finally, the proposed approach provides a better solution for image classification.Putro et al. [13] introduced the CNN-based face detector with a tiny architecture. The proposed approach contains two types of features, unique and multimodal facial features that help predict the multiple variations. The training mainly improves the outcomes by managing the loss and twitch on training configuration. The proposed approach solves the wider face detection issues by integrating the training features with the proposed method. Finally, the proposed system executes the 53 frames per second by processing high-resolution videos. Vu et al. [14] proposed the fast palm ROI extraction technique to tackle the complex issues. The proposed approach follows the contactless platform by using the free hand posture that finds the problems and increases the accuracy for ROI. Finally, the proposed method obtained high accuracy based on the extraction of palm ROI. Wu et al. [15] proposed a sophisticated denoisingResNet model, which combines the wavelet denoising (WD) and squeeze-and-excitation ResNet18 (SER) approaches. Skin noise and optical discoloration were removed from palm vein images using the WD method. The WD approach uses residual learning technology to improve the tiny-frequency feature into the DL feature. Finally, the suggested approach had been verified via a series of investigations. Sun et al. [16] introduced the plan view detection algorithm based on NPE and KELM. The proposed approach is integrated with preprocessing, feature extraction, and dimensionality reduction based on classification and recognition. Finally, the results show better performance compared with existing systems. Li et al. [17] introduced a novel prevention system called VeinGuard. The novel approach combines DL algorithms, designed using the local-GAN, which prevent adversarial palm-vein image attacks. This approach contains the input images from different attacks that show the malicious attacks and reduce the computation time. Finally, the proposed system offers better accuracy based on adversarial attacks. Qin et al. [18] introduced the SSPP called the PVI approach, which contains multi-stage and multi-direction AGAN. An advanced data augmentation is used to find the patches in one image. The proposed approach includes the total conv GAN layers that distribute the input in different directions and detect the PVI. The proposed shows the accuracy, which is high compared with existing approaches. Yang et al. [19] introduced the low-rank initialization that extracts significant data from finger vein images. Various image-based fluctuations weaken the correlation among the original pictures and destroy the low-rank initialization. Finally, training aids in redesigning the low-rank coefficient to improve finger vein recognition performance. Qin et al. [20] introduced the iterative DBN that helps extract the vein features from the label data, and it is dynamically generated by using limited time to modify the DBN. The proposed system is integrated with several steps, like segmentation of input images by identifying the pixels of the image. The redesigned training model is used to predict the chances of the existing approach. The proposed method achieved accurate hand-vein recognition, showing better accuracy.Yang et al. [21] proposed a novel FV-GAN approach that solves various issues, such as extracting the accurate FV patterns with low-contrast IR finger images with less awareness. DL models such as CNN and others show a massive response on detection FV, but due to the more significant number of layers and size of the finger image, it takes more processing time. CNN has the drawback of showing less feature initialization because of low-quality FV image that contains eccentric and vessel breaks. To overcome all these issues, a novel approach, FV-GAN, offers high performance in reducing processing time. Das et al. [22] proposed the CNN-based FV identification system that analyzes the functionalities of the developed network from four publicly available networks. The proposed approach aims to provide constant and high accuracy when implemented with various qualities. The evaluation results show that the proposed CNN model obtained a high accuracy above 95.1%, which is high compared with existing models. Pan et al. [23] proposed the advanced DCNN approach that helps detect the Multi-Scale Deep Representation Aggregation (MSDRA) model using a pre-trained model DCNN to see finger veins. The proposed approach combines various techniques, such as detecting the multi-stage features with classification using SVM. The proposed method uses two datasets for the evaluation of results. Finally, the MSDRA shows better accuracy based on finger vein recognition.Song et al. [24] introduced the CNN model that contains two methods: difference image and measuring the distance among the feature vector obtained from the CNN. The difference between images is vulnerable to noise, and differences in pixel values create them. The proposed approach solves the issues developed by DenseNet, which shows massive performance in

terms of accuracy. To address issues with massive noise, Shen et al. [25] introduced a lightweight CNN model for finger vein image recognition and matching, obtaining the ROI and finger vein pattern feature. The proposed system solves the problems identified based on the lack of accuracy and more computation time. The proposed system received 99.4% and 99.45% accuracy, which is better than existing models. Hou et al. [26] proposed a new loss technique that learns the inter-class and intra-class data simultaneously by using the CNN for finger vein verification. Finally, the proposed approach obtained the efficiency and effectiveness of the proposed method. Kuzu et al. [27] introduced advanced finger vein pattern detection that captures accurate patterns using the CNN. The dataset is from videos with different times from 101 subjects. The proposed approach obtained an accuracy of 99.34%, better than existing approaches. Huang et al. [28] proposed the observation technique called joint attention that enables vigorous adaptation and data accumulation to extract the fine-grained details, which improves the finger vein patterns to obtain the identity features. The dimensionality reduction removes the feature maps and outcomes with highly significant features. Finally, the proposed approach, JAFVNet, shows effective performance in terms of accuracy. Yang et al. [29] developed a new biometric security system based on finger vein biometrics. The proposed BDD-ML-ELM looks at safeguarding the initial finger-vein structure even if its altered revision is affected; ultimately, the BDD-ML-ELM enhanced attack detection reliability over the current techniques.Yang et al. [30] presented the FVRAS-Net, a significant CNN approach using an MTL model to achieve high security and real-time performance. Finally, the proposed method achieves accuracy in terms of Finger-Vein Recognition rate.

III. METHODOLOGY

This section explains the methodology of this paper. The proposed approach is an integrated approach that recognizes the palm vein by using integrated learning approach (ILA). ILA contains various methods that help in fine tune the results for better outcomes. A pre-trained model EfficientNetB1is used as training model as a first step. Second, various preprocessing or noise filter approaches are used to remove the noise from input images. A feature extraction method feature-level fusion combined with several methods obtains the accurate features from the palm vein dataset. Finally the transfer learning is used to combine with LightGBM gives the accurate output for the development of Contactless Palm Vein Recognition System (CPVR) which is shown in Fig. 1. Fig. 1 also represents the methods that are used in this work.

A. Pre-trained Model for Contactless Palm Vein Recognition System (CPVR)

EfficientNet is a DL-based pre-trained model released in 2019. EfficientNetB1 is a neural network architecture developed in 2019 by Google researchers. These architectures are known for being efficient in terms of computational resources and model size, making them suitable for a wide range of computer vision tasks. The EfficientNet family is based on a concept known as compound scaling, which involves simultaneously scaling the network's depth, width, and resolution. It allows for a balance between model capacity and computational cost. EfficientNetB1 is one of the models in this family, with fewer computational resources than more significant variants like B2, B3, and so on. To reduce computational complexity while maintaining representation power, it employs a variety of efficient building blocks, including depthwise separable convolutions. This paper employs the EfficientNetB1 as a pre-trained model and trains the Palm Vein using datasets such as polyu 2D images.

The following key factors of EfficientNetB1 are given below:

B. Width Scaling Factor (ϕ)

The width scaling factor, often denoted as phi (ϕ), determines the width of the network and affects the number of channels in each layer. It's typically chosen from a set of predefined values. For EfficientNetB1, ϕ is set to 1.0.

C. Depth Scaling Factor (α)

The depth scaling factor, alpha (α), determines the number of layers in the network. It's also chosen from a predefined set. For EfficientNetB1, alpha is set to 1.1.

D. Resolution Scaling Factor (Resolution)

The resolution scaling factor is used to adjust the input image resolution. Given that the base resolution is 224x224 for EfficientNetB1, you can calculate the new resolution using this formula:

$$\text{New Resoulution} = 224 * \text{resolution} \quad (1)$$

where "resolution" is a value greater than 1, which increases the input resolution, or less than 1, which decreases it.

E. Number of Layers (N)

In ever block the no of layers of EfficientNetB1 can be calculated using the depth scaling factor alpha (α) as follows:

$$N = 1.33^{\propto} \quad (2)$$

where$\propto$ represents exponentiation.

F. Number of Channels in Each Layer (C)

In ever block the no of layers of EfficientNetB1 can be calculated using the width scaling factor phi (ϕ) as follows:

$$C = 32 * \text{round}(\phi * 1.0) \quad (3)$$

G. Number of Blocks in Each Stage (S)

The number of blocks in each stage is a design choice, and for EfficientNetB1, it typically consists of 4, 6, or 8 blocks in each stage. This value can vary depending on specific implementations.

H. Preprocessing Technique

Preprocessing is a critical step in removing noise from input PVR images. Noise removal is an important step in image processing that improves image quality by removing unwanted artifacts or distortions caused by various sources of noise. To reduce noise in the palm vein image, median filtering is used. Uneven illumination or sensor artifacts can both cause noise. The median filter is effective at removing salt-and-pepper noise, which causes some pixels to randomly

change to maximum or minimum values. It replaces each pixel with the median value of a specified neighborhood. Another method for removing noise is the Wavelet transform, which divides an image into multiple scales and allows for separate processing of different frequency components. Wavelet denoising techniques use threshold to remove noise while preserving image features. The Gabor Filter is also used to enhances the texture features can highlight the vein patterns.

The median filter for every input is represented as:

I. Median Filter

Let $I(a, b)$ initializes the actual image.

Let $I_{filtered}(a, b)$ represents the final output image.

1) Operation of median filter

- The median filter operation is carried out by dragging a window of a given size (MN) across the entire image.
- The median filter calculates the pixel values within the window centered at (a, b) for each pixel in the input image at coordinates (a, b).
- The corresponding pixel in the output (filtered) image at the exact coordinates (a,b) is then assigned the median value.

2) Median calculation: Sort the pixel values in the window in ascending order and choose the middle value to find the median value in a pixel's neighborhood. If the window size is even, the average of the two central values will be used. The median calculation for a 33 is as follows:

$$I_{filtered}(a, b) = \text{median}(I(a-1)(b-1), I(a, b-1), I(x+1, y-1), I(a-1, b), I(a, b), I(a), I(b) \quad (4)$$

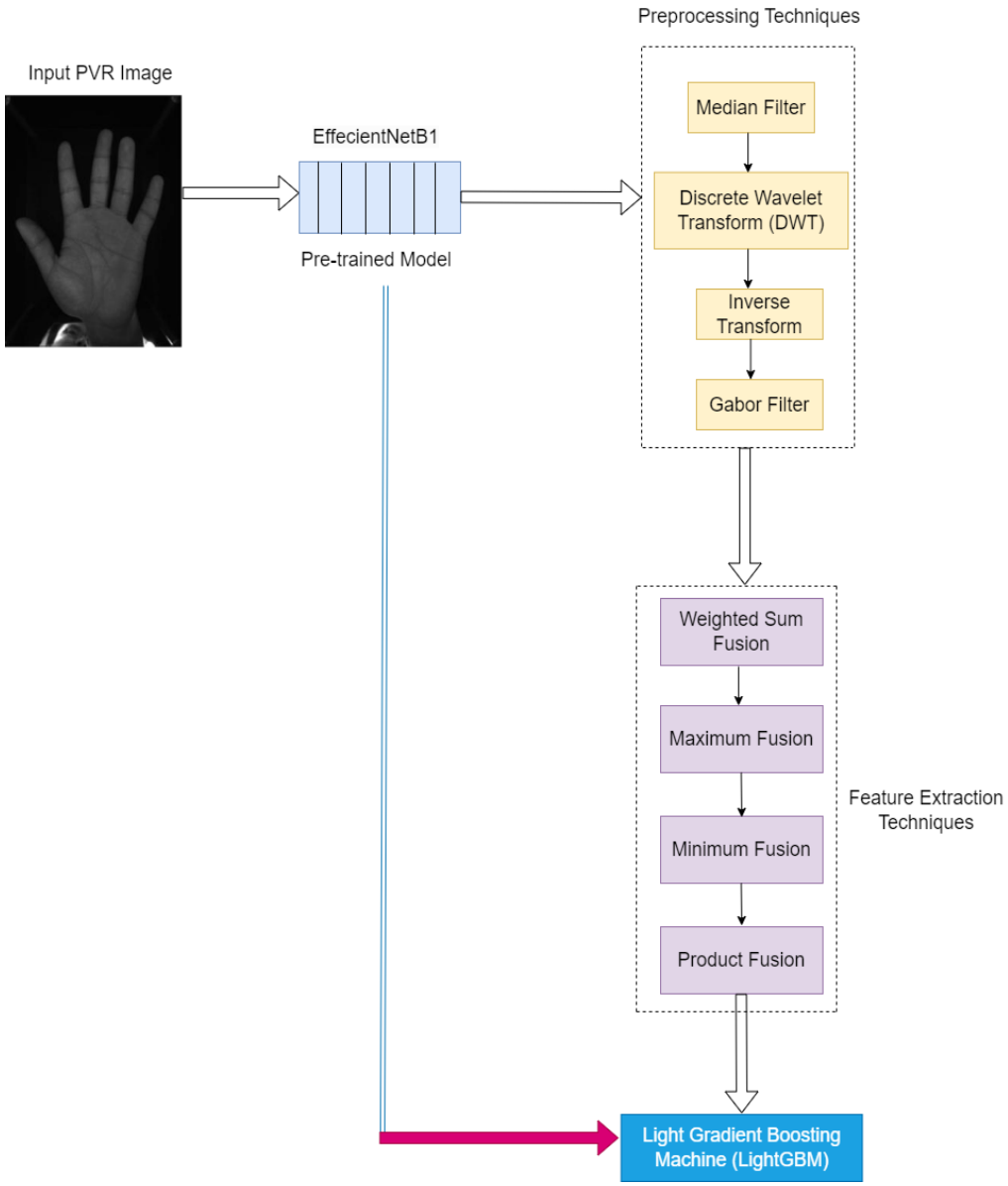


Fig. 1. Architecture diagram.

IV. DISCRETE WAVELET TRANSFORM (DWT)

The DWT is a mathematical technique for analyzing and decomposing two-dimensional images into various frequency components. A signal is divided into two parts in the DWT: approximation (low-frequency components) and detail (high-frequency components). A low-pass filter (LPF) and a high-pass filter (HPF) are used to achieve this decomposition. These filters are also known as analysis filters. The DWT equations are written as follows:

In the first step, the decomposition of input signals represented asx[n]:

Apply the LPF to achieve approximation coefficients (cX):

$$cX[j] = \sum_k x[k] * ho[k - 2j] \quad (5)$$

where, h0[k] is the impulse response of the LPF.

Apply the HPF to achieve the detail coefficients (cY):

$$cY[j] = \sum_k x[k] * h1[k - 2j] \quad (6)$$

where, h1[k] is the impulse response of the HPF.

Here, j represents the scale or level of the wavelet transform, and cX[j] and cY[j] are the estimation and specific coefficients at scale j, respectively.

A. Inverse Transform

The synthesis filters and formulas can reconstruct the actual signal from the estimation and specific coefficients. Synthesis filters are frequently the "dual" filters of analysis filters.

Apply the LPF to achieve approximation coefficients (cX) to get the approximation at the previous scale ($cX[j-1]$):

$$cX[j-1] = \sum_k cX[k] * g0[2(j-1) - k] \quad (7)$$

where,$g0[j]$is the impulse response of the syntheis LPF.

Apply the LPF to achieve approximation coefficients (cY) to get the approximation at the previous scale ($cY[j-1]$):

$$cY[j-1] = \sum_k cY[k] * g1[2(j-1) - k] \quad (8)$$

where,$1[k]$is the impulse response of the syntheis HPF.

Iteratively repeat these reconstruction steps, beginning with the finest scale (j = J) and ending with the coarsest scale (j = 1) until you obtain the reconstructed signal x[n].

B. Gabor Filter (GF)

It is one of the significant filters that used for tasks such as Palm Vein Recognition (PVR). These filters capture an image's texture and spatial frequency characteristics. Gabor filters can be used in Palm Vein Recognition to extract features unique to the palm veins. The equations and formulas for Gabor filters and their application in Palm Vein Recognition are as follows:

The 2D GF is defined as:

$$G(x, y; \lambda, \theta, \sigma, \gamma = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\lambda} + \theta\right) \quad (9)$$

where:

x and y are the coordinates of the pixel in the image.

$\lambda$is the wavelength of the sinusoidal component of the filter.

θRepresentsinclination of the GF.

σRepresents standard deviation.

$\gamma$is the spatial ratio, controlling the ellipticity of the filter.

$x' = x\cos(\theta) + y\sin(\theta)$and $y' = -x\sin(\theta) + y\cos(\theta)$ are the rotated coordinates.

C. Gabor Filter Bank:

A GF bank is created by varying the values of and capturing information at multiple scales and orientations. The filter responses for each parameter combination are computed for the input image.

V. FEATURE EXTRACTION TECHNIQUE FOR CPVR USING FEATURE-LEVEL FUSION

In Palm vein Recognition (PVR), feature-level fusion combines multiple palm vein features extracted from different sources or sensors to improve overall recognition accuracy. Various mathematical operations are used to achieve the fusion process. The following fusion methods used to extract the features from input images:

Weighted Sum Fusion: This method assigns weights to each feature source and combines them linearly.

$$\text{Fused Feature} = \sum(W_i * \text{Feature}_i) \quad (10)$$

Maximum Fusion: This method selects the maximum value for each feature dimension across all sources.

$$\text{FusedFeature}[j] = \max\left(\text{Feature}_{1[j]}, \text{Feature}_{2[j]}, \ldots, \text{Feature}_{N[j]}\right) \quad (11)$$

Minimum Fusion: Similar to Maximum Fusion, but it selects the minimum value for each feature dimension.

$$\text{FusedFeature}[j] = \min\left(\text{Feature}_{1[j]}, \text{Feature}_{2[j]}, \ldots, \text{Feature}_{N[j]}\right) \quad (12)$$

Product Fusion: This method multiplies the feature vectors element-wise from different sources.

$$\text{FusedFeature}[j] = \text{Feature}_{1[j]}, \text{Feature}_{2[j]}, \ldots, \text{Feature}_{N[j]} \quad (13)$$

where, N is the number of feature sources.

A. Light Gradient Boosting Machine (LightGBM) with Transfer Learning

LightGBM is a well-known gradient boosting framework that can be used for various machine learning tasks such as Palm Vein Recognition. LightGBM is a gradient-boosting framework that predicts using an ensemble of decision trees. It is intended to be fast, memory efficient, and scalable. In machine learning, transfer learning is typically defined as using knowledge gained from one task or dataset to improve performance on a related job or dataset. On the other hand, transfer learning is not commonly applied to gradient-boosting algorithms like LightGBM, which are primarily used for

tabular data and need a simple mechanism for incorporating knowledge from other domains or tasks. The following steps help to provide better outcomes for CPVR.

Step 1: Loss Function: For regression and classification tasks, LightGBM employs a variety of objective functions. During training, these functions are optimized.

$$\text{For Regression: } L(y, F) = \frac{1}{2} * \sum (y_i - F_i)^2 \quad (14)$$

Step 2: For binary classification, the logistic loss is commonly used

$$\text{For Regression: } L(y, p) = \sum \log\left(1 + e^{(-pi)}\right) + n\,(1 - y_i) * \log(1 + e^{pi})) \quad (15)$$

L is the loss function.

y is the true label.

F or p is the predicted output of the model.

Step 3: Gradient and Hessian Calculation

LightGBM computes the gradients and Hessians of the loss function concerning the predicted values. These gradients and Hessians determine the magnitude and direction of the model's parameter updates (trees). The gradients and Hessians are unique to the loss function.

Scalars function $f(x)$ where x is a vector of variables.

Gradient (∇f): The gradient of f is a vector that consists of partial derivatives of f with respect to every variable in x, it is initialized as $\frac{\partial f}{\partial x}$. Each component of the gradient is measured as follows:

$$\frac{\partial f}{\partial x_i} \quad (16)$$

where, x_i is the i-th variable in the vector x. Thus, the full gradient vector is:

$$\nabla f = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right] \quad (17)$$

Hessian $(\nabla^2 f)$: The Hessian matrix is a matrix of squares containing the derivatives that are partial to f about each variable in x. It is denoted as $\frac{\partial^2 f}{\partial x^2}$. the elements of the Hessian matrix are computed as follows:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} \quad (18)$$

where, x_i and x_j are variables in the vector x. So, the full Hessian matrix is an n×n matrix, where n is the number of variables, and its elements are computed for all combinations of i and j.

Step 4: Tree Building: LightGBM builds trees in a leaf-wise manner. The algorithm selects the leaf node that results in the maximum reduction in the loss function. The leaf-wise growth strategy is different from traditional depth-first or level-wise strategies used in other gradient boosting algorithms.

Step 5: Leaf Value Calculation: When a tree node is split, LightGBM calculates the value assigned to each leaf node. This calculation aims to optimize the loss function, taking into account the gradients and Hessians. The optimal leaf values are used to update the predictions.

Step 6: Shrinkage (Learning Rate): LightGBM typically uses a shrinkage parameter (learning rate) to control the step size of updates during the optimization process. This parameter prevents the model from making large adjustments in each iteration.

B. Dataset Description

The CASIA multi-spectral palm print image database was compiled from online sources and includes 3600 samples for training and 3600 samples for testing collected from 100 people [31]. Here six types of palm print images are shown in Fig. 2.

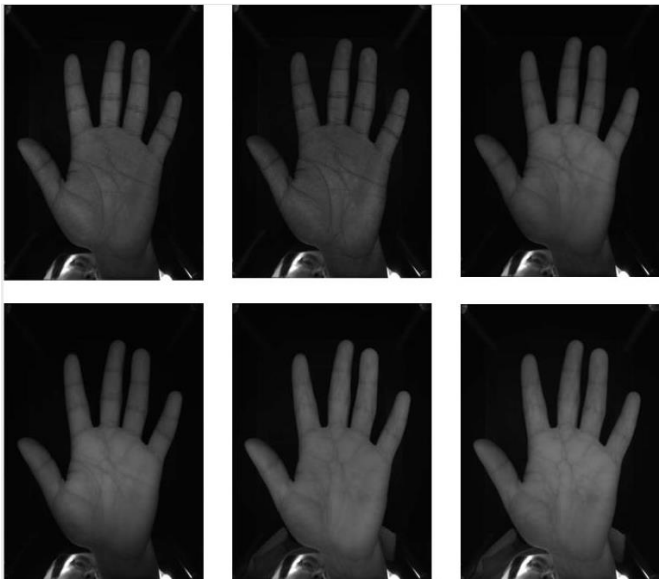


Fig. 2. Six types of palm vein images.

C. Performance Metrics

False Acceptance Rate (FAR): FAR calculates the likelihood of the system misidentifying an unauthorized user as authorized. A lower FAR denotes greater security.

$$\text{FAR} = \frac{\text{No of False Acceptances}}{\text{Total No of Identification Attempts}} \times 100\% \quad (19)$$

where:

"Number of False Acceptances" represents total no of incorrectly accepts an imposter.

"Total Number of Identification Attempts" is the total number of authentication or identification attempts, including both genuine and impostor attempts.

False Rejection Rate (FRR): The FRR calculates the likelihood of the system not accepting an authorized user. A lower FRR is preferable for user ease.

$$\text{FRR} = \frac{\text{No of False Rejections}}{\text{Total No of original Attempts}} \times 100\% \quad (20)$$

"Number of False Rejections" refers to the instances where the system incorrectly rejects a valid input or user.

"Total Number of Genuine Attempts" is the total number of times the system was presented with valid inputs or users.

Equal Error Rate (EER): The EER is a case where FAR and FRR are equivalent. Lower EER values suggest better system efficiency overall.

$$EER = \frac{(FAR+FRR)}{2} \quad (21)$$

D. Experimental Results

This section focused on evaluation results of the LightBGM with transfer learning of PVRC. Python is powerful programming language that provides the better libraries to implement the proposed algorithm. An advanced hardware requirements like 32 GB RAM and I7 processor is needed to the system to execute the large PVR image dataset. The performance of pre-trained models such as EfficientNetB1 compared with several pre-trained models achieved the better results. The training and testing loss of EffecientNetB1 is explained in Fig. 3 and Fig. 4.

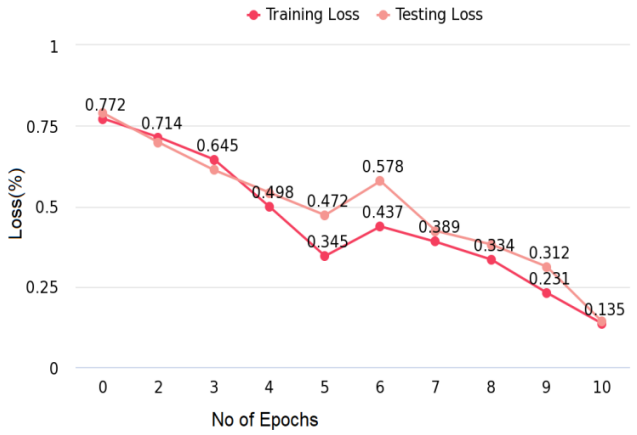


Fig. 3. The performance of EffecientNetB1 in terms of training and testing loss.

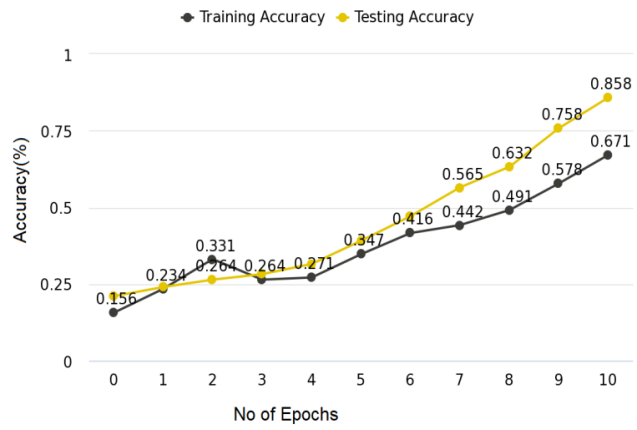


Fig. 4. The performance of EffecientNetB1 in terms of training and testing accuracy.

Fig. 3 shows the training and testing loss of EffecientB1 pre-trained model. Totally for 10 Epochs the loss is about 0.135% for 1st Epoch. There is Epoch iteration is starts from 10 Epoch to zeroEpoch. Here the training loss is low with 0.772% represents the lowest error rate. The training loss is about 0.743% which is low compare with testing loss and it prevents the over fitting. Fig. 4 shows the training and testing accuracy the training accuracy is about 0.671% and testing accuracy is about 0.858%.

Table I shows the comparative performance of various existing and proposed pre-trained models based on the FAR parameters. From the comparison it is shown that the proposed model shows the better outcomes compare with existing models. Table II shows the performance of several pre-trained models compared with the proposed model and obtained the outcomes based on the given parameter FRR. The FRR(%) shows very low compare with existing models that represents the better performance. Finally, these models help the proposed approach to show significant outcomes.

Table III shows the comparative outcomes of various pre-trained models for analyzing the performance in terms of ERR. Among all the models the proposed model shows the low error rate with high performance. Table IV shows the high performance in terms of all the parameters. Fig. 5 shows the overall performances of list of algorithms for CPVR.

TABLE I. PERFORMANCE OF PRE-TRAINED MODELS IN TERMS OF FAR

Algorithms		Year	FAR (%)
MobileNet_v3	Howard et al.[32]	2019	7.56
GhostNet	Han et al.[33]	2020	6.78
RESNET	Zhang et al.[34]	2020	5.97
EffecientNetB1	Ours	-	4.54

TABLE II. PERFORMANCE OF PRE-TRAINED MODELS IN TERMS OF FRR

Algorithms		Year	FRR (%)
MobileNet_v3	Howard et al.[32]	2019	6.23
GhostNet	Han et al.[33]	2020	5.23
RESNET	Zhang et al.[34]	2020	5.97
EffecientNetB1	Ours	-	3.76

TABLE III. PERFORMANCE OF PRE-TRAINED MODELS IN TERMS OF ERR

Algorithms		Year	ERR (%)
MobileNet_v3	Howard et al.[32]	2019	6.895
GhostNet	Han et al.[33]	2020	6.55
RESNET	Zhang et al.[34]	2020	5.97
EffecientNetB1	Ours	-	4.15

TABLE IV. THE OVERALL PERFORMANCES OF LIST OF ALGORITHMS WITH NOISE FILTERS

Algorithms		Year	EER (%)	FRR (%)	ERR (%)
PCANet with DL	Meraoumia et al.[35]	2021	0.949	1.789	1.567
CNN with Auto Encoder	Thapar et al. [36]	2017	3.71	2.987	2.123
CNN+ Bayesian Optimization + Jerman Filter	Obayya et al.[37]	2020	0.0683	0.0765	1.765
EffecientNetB1+ LightBGM with Transfer Learning	Ours	-	0.0541	0.0345	0.0235

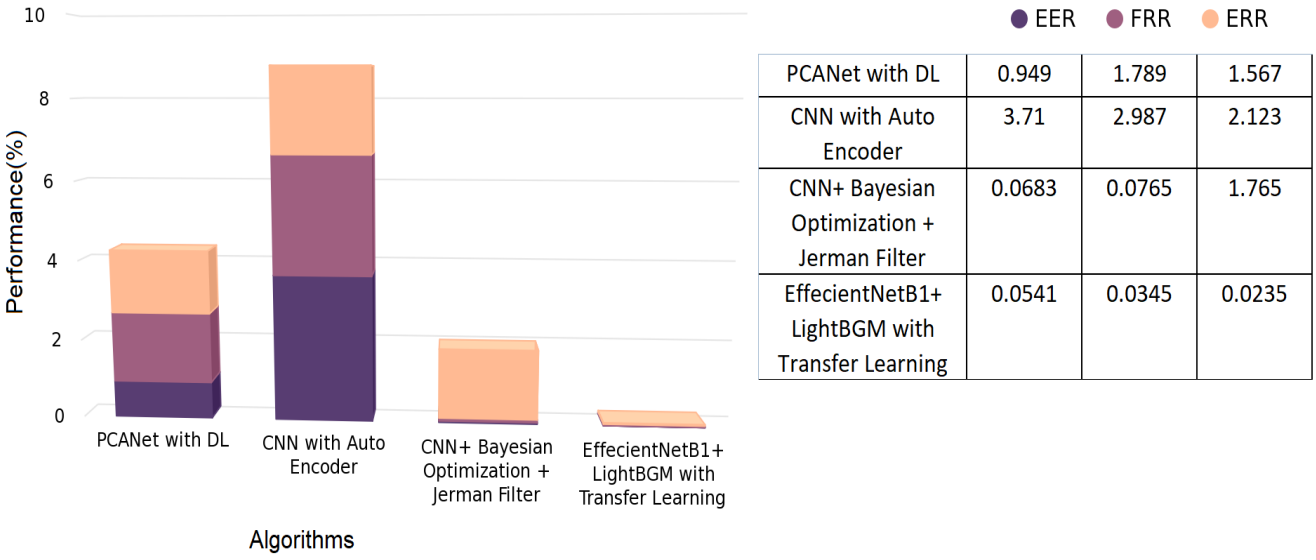


Fig. 5. The overall performances of list of algorithms for CPVR.

VI. CONCLUSION

The Integrated Learning Approach (ILA) for CPVR was introduced in this paper as a promising advancement in biometric security technology. Palm Vein Recognition is inherently secure because it relies on an individual's palm's unique vascular patterns. Contactless systems, in particular, reduce contamination risk while improving user convenience. The ILA combines several algorithms and methods to improve the system's accuracy and reliability. It ensures the system can produce consistent results even in various environmental conditions. Contactless systems are user-friendly and non-intrusive because they do not require physical contact. It can boost user acceptance and make the system suitable for various applications. Palm Vein Recognition is generally fast, allowing for quick and efficient access control or authentication. ILA improves this speed even further by optimizing the recognition process. Finally, ILA holds great promise in enhancing security, accuracy, and user experience across various applications. Its combination of contactless technology and advanced algorithms makes it an appealing option for businesses looking for dependable and secure biometric authentication solutions. However, as with any technology, factors such as implementation, user education, and ongoing maintenance must be considered to maximize the benefits of such systems while ensuring user privacy and data protection.

REFERENCES

[1] W. Kang, H. Liu, W. Luo and F. Deng, "Study of a full-view 3D finger vein verification technique", IEEE Trans. Inf. Forensics Security, vol. 15, pp. 1175-1189, 2019.

[2] B. Nakisa, F. Ansarizadeh, P. Oommen and S. Shrestha, "Technology Acceptance Model: A Case Study of Palm Vein Authentication Technology," in IEEE Access, vol. 10, pp. 120436-120449, 2022, doi: 10.1109/ACCESS.2022.3221413.

[3] S. Zhao and B. Zhang, "Learning salient and discriminative descriptor for palmprint feature extraction and identification", IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 12, pp. 5219-5230, Dec. 2020.

[4] Wu, W., Elliott, S.J., Lin, S., Sun, S. and Tang, Y. (2020), Review of palm vein recognition. IET Biom., 9: 1-10.

[5] O. Toygar, F. O. Babalola and Y. Bitirim, "FYO: A novel multimodal vein database with palmar dorsal and wrist biometrics", IEEE Access, vol. 8, pp. 82461-82470, 2020.

[6] Z. Pan, J. Wang, Z. Shen, X. Chen and M. Li, "Multi-layer convolutional features concatenation with semantic feature selector for vein recognition", IEEE Access, vol. 7, pp. 90608-90619, 2019.

[7] G. Wang, C. Sun and A. Sowmya, "Multi-weighted co-occurrence descriptor encoding for vein recognition", IEEE Trans. Inf. Forensics Security, vol. 15, pp. 375-390, 2020.

[8] H. Qin, M. A. El Yacoubi, J. Lin and B. Liu, "An iterative deep neural network for hand-vein verification", IEEE Access, vol. 7, pp. 34823-34837, 2019.

[9] R. S. Kuzu, E. Maiorana and P. Campisi, "Vein-based biometric verification using densely-connected convolutional autoencoder", IEEE Signal Process. Lett., vol. 27, pp. 1869-1873, 2020.

[10] S. -J. Horng, D. -T. Vu, T. -V. Nguyen, W. Zhou and C. -T. Lin, "Recognizing Palm Vein in Smartphones Using RGB Images," in IEEE

Transactions on Industrial Informatics, vol. 18, no. 9, pp. 5992-6002, Sept. 2022, doi: 10.1109/TII.2021.3134016.

[11] S. Cho, B. Oh, K. Toh and Z. Lin, "Extraction and cross-matching of palm vein and palmprint from the RGB and the NIR spectrums for identity verification", IEEE Access, vol. 8, pp. 4005-4021, 2020.

[12] Z. Xi, Y. Niu, J. Chen, X. Kan and H. Liu, "Facial expression recognition of industrial internet of things by parallel neural networks combining texture feature", IEEE Trans. Ind. Informat., vol. 17, no. 4, pp. 2784-2793, Apr. 2021.

[13] M. D. Putro, L. Kurnianggoro and K.-H. Jo, "High performance and efficient real-time face detector on central processing unit based on convolutional neural network", IEEE Trans. Ind. Informat., vol. 17, no. 7, pp. 4449-4457, Jul. 2021.

[14] D.-T. Vu, T.-V. Nguyen and S.-J. Horng, "Rotation-invariant palm ROI extraction for contactless recognition" in Advances in Computer Vision and Computational Biology, Cham, Germany::Springer, pp. 701-715, 2021.

[15] [15] W. Wu et al., "Outside Box and Contactless Palm Vein Recognition Based on a Wavelet Denoising ResNet," in IEEE Access, vol. 9, pp. 82471-82484, 2021, doi: 10.1109/ACCESS.2021.3086811.

[16] S. Sun, X. Cong, P. Zhang, B. Sun and X. Guo, "Palm Vein Recognition Based on NPE and KELM," in IEEE Access, vol. 9, pp. 71778-71783, 2021, doi: 10.1109/ACCESS.2021.3079458.

[17] Y. Li, S. Ruan, H. Qin, S. Deng and M. A. El-Yacoubi, "Transformer Based Defense GAN Against Palm-Vein Adversarial Attacks," in IEEE Transactions on Information Forensics and Security, vol. 18, pp. 1509-1523, 2023, doi: 10.1109/TIFS.2023.3243782.

[18] H. Qin, M. A. El-Yacoubi, Y. Li and C. Liu, "Multi-scale and multi-direction GAN for CNN-based single palm-vein identification", IEEE Trans. Inf. Forensics Security, vol. 16, pp. 2652-2666, 2021.

[19] L. Yang, G. Yang, K. Wang, F. Hao and Y. Yin, "Finger vein recognition via sparse reconstruction error constrained low-rank representation", IEEE Trans. Inf. Forensics Security, vol. 16, pp. 4869-4881, 2021.

[20] H. Qin, M. A. E. Yacoubi, J. Lin and B. Liu, "An iterative deep neural network for hand-vein verification", IEEE Access, vol. 7, pp. 34823-34837, 2019.

[21] W. Yang, C. Hui, Z. Chen, J.-H. Xue and Q. Liao, "FV-GAN: Finger vein representation using generative adversarial networks", IEEE Trans. Inf. Forensics Security, vol. 14, no. 9, pp. 2512-2524, Sep. 2019.

[22] R. Das, E. Piciucco, E. Maiorana and P. Campisi, "Convolutional neural network for finger-vein-based biometric identification", IEEE Trans. Inf. Forensics Security, vol. 14, no. 2, pp. 360-373, Jun. 2019.

[23] Z. Pan, J. Wang, G. Wang and J. Zhu, "Multi-scale deep representation aggregation for vein recognition", IEEE Trans. Inf. Forensics Security, vol. 16, pp. 1-15, 2021.

[24] J. M. Song, W. Kim and K. R. Park, "Finger-vein recognition based on deep DenseNet using composite image", IEEE Access, vol. 7, pp. 66845-66863, 2019.

[25] J. Shen et al., "Finger vein recognition algorithm based on lightweight deep convolutional neural network", IEEE Trans. Instrum. Meas., vol. 71, pp. 1-13, 2022.

[26] B. Hou and R. Yan, "Arcvein-arccosine center loss for finger vein verification", IEEE Trans. Instrum. Meas., vol. 70, pp. 1-11, 2021.

[27] R. S. Kuzu, E. Piciucco, E. Maiorana and P. Campisi, "On-the-fly finger-vein-based biometric recognition using deep neural networks", IEEE Trans. Inf. Forensics Security, vol. 15, pp. 2641-2654, 2020.

[28] J. Huang, M. Tu, W. Yang and W. Kang, "Joint attention network for finger vein authentication", IEEE Trans. Instrum. Meas., vol. 70, pp. 1-11, 2021.

[29] W. Yang, S. Wang, J. Hu, G. Zheng, J. Yang and C. Valli, "Securing deep learning based edge finger vein biometrics with binary decision diagram", IEEE Trans. Ind. Informat., vol. 15, no. 7, pp. 4244-4253, Jul. 2019.

[30] W. Yang, W. Luo, W. Kang, Z. Huang and Q. Wu, "FVRAS-Net: An embedded finger-vein recognition and antispoofing system using a unified CNN", IEEE Trans. Instrum. Meas., vol. 69, no. 11, pp. 8690-8701, Nov. 2020.

[31] Ying Hao Zhenan Sun,Tieniu Tan and Chao Ren, "Multi-spectral palm image fusion for accurate contact-free palmprint recognition", Proceedings of IEEE International Conference on Image Processing, 2008, pp.281-284, USA

[32] A. Howard, M. Sandler, B. Chen. W. J. Wang, L. C. Chen, M. X. Tan, G. Chu, V. Vasudevan, Y. K. Zhu, R. M. Pang, H, Adam, Q. Le. Searching for mobileNetV3. In Proceedings of IEEE/CVF International Conference on Computer Vision, IEEE, Seoul, South Korea, pp.1314–1324, 2019. DOI: 10.1109/ICCV.2019.00140.

[33] K. Han, Y. H. Wang, Q. Tian, J. Y. Guo, C. J. Xu, C. Xu. GhostNet: More features from cheap operations. [Online], Available: https://arxiv.org/abs/1911.11907, 2019.

[34] H. Zhang, C. R. Wu, Z. Y. Zhang, Y. Zhu, Z. Zhang, H. B.Lin, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, A.Smola. ResNeSt: Split-attention networks. [Online], Available: https://arxiv.org/abs/2004.08955, 2020.

[35] A. Meraoumia, F. Kadri, H. Bendjenna, S. Chitroub, and A. Bouridane, "Improving biometric identification performance using PCANet deep learning and multispectral palmprint," in Signal Processing for Security Technologies. Cham, Switzerland: Springer, 2017, pp. 51–69.

[36] D. Thapar, G. Jaswal, A. Nigam, and V. Kanhangad, "PVSNet: Palm vein authentication siamese network trained using triplet loss and adaptive hard mining by learning enforced domain specific features," in Proc. IEEE 5th Int. Conf. Identity, Secur., Behav. Anal. (ISBA), Jan. 2019, pp. 1–8, doi: 10.1109/ISBA.2019.8778623.

[37] M. I. Obayya, M. El-Ghandour and F. Alrowais, "Contactless Palm Vein Authentication Using Deep Learning With Bayesian Optimization," in IEEE Access, vol. 9, pp. 1940-1957, 2021, doi: 10.1109/ACCESS.2020.3045424.

Linear and Nonlinear Analysis of Photoplethysmogram Signals and Electrodermal Activity to Recognize Three Different Levels of Human Stress

Yan Su[1], Yuanyuan Li[2*], Shumin Zhang[3], Hui Wang[4]
Mental Health Service Center, Cangzhou Medical College, Cangzhou 061001, Hebei, China[1, 3]
Student Affairs Department, Cangzhou Medical College, Cangzhou 061001, Hebei, China[2]
Department of Health Management and Service, Cangzhou Medical College, Cangzhou 061001, Hebei, China[4]

***Abstract*—All human beings experience different levels of psychological stress during their daily activities, and stress is an integral part of human life. So far, few studies have attempted to identify different levels of stress by analyzing physiological signals. However, it should be noted that developing a practical system for detecting multiple stress levels is a challenging task, and no standard system has been developed for this purpose. Therefore, in the current study, we propose a new detection system based on linear and nonlinear analysis of photoplethysmogram (PPG) and electrodermal activity (EDA) signals to classify three levels of stress (low, medium and high). In the current study, we recorded the physiological signals of EDA and PPG during three trials of a Stroop color word test that induced three levels of stress in 42 healthy male volunteers. Mean, median, standard deviation, variance, skewness, kurtosis, minimum, maximum, and RMS features in the time domain were calculated from physiological signals as linear features. Also, approximate entropy, sample entropy, permutation entropy, Hurst exponent, Katz fractal dimension, Higuchi fractal dimension, Petrosian fractal dimension, detrended fluctuation analysis (DFA), and embedding dimension and time delay parameters from phase space reconstruction of the signals were calculated as nonlinear features. The combination of nonlinear and linear features extracted from both PPG and EDA signals resulted in the highest mean accuracy (88.36%), intraclass correlation (ICC) (98.82%) and F1 (89.24%) values in the classification of three levels of mental stress through multilayer perceptron neural network. Our findings showed that the combination of nonlinear and linear approaches for biological data analysis (PPG and EDA) could help to develop a stress detection system.**

Keywords*—Stress detection; biological signal; linear analysis; nonlinear analysis; classification

I. INTRODUCTION

All human beings experience different levels of psychological stress during their daily life activities, and stress is an integral part of human life. Stress refers to situations and feelings in which people perceive expectations to be beyond their capabilities [1]. In fact, stress can be defined as the mind or body's response to any need for change [2]. Human stress is controlled by the activation of the limbic system and the hypothalamus-pituitary-adrenal axis, which control the release of adrenaline and cortisol (stress hormones) in the bloodstream [3]. The circulation of these hormones in the human body through the bloodstream leads to different physiological variations. As a result, the heart rate begins to increase relative to the normal condition, increasing blood pumping to the muscles and various organs. Therefore, blood pressure and breathing rate increases [4]–[6]. In addition, adrenaline causes the release of stored fat and glucose into the bloodstream, preparing the body to respond to stress [7]. Furthermore, it has been shown that different areas of the human brain, such as the prefrontal cortex, play an important role in regulating various signs of the body during stress [8]. All this cumulative evidence shows that physiological systems and signals undergo changes during psychological stress.

On the other hand, it should be noted that excessive stress affects people's health. It has been introduced as a risk factor involved in the occurrence of major psychiatric diseases such as schizophrenia, depression and anxiety disorders [9]– [11]. Mental stress affects the ability for problem-solving, creativity, work memory and decision-making in humans. In addition, it can be a risk factor for various physical illnesses such as strokes, diabetes, and cardiovascular diseases [12], [13]. Therefore, determining the level of stress in different situations can help people to control it using stress reduction techniques and avoid the unpleasant consequences of excessive stress on health. Accordingly, in recent years, many pattern recognition methods have been developed to detect different emotions and their levels from biological signals [14]–[16]. Electroencephalogram (EEG), electrocardiogram (ECG), electromyogram (EMG), electrodermal activity (EDA), respiratory signal (RSP), blood volume pulse (BVP) and photoplethysmogram (PPG) are among the biological and physiological signals that have been computationally analyzed for this purpose. However, the main challenge in stress detection systems is the fact that each person is unique and shows emotions in different ways. This makes the topic of research hot. Although most studies on emotion recognition used ECG and EEG signals, we attempted to record and analyze PPG and EDA signals in the current study because they can be recorded via two-finger electrodes on the non-

dominant hand without compromising privacy and comfort. Moreover, some early studies demonstrated that PPG and EDA are good indicative tools to assess emotions.

II. Related Works

Paul Ekman was the first researcher who tried to recognize different emotions through physiological signal analysis [17]. Later, several researchers tried to continue his interesting path by analyzing different physiological signals. However, most studies focused on emotions like joy, fear, sadness, disgust, anger and surprise, and very few studies attempted to detect different levels of mental stress. Healey and Pickard induced three levels of stress (low, moderate, and high) during a driving task and analyzed ECG, EMG, RSP, and EDA signals recorded from healthy participants. They used linear frequency analysis and a linear discriminant analysis (LDA) classifier and reported a good accuracy of 97% for distinguishing three levels of human stress [18]. Shirvan et al. proposed a computational technique based on different linear and nonlinear analyses (including statistical analysis, fractal dimension analysis and detrended fluctuation analysis) of functional near-infrared spectroscopy (fNIRS) signals to detect low and high levels of stress. They used a feature selection method and support vector machine (SVM) classifier at both individual and group settings for stress levels classification and reported an accuracy of 88.72% in this regard [19]. Yannakakis and Hallam induced two levels of fun (low and high) in healthy participants through an interactive game and analyzed the recorded ECG, EDA and BVP signals by linear statistical analysis. They used SVM and Artificial Neural Networks (ANN) in the classification stage and reported 70% accuracy in recognizing two levels of fun [20]. Katsis et al. induced low stress, high stress, euphoria and disappointment in subjects through a driving task and analyzed the linear dynamics of the recorded ECG, EMG, EDA and RSP signals. In the classification stage, they used SVM and a neuro-fuzzy inference system and achieved 79.3% accuracy for the classification of the four states [21]. Valenza et al. induced multiple levels and valence and arousal in healthy volunteers through an international affective picture system and analyzed the nonlinear dynamics of the recorded ECG, EDA and RSP signals. In the classification stage, they used a quadratic discriminant classifier and achieved more than 90% accuracy in affective arousal and valence recognition [14].

As mentioned, few studies have attempted to identify different levels of stress by analyzing physiological signals. However, it should be noted that developing a practical system for detecting multiple stress levels is challenging, and no standard system has been developed for this purpose. Therefore, in the current study, we propose a new detection system based on linear and nonlinear analysis of PPG and EDA signals to classify three levels of stress (low, medium and high).

III. Materials and Methods

A total of 42 healthy male volunteers participated in the research with an average age of 26.31 ± 5.12 years. The research method was first explained to all participants, and informed consent was obtained from them before beginning the experiment. All subjects had a normal or normalized vision. A psychiatric interview was conducted by a psychiatrist to ensure the mental health of all participants to have no symptoms of major psychiatric disorders, cognitive problems, insomnia, anxiety, or social dysfunction. In addition, participants had no history of major physical illnesses, drug or alcohol abuse, and neurological disorders.

A. Stress Induction

In the current study, we utilized the Stroop color word test in a visual basic windows environment to induce three levels of stress in the participants. This test comprises three different experiments: preliminary experiment, congruent or non-conflict experiment, and non-congruent or conflict experiment. In the preliminary experiment, the color of the word appeared black. In the congruent experiment, the color of the work that appeared is similar relative to the color in the written word. In the non-congruent experiment, the color of the word appeared in different colors relative to the written word. Fig. 1 shows the Stroop color word test used in the current study for inducing stress. In each task and experiment, participants should indicate the color of the word. Each experiment lasted three minutes. In all experiments, each trial was displayed for three seconds, and participants were asked to respond to each trial using a mouse. Previous studies have shown that this test can reliably elicit three levels of stress in human subjects. Indeed, the preliminary experiment induces a low-stress level, the congruent experiment induces a medium stress level, and the non-congruent experiment induces a high-stress level [22], [23]. A 16-inch monitor was placed in front of participants at a 70-cm distance from them to perform the Stroop test.

B. Physiological Signal Acquisition

As mentioned, in the current study, we captured the EDA and PPG physiological signals during three experiments of the Stroop color word test. EDA indicates the variations in the electrical properties of the skin because of mentally induced sweat gland activities upon external stimuli. Skin resistance varies with the status of sweat glands in the skin. Sweating is controlled by the sympathetic nervous system, and thus, skin resistance is an indication of psychological arousal [24]. On the other hand, PPG is a simple optical non-invasive method to determine volumetric changes in blood in peripheral circulation that has been shown to be related to affective states [25].

In the current study, the Shimmer3 EDA+ module, along with an optical pulse sensor, was used for recording the EDA and PPG signals. This device is an extensible wireless sensor platform for recording sampled EDA data in real-time. The optical pulse sensor attached to this module also can record a PPG signal from a finger. This module digitized data at a 250 Hz sampling rate and streamed the data to a host PC in real-time. Two dry electrodes, along with the optical pulse probe, were attached to the fingers of subjects' non-dominant hands to record the EDA and PPG signals. The Shimmer3 module has shown to be an accurate and reliable wearable sensor platform for capturing physiological signals, which can be utilized for biomedical research applications [26].

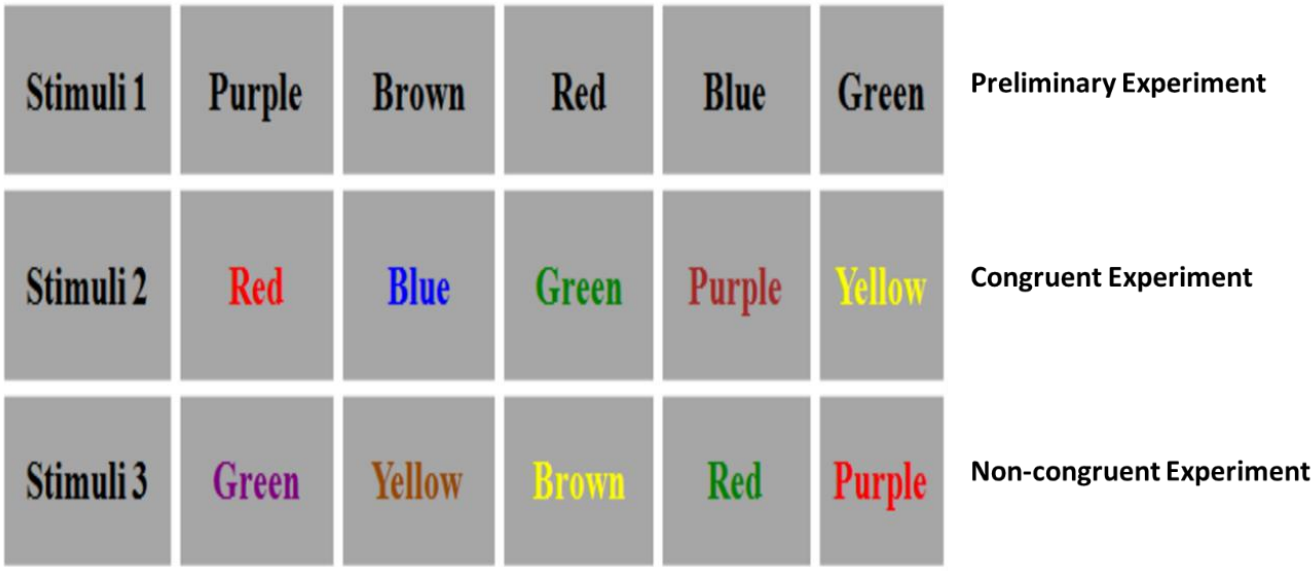


Fig. 1. The Stroop color word test was used in the current study to induce three levels of stress.

C. Linear and Nonlinear Analysis for Feature Extraction

The linear features extracted from EDA and PPG signals include mean, median, standard deviation, variance, skewness, kurtosis, minimum, maximum, and RMS in the time domain. These linear statistical features have a low computational cost which has been shown to be effective in various biomedical research applications. Mathematical definitions of these features and their details can be found in [27], [28]. On the other hand, the nonlinear features extracted from EDA and PPG signals include approximate entropy, sample entropy, permutation entropy, Hurst exponent, Katz fractal dimension, Higuchi fractal dimension, Petrosian fractal dimension, detrended fluctuation analysis (DFA), and embedding dimension and time delay parameters from phase space reconstruction of the signals (see Table I). Mathematical definitions of these features and their details can be found in [29], [30].

TABLE I. LIST OF LINEAR AND NONLINEAR FEATURES EXTRACTED FROM PPG AND EDA SIGNALS

Analysis	Extracted features from PPG and EDA signals
Linear	Mean, median, standard deviation, variance, skewness, kurtosis, minimum, maximum, RMS
Nonlinear	Approximate entropy, sample entropy, permutation entropy, Hurst exponent, Katz fractal dimension, Higuchi fractal dimension, Petrosian fractal dimension, detrended fluctuation analysis, embedding dimension, time delay

IV. RESULTS

Before feature extraction, we first applied a simple segmentation method to the recorded signals through a rectangular window with a length of 45 seconds. Considering the sampling frequency of 250 Hz, each segment contained 11250 data points. Also, each segment had a label to show the individual's stress level. All the above features were extracted from each segment, and the average values extracted for all segments with the same label were defined as the main feature in the classification step. Linear features were first extracted from PPG and EDA time series, and then, nonlinear features were estimated from the signals through the described nonlinear dynamic algorithms. Fig. 2 and Fig. 3 show examples of PPG and EDA signals recorded at three stress levels, respectively. Also, Fig. 4 depicts the histogram of time delays obtained from PPG and EDA signals.

In the classification stage, 70% of features were utilized to train a multilayer perceptron (MLP) neural network, 10% was utilized for model validation, and the remaining 20% was used to test the MLP. In the validation stage, the leave-one-subject-out approach was utilized to estimate the performance of MLP. We investigated different combinations of features and signals (i.e., EDA and PPG) to arrive at the optimal way to detect the stress level. In other words, we utilized various feature combinations for MLP modeling and assessed the classification results of each combination to obtain the best solution for this three-class classification problem. Assessment metrics utilized in the current study for evaluating different strategies were accuracy, intra-class correlation coefficient (ICC) and F1-measure.

Fig. 5 to Fig. 7 show mean classification accuracies, F1-measures and ICC values obtained for each feature combination by MLP classifier. As shown, the best accuracy of 79.5% was obtained by nonlinear features extracted from PPG signals. The best accuracy of 80.42% was obtained by combined features (i.e., linear and nonlinear features) extracted from EDA signals. In addition, the best accuracy of 88.36% was obtained by combining features extracted from PPG and EDA signals. Indeed, the combination of nonlinear and linear features extracted from both PPG and EDA signals resulted in the highest mean accuracy (88.36%), ICC (98.82%) and F1 (89.24%) values in the classification of three levels of mental stress.

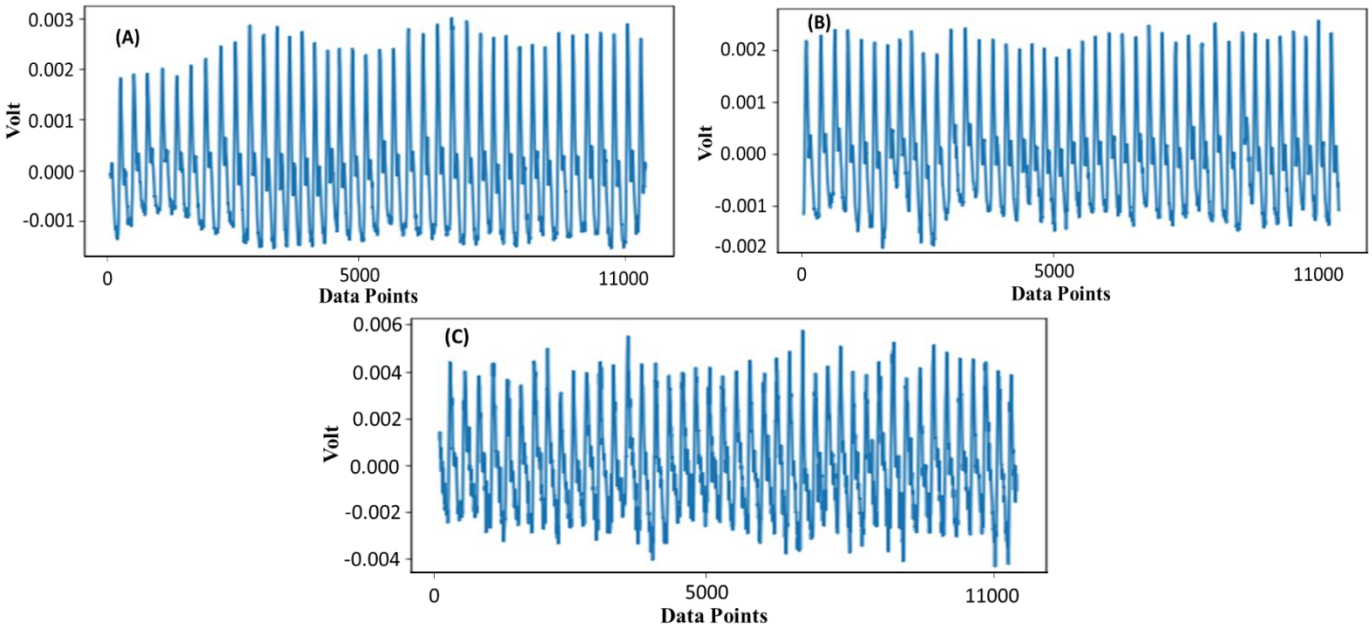


Fig. 2. Example of recorded PPG signals for (A) low-stress level, (B) medium stress level, and (C) high-stress level.

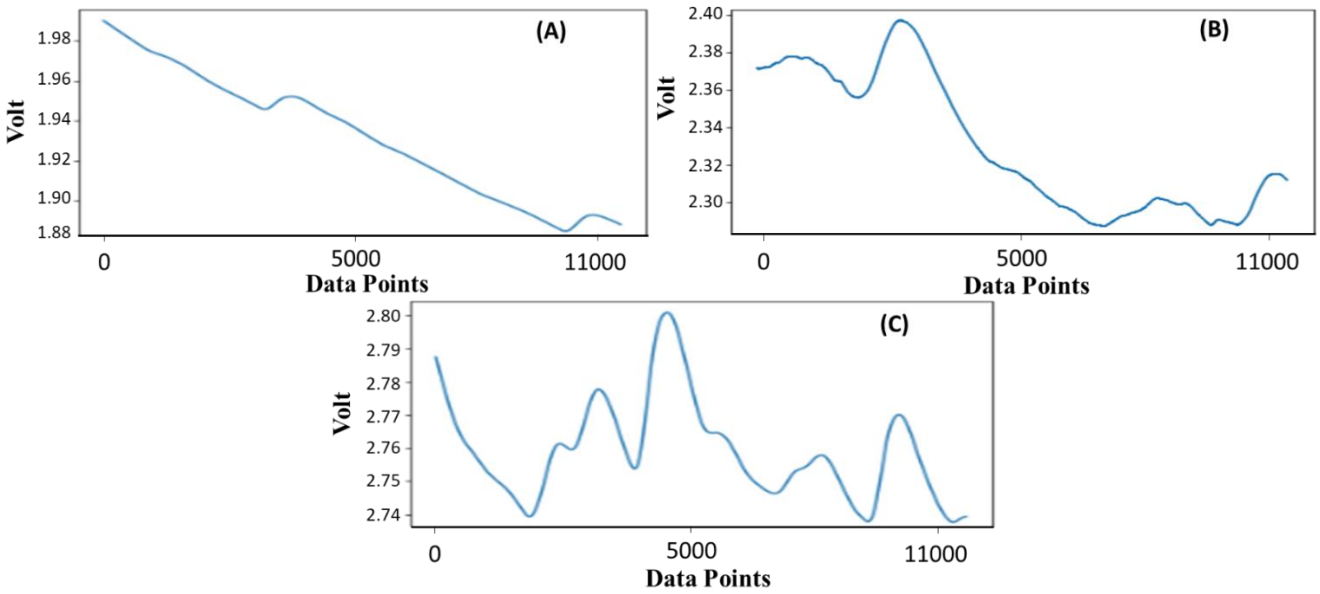


Fig. 3. Example of recorded EDA signals for (A) low-stress level, (B) medium stress level, and (C) high-stress level.

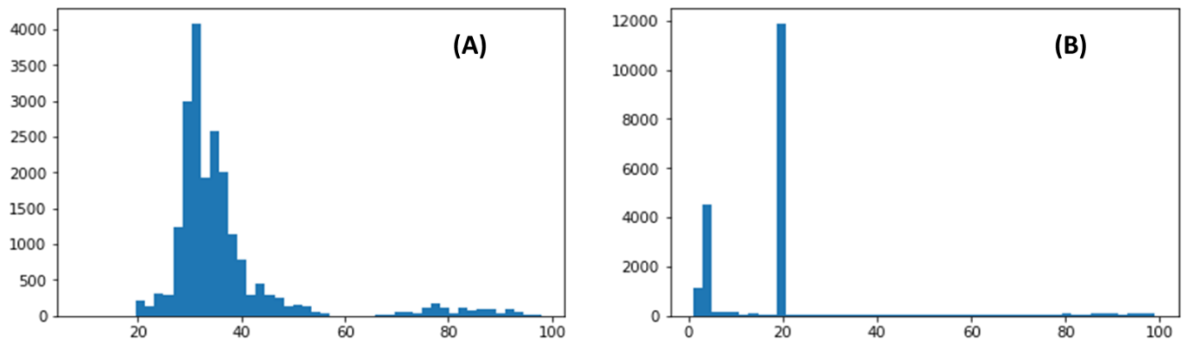


Fig. 4. Histogram of time delays obtained from (A) PPG and (b) EDA signals.

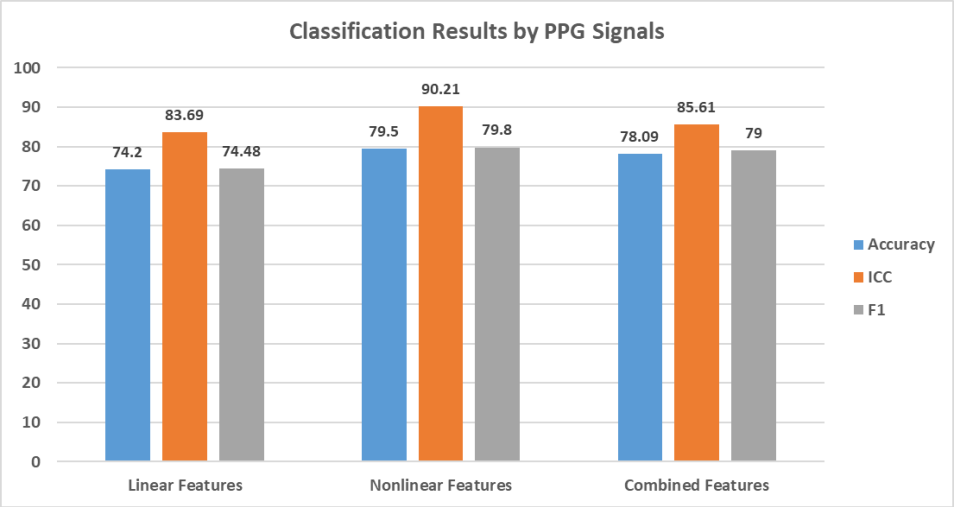


Fig. 5. Averaged classification results were obtained for different features extracted from PPG signals.

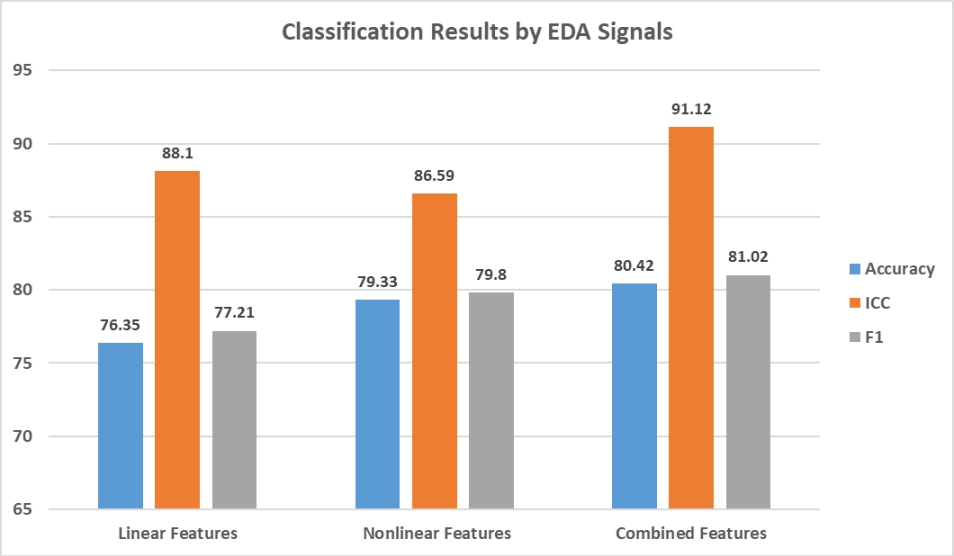


Fig. 6. Averaged classification results were obtained for different features extracted from EDA signals.

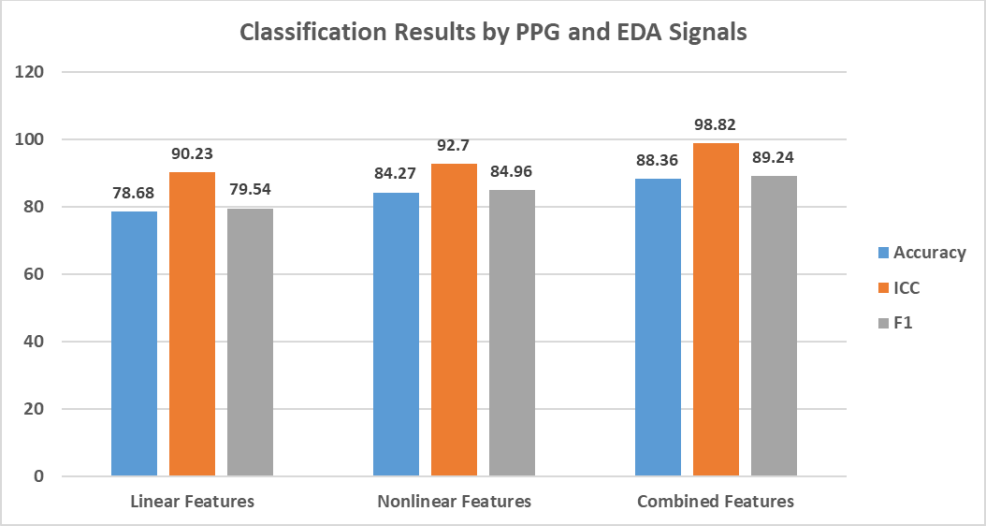


Fig. 7. Averaged classification results were obtained for different features extracted from PPG and EDA signals.

V. DISCUSSION

In the current study, we explored the possibility of detecting and classifying three levels of human stress (low, medium and high-stress levels) in 42 healthy people. Our findings showed that classification with linear and nonlinear features extracted from PPG and EDA signals is a good strategy for achieving an artificial intelligence-based automated system for detecting closed levels of human stress. Therefore, the nonlinear and linear dynamics of biological signals play a vital role in recognizing stress levels. This shows that the biological signals are not purely stochastic and random and follow a deterministic nonlinear behavior in response to different conditions. However, it should be noted that these results were obtained in laboratory conditions. All participants experienced a fixed setup with a noiseless environment, and different stress levels were induced through an executive cognitive task. However, this situation is totally different from real-life situations and the stress of everyday life, and this is the main limitation of this work. The use of wearable devices and virtual reality environments may alleviate this important limitation that future studies should consider.

Moreover, it should be noted that human stress may be unstable and temporary [31]. Therefore, it is very important to design a fast real-time system to detect stress levels in such situations. In addition, emotions and mental stress may be influenced by different physical and mental disorders. However, here, we only worked on healthy subjects. Consequently, our proposed system should be used with caution. On the other hand, our findings can be used in the field of psychiatry and psychology, and future studies should investigate the ability of our system to detect different levels of stress in psychiatric patients. Overall, our proposed automated stress detection system can be used in a wide range of settings to improve safety, health, and performance, including workplace safety, healthcare, education, sports, and transportation. For example, our stress detection system can be used in sports to monitor athletes' stress levels and provide feedback on how to manage stress during competition. This can help improve performance and reduce the risk of injury. Also, our stress detection system can be used in schools to monitor students' stress levels and provide support if necessary. This can help improve academic performance and reduce absenteeism.

Our proposed system showed good performance compared to previous studies. Healey and Picard proposed an automated system to distinguish three stress levels caused by a driving task through EMG, ECG, EDA and RSP signals and achieved an accuracy of 97% for this purpose [18]. However, the variety of biological signals in their work has led to a large increase in the cost of computations and the practical limitation of their proposed system. Zhai and Barreto reported an accuracy of 90% in distinguishing two stress levels using different biological signals and support vector machines [32]. This is despite the fact that we achieved the same accuracy as their work in our three-class problem. Furthermore, our proposed system outperformed the system introduced by Katsis et al., which achieved 79% accuracy in detecting three stress levels [21]. However, an important point that should be mentioned when comparing different systems is the lack of a comprehensive and public database for a more accurate evaluation and comparison of the systems proposed by researchers in this field.

VI. CONCLUSION

To sum up, we dealt with the crucial stages of an automated recognition system for three levels of human stress using biological signals of EDA and PPG, from signal recording to the classification step, and investigated the results from each stage of this system. Using the proposed system, we achieved a mean detection accuracy of 88%, which provides evidence to show autonomic nervous system differences among different stress levels. A range of biological features from linear and nonlinear analyzes was calculated to obtain the optimum stress-related features. Our findings showed that combining nonlinear and linear approaches for biological data analysis (PPG and EDA) could help develop a stress detection system. At the end, we call to action for a comprehensive, publicly accessible database of physiological signals to evaluate and compare stress detection systems rigorously.

REFERENCES

[1] A. A. Noorbala, H. Rafiey, F. Alipour, and A. Moghanibashi-Mansourieh, “Psychosocial stresses and concerns of people living in Tehran: a survey on 6000 adult participants,” Iran J Psychiatry, vol. 13, no. 2, p. 94, 2018.

[2] H. M. Al-Kuraishy and A. I. Al-Gareeb, “Eustress and malondialdehyde (MDA): Role of Panax ginseng: Randomized placebo-controlled study,” Iran J Psychiatry, vol. 12, no. 3, p. 194, 2017.

[3] E. R. De Kloet, M. Joëls, and F. Holsboer, “Stress and the brain: from adaptation to disease,” Nat Rev Neurosci, vol. 6, no. 6, pp. 463–475, 2005.

[4] C. Tsigos, I. Kyrou, E. Kassi, and G. P. Chrousos, “Stress: endocrine physiology and pathophysiology,” Endotext [Internet], 2020.

[5] M. S. Saddiq et al., “Effect of salinity stress on physiological changes in winter and spring wheat,” Agronomy, vol. 11, no. 6, p. 1193, 2021.

[6] M. C. Pascoe, D. R. Thompson, and C. F. Ski, “Yoga, mindfulness-based stress reduction and stress-related physiological measures: A meta-analysis,” Psychoneuroendocrinology, vol. 86, pp. 152–168, 2017.

[7] I. C. Maduka, E. E. Neboh, and S. A. Ufelle, “The relationship between serum cortisol, adrenaline, blood glucose and lipid profile of undergraduate students under examination stress.,” Afr Health Sci, vol. 15, no. 1, pp. 131–136, 2015.

[8] E. B. Bloss, W. G. Janssen, B. S. McEwen, and J. H. Morrison, “Interactive effects of stress and aging on structural plasticity in the prefrontal cortex,” Journal of Neuroscience, vol. 30, no. 19, pp. 6726–6731, 2010.

[9] M.-A. Besharat, H. Khadem, V. Zarei, and A. Momtaz, “Mediating role of perceived stress in the relationship between facing existential issues and symptoms of depression and anxiety,” Iran J Psychiatry, vol. 15, no. 1, p. 80, 2020.

[10] M. H. Largani et al., “Depression, Anxiety, Perceived Stress and Family Support in COVID-19 Patients,” Iran J Psychiatry, vol. 17, no. 3, pp. 257–264, 2022.

[11] S. Mohammadkhani, A. Foroutan, M. Akbari, and M. Shahbahrami, “Emotional Schemas and Psychological Distress: Mediating Role of Resilience and Cognitive Flexibility,” Iran J Psychiatry, vol. 17, no. 3, p. 284, 2022.

[12] P. J. Gianaros and T. D. Wager, “Brain-body pathways linking psychological stress and physical health,” Curr Dir Psychol Sci, vol. 24, no. 4, pp. 313–321, 2015.

[13] M. A. Stults-Kolehmainen and R. Sinha, “The effects of stress on physical activity and exercise,” Sports medicine, vol. 44, pp. 81–121, 2014.

[14] G. Valenza, A. Lanata, and E. P. Scilingo, "The role of nonlinear dynamics in affective valence and arousal recognition," IEEE Trans Affect Comput, vol. 3, no. 2, pp. 237–249, 2011.

[15] A. Greco, G. Valenza, L. Citi, and E. P. Scilingo, "Arousal and valence recognition of affective sounds based on electrodermal activity," IEEE Sens J, vol. 17, no. 3, pp. 716–725, 2016.

[16] X. Niu, L. Chen, H. Xie, Q. Chen, and H. Li, "Emotion pattern recognition using physiological signals," Sensors & Transducers, vol. 172, no. 6, p. 147, 2014.

[17] L. Berkowitz, T. Dalgleish, and M. Power, "Handbook of cognition and emotion." Chichester, UK: Wiley, 1999.

[18] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," IEEE Transactions on intelligent transportation systems, vol. 6, no. 2, pp. 156–166, 2005.

[19] R. Arefi Shirvan, S. K. Setarehdan, and A. Motie Nasrabadi, "Classification of mental stress levels by analyzing fNIRS signal using linear and non-linear features," International Clinical Neuroscience Journal, vol. 5, no. 2, pp. 55–61, 2018.

[20] G. N. Yannakakis and J. Hallam, "Entertainment modeling through physiology in physical play," Int J Hum Comput Stud, vol. 66, no. 10, pp. 741–755, 2008.

[21] C. D. Katsis, N. S. Katertsidis, and D. I. Fotiadis, "An integrated system based on physiological signals for the assessment of affective states in patients with anxiety disorders," Biomed Signal Process Control, vol. 6, no. 3, pp. 261–268, 2011.

[22] P. Karthikeyan, M. Murugappan, and S. Yaacob, "A review on stress inducement stimuli for assessing human stress using physiological signals," in 2011 IEEE 7th International Colloquium on Signal Processing and its Applications, IEEE, 2011, pp. 420–425.

[23] B. Pehlivanoğlu, N. Durmazlar, and D. Balkancı, "Computer adapted Stroop colour-word conflict test as a laboratory stress model," Erciyes Medical Journal, vol. 27, no. 2, pp. 58–63, 2005.

[24] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxEDA: A convex optimization approach to electrodermal activity processing," IEEE Trans Biomed Eng, vol. 63, no. 4, pp. 797–804, 2015.

[25] G. Udovičić, J. Đerek, M. Russo, and M. Sikora, "Wearable emotion recognition system based on GSR and PPG signals," in Proceedings of the 2nd international workshop on multimedia for personal health and health care, 2017, pp. 53–59.

[26] A. Burns et al., "SHIMMERTM: an extensible platform for physiological signal capture," in 2010 annual international conference of the IEEE engineering in medicine and biology, IEEE, 2010, pp. 3759–3762.

[27] A. Khaleghi et al., "EEG classification of adolescents with type I and type II of bipolar disorder," Australas Phys Eng Sci Med, vol. 38, pp. 551–559, 2015.

[28] A. Khaleghi, P. M. Birgani, M. F. Fooladi, and M. R. Mohammadi, "Applicable features of electroencephalogram for ADHD diagnosis," Research on Biomedical Engineering, vol. 36, pp. 1–11, 2020.

[29] M. R. Mohammadi, A. Khaleghi, A. M. Nasrabadi, S. Rafieivand, M. Begol, and H. Zarafshan, "EEG classification of ADHD and normal children using non-linear features and neural network," Biomed Eng Lett, vol. 6, pp. 66–73, 2016.

[30] H. Zarafshan, A. Khaleghi, M. R. Mohammadi, M. Moeini, and N. Malmir, "Electroencephalogram complexity analysis in children with attention-deficit/hyperactivity disorder during a visual cognitive task," J Clin Exp Neuropsychol, vol. 38, no. 3, pp. 361–369, 2016.

[31] G. Russell and S. Lightman, "The human stress response," Nat Rev Endocrinol, vol. 15, no. 9, pp. 525–534, 2019.

[32] J. Zhai and A. Barreto, "Stress detection in computer users based on digital signal processing of noninvasive physiological variables," in 2006 international conference of the IEEE engineering in medicine and biology society, IEEE, 2006, pp. 1355–1358.

Development of a Framework for Classification of Impulsive Urban Sounds using BiLSTM Network

Nazbek Katayev[1], Aigerim Altayeva[2], Bayan Abduraimova[3], Nurgul Kurmanbekkyzy[4], Zhumabay Madibaiuly[5], Bakhytzhan Kulambayev[6]
Kazakh National Women's Teacher Training University, Almaty, Kazakhstan[1]
Al-Farabi Kazakh National University, Almaty, Kazakhstan[2]
L. N. Gumilyov Eurasian National University, Astana, Kazakhstan[3]
Kazakh-Russian Medical University, Almaty, Kazakhstan[4]
Joldasbekov Institute of Mechanics and Engineering, Almaty, Kazakhstan[5]
Academy of Logistics and Transport, Almaty, Kazakhstan[5]
Turan University, Almaty, Kazakhstan[6]

***Abstract*—Urban environments are awash with myriad sounds, among which impulsive noises stand distinct due to their brief and often disruptive nature. As cities evolve and expand, the accurate classification and management of these impulsive sounds become paramount for urban planners, environmental scientists, and public health advocates. This paper introduces a novel framework leveraging the Bidirectional Long Short-Term Memory (BiLSTM) Network for the systematic categorization of impulsive urban sounds. Traditional methodologies often falter in recognizing the nuanced intricacies of such noises. In contrast, the presented BiLSTM-based approach adapts to the temporal variability intrinsic to these sounds, thereby enhancing classification accuracy. The research harnesses an expansive dataset, curated from various urban settings, to train and validate the model. Preliminary findings suggest that our BiLSTM framework outperforms existing models, with a marked increase in both specificity and sensitivity metrics. The outcome of this study holds profound implications for city acoustics management, noise pollution control, and urban health interventions. Moreover, the framework's adaptability paves the way for its application across diverse acoustic landscapes beyond the urban realm. Future endeavors should seek to further optimize the model by integrating more diverse soundscapes and addressing potential biases in data collection.**

***Keywords*—Impulsive sound; machine learning; deep learning; CNN; LSTM; classification**

I. Introduction

In the vibrant tapestry of urban life, sounds and noises play an integral role, shaping the auditory landscapes that city inhabitants navigate daily. The urban soundscape, a combination of ambient noises, human interactions, vehicular movements, and sudden, impulsive sounds, constitutes an integral aspect of urban living [1]. These sounds, particularly the impulsive varieties, serve as a double-edged sword [2]. On one hand, they contribute to the character and ambiance of a city, often evoking deep-seated memories and emotional responses among its residents. On the other hand, unchecked and discordant impulsive noises can deteriorate the quality of life, leading to stress, sleep disturbances, and even chronic health issues [3]. Consequently, the significance of identifying, classifying, and managing these sounds in urban spaces cannot be overstated.

While a plethora of research has focused on the broad soundscape of cities, the niche area of impulsive urban sounds has traditionally been underserved. Defined by their short, abrupt nature, these sounds—be it the honk of a car, the clang of a dropped tool, or the burst of fireworks—pose unique challenges to classification systems [4]. Traditional audio classification models, built primarily for longer and more consistent sounds, often struggle to capture the fleeting nuances of impulsive noises [5]. The rapid onset and offset of these sounds, combined with their varied frequency range, demand an approach that is both sensitive to temporal dynamics and adaptable to a broad acoustic spectrum.

Enter the realm of neural networks, which in recent years, has revolutionized the domain of sound classification. Among neural architectures, the Long Short-Term Memory (LSTM) network [6], a type of recurrent neural network, has shown promise in handling sequences and time-series data, making it a suitable contender for our auditory challenge. However, a unidirectional LSTM processes data in its input sequence order, potentially overlooking patterns that emerge from the reverse sequence of sounds [7]. Recognizing this limitation, and drawing inspiration from the bidirectional nature of human auditory processing where sounds are often understood in the context of both preceding and following sounds, this research innovatively employs the Bidirectional Long Short-Term Memory (BiLSTM) Network [8]. The BiLSTM, by virtue of processing an input sequence in both forward and backward directions, stands poised to capture the intricate patterns and characteristics intrinsic to impulsive urban sounds, offering a comprehensive understanding of their structure.

This paper, therefore, sets forth with a dual agenda. Firstly, it seeks to elucidate the significance and complexity of impulsive urban sounds, grounding its arguments in both auditory science and urban studies. Secondly, it embarks on a journey to explore the efficacy of the BiLSTM framework in classifying these sounds, aiming to bridge the gap between neural network research and urban acoustic management. In doing so, this research not only endeavors to advance the field

of auditory classification but also aspires to have a tangible impact on urban planning, noise pollution control measures, and public health interventions.

As we delve deeper into this exploration, it becomes imperative to understand the broader context within which urban sounds exist, the technological advancements in neural networks, and the potential applications of an effective classification system. Through this multi-faceted lens, this research hopes to offer a comprehensive view of the challenges and opportunities that lie at the intersection of urban acoustics and advanced neural architectures.

II. Related Works

Amid the bustling panorama of urban existence, the cacophony of sounds emerges not just as an incidental backdrop, but as an active participant shaping the dynamics of city life. The auditory fabric of urban centers is woven with diverse threads, ranging from the rhythmic footfalls on pavements to the occasional discordant blare of car horns [9]. Within this intricate web, impulsive urban sounds—transient, unexpected, and often sharp in nature—hold a unique place [10]. Their fleeting existence and unpredictable onset present both an auditory intrigue and a challenge that merit academic and scientific exploration.

As cityscapes continue to evolve, converging towards a future that's increasingly urbanized, the sonic environment they foster becomes an indispensable area of study. The implications of these sounds stretch across various dimensions: psychological, sociological, environmental, and even physiological [11]. For instance, while a distant church bell or a street performer's melody might evoke feelings of nostalgia or joy, the sudden screech of brakes or a loud explosion can trigger stress or anxiety [12]. The dichotomy of these reactions underscores the relevance of understanding and classifying urban sounds, especially those of an impulsive nature [13]. Given their impact on the well-being of city dwellers, mental health, and the broader urban experience, a systematic study becomes not just an academic endeavor but a societal imperative.

Historically, the academic arena has demonstrated a sustained interest in urban noises, resulting in extensive literature on the general soundscape of cities [14]. However, when it comes to the niche area of impulsive sounds, the scholarly attention seems somewhat disproportionate [15]. This relative dearth is surprising, given that impulsive noises, by virtue of their sudden onset and varied frequency profiles, pose unique challenges. Traditional auditory classification models, designed with an inclination towards consistent and prolonged sounds, falter when faced with the erratic nature of impulsive noises [16]. The fleeting presence and diverse acoustic characteristics of these sounds necessitate an approach that's not only nimble but also adept at capturing rapid temporal fluctuations [17].

Enter the world of advanced neural networks—a domain that has, in recent times, transformed numerous fields, including audio processing [18]. Among the neural architectures on offer, the Long Short-Term Memory (LSTM) network, a subtype of recurrent neural networks, has emerged as a front-runner for tasks involving sequence or time-series data [19]. Given its prowess in handling sequential data, LSTM offers a glimmer of hope for the impulsive sound conundrum. However, traditional LSTM, being unidirectional, processes sequences in the order they are presented, potentially missing out on valuable insights that could be gleaned from reverse order processing.

It's against this backdrop that this research introduces the Bidirectional Long Short-Term Memory (BiLSTM) [20] Network to the equation. Drawing parallels from human auditory processing, which inherently understands sounds based on both their preceding and succeeding context, the BiLSTM processes sequences bidirectionally—both forwards and backwards. This bidirectional approach promises a more holistic grasp of impulsive urban sounds, capturing nuances that might escape unidirectional models [21]. By processing sounds in this dual manner, the BiLSTM aspires to straddle the intricate patterns and temporal dynamics intrinsic to impulsive noises.

This paper embarks on a journey with twofold objectives. First, it aims to contextualize the importance and intricacies of impulsive urban sounds within the broader discourse of urban studies and auditory science [22]. It strives to illustrate why these sounds, often sidelined in scholarly pursuits, deserve focused attention. Second, the research delves into the technical and empirical exploration of the BiLSTM framework, investigating its potential as the much-needed solution to the challenges posed by impulsive sounds. Through this synthesis, the paper hopes to create a bridge—linking the often disparate worlds of neural network research and urban acoustic management.

As we venture further into this academic exploration, we're invited to reflect upon a myriad of interconnected themes: the transformative power of neural networks, the complex tapestry of urban soundscapes, and the potential societal ramifications of effective sound classification. Through this kaleidoscopic lens, this research endeavors to provide a comprehensive and nuanced perspective, setting the stage for groundbreaking revelations in the crossroads of urban acoustics and neural network technology.

III. Materials and Methods

In line with our initial conceptual framework, there is a two-fold requirement: first, to register the designated sound analysis apparatus, and second, to subject it to rigorous training [23]. This machine, once operational, deciphers the ingested auditory data, which can span a gamut of audio formats, a notable example being the mp3 format. Upon the reception of such an audio file, the machine subsequently crafts its associated spectrogram. A spectrogram, often interchangeably termed a sonogram (Fig. 1), is a graphical rendering that elucidates the relationship between the spectral density of a signal's power and its temporal progression [24]. Historically, spectrograms have found multifaceted applications across disciplines. They play a pivotal role in areas such as speech recognition, analysis of animal vocal patterns, diverse musical domains, radio and sonar technologies, linguistic signal processing, seismological research, and several other specialized fields.

A. Searching and Selecting Dataset

Each model training task requires a lot of input data, and the quality of our model will essentially depend on them. Therefore, the choice of dataset is an important part when building a model. Often the data also needs to be filtered or "cleaned up" in case some of the samples contain misrepresentations or false sounds for the class.

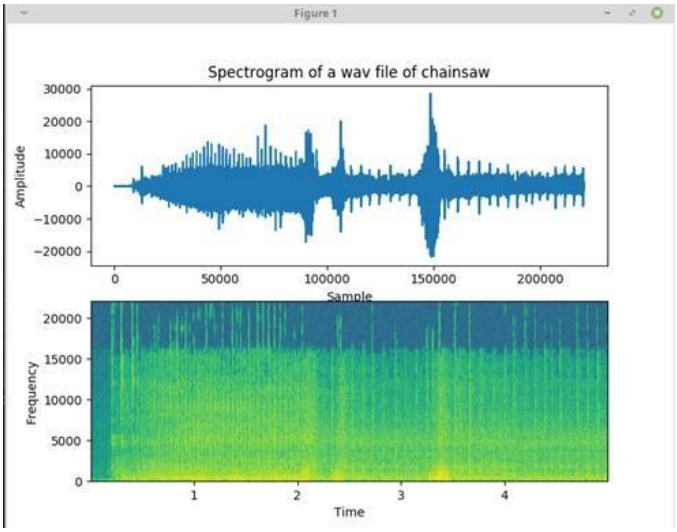


Fig. 1. Example of spectrogram of chainsaw.

We found two very interesting datasets, the first is the UrbanSounds-8K and second is UrbanAudioDataset. But firstly, let's see the first one.

UrbanSounds-8K. This dataset [25] contains about ~900 ".wav" sound files for each 10 classes such as:

- Air conditioners
- Car horns
- Children playing
- Dog barking
- Drilling
- Engine idling
- Gun shots
- Jackhammers
- Sirens
- Street music

Total 8732 audio files. This dataset, is pretty big (about 6,6 gigabytes) but as we know, the more data, the more we can train the model. That's why we believe that it will perfectly show the advantages and disadvantages of model preprocessing. And in the first part of our research we will use it.

Urban Audio Dataset. The second dataset [26] was collected from various resources and consist about 10000 samples for next eight classes:

- Crying
- Dog barking
- Emergency alarms
- Explosions
- Fire
- Glass breaking
- Screaming
- The sound of a weapon firing(gun shots)

However, the data in it requires normalization, since the data format is extremely different (with formats like: .mp3, .aiff, .flac, .wav, .m4a), also weights about 30 gigabytes! Therefore, we will use this dataset only after checking the main model. And this dataset is more suitable for our problem, since these sounds are more suitable for alerting danger. But again, we will talk about this dataset in more detail later in the next parts of the research.

B. Environment Selection

After the datasets, let's think about the hardware environment. Initially, the development was carried out on a virtual machine in the VirtualBox image on the Linux Mint system, which was sharpened for computer vision tasks, and in particular OpenCV. It fit the prototype, but due to the limitations of virtualization, it was decided to transfer the project to the main machine (host machine) on the Windows 10 operating system.

Now let's decide for the environment itself. Machine learning and deep learning in general are very widely used in the Python language due to ease of use, however, it is worth mentioning that for performance, you can try developing in C / C ++ if the issue of performance will play a key role. But we believe that Python 3.7 is enough for this task.

For analysis and research, we will use Jupyter Notebook. It is on it that we can write lines of code that we can interpret in different ways and thereby observe changes in individual cells of the program launch.

For the final part of development in production, we will be using PyCharm by JetBrains (Community Edition) and IntelliJ IDEA with Java Spring Framework for backend. But about interacting with application we will explain later in next parts.

C. Sound Processing

In this section, we will talk about the part about digital sounds. We decided that it was very important to fully understand how a computer can pick up sound, how we can work with it, and how to adapt it for classification. In this subsection, we will try to answer these questions.

Digital audio is the result of converting an analog audio signal into a digital audio format. There are a lot of audio formats at the moment, such as .ogg, .wav, .mp3, .flac and more. They differ in their storage and playback properties, but all alone cope with their task - they transmit an audio signal within a digital system. In our case, the sound will be displayed as a graph like in Fig. 2:

From which we can later obtain values after sampling when converting from an analog signal to digital. In other words, the digital sound that is already in the computer is already converted and we can work directly with it.

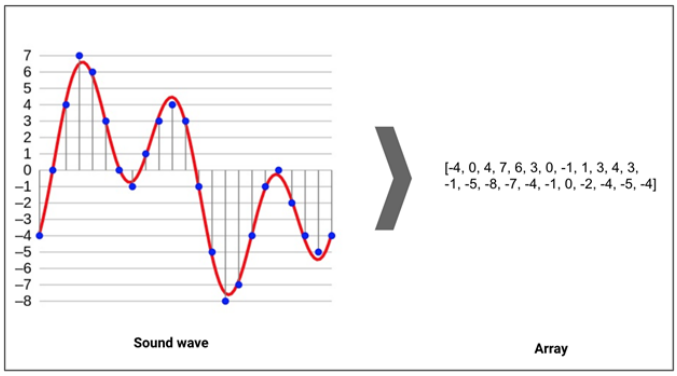


Fig. 2. Example of converting sound wave to array type.

D. Sound Analysis and Visualization

Audio has its own sound transmission channels, these channels called "Left-Right" output [27], in simple words, the output of the whole sound goes immediately to both the left ear and the right. This is called a mono channel. And the recording of such sound is carried out only from one input device, for example, a conventional microphone.

However, when we need to add more different kinds of sounds/effects, this is where stereo sound comes to us. Its fundamental difference lies in the fact that the received sound does not go to both Left-Right channels, but specifically to the Left and separately to the Right [28]. Thus, the sound in the channels acquires a certain volume in the sound. For a more comparative analysis, you can see their display in Fig. 3.

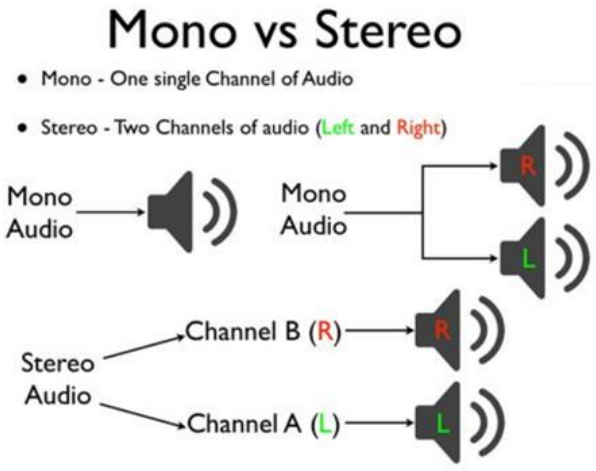


Fig. 3. Mono and stereo comparison.

And in order to see the difference for ourselves, we can take a different sound and compare their graphs, it is enough to display them through the Matplotlib library, which will allow us to do this [29]. To do this, we import and take another example with a mono channel for comparison, this will be a barking dog. Now let's display in Fig. 4:

In the graph, we can see that mono sound is displayed as one color, when the colors are displayed differently in stereo. Also, to check the channel, you can write a function that, using .shape, will show us a mono or stereo channel. This will especially help in the analysis of the second dataset.

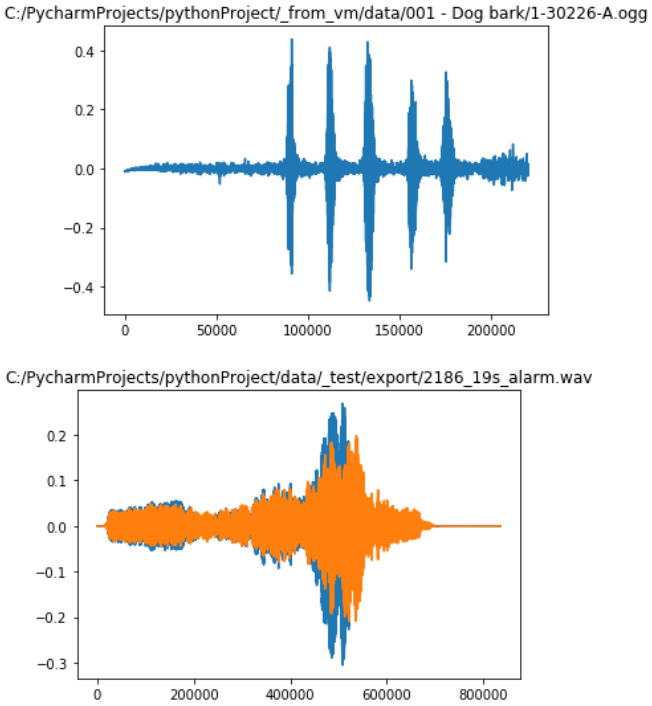


Fig. 4. Visualization of mono sound (upper), stereo sound (lower).

E. Proposed Model

There are deep learning techniques that can be applied in different areas as sound processing, images or video processing [30]. Urban environments, marked by their dynamic interactions and complexities, continuously emanate a diverse array of sounds, ranging from the benign murmurs of daily life to potentially dangerous noises that can indicate emergent situations or hazards [31]. Accurately discerning and classifying these dangerous sounds is not only paramount for the enhancement of urban safety but also imperative for proactive response mechanisms in smart cities. Traditional sound classification techniques often fall short in recognizing these transitory yet critical sounds due to their inherent limitations in capturing temporal relationships [32]. Enter the Bidirectional Long Short-Term Memory (BiLSTM) model, a sophisticated neural network architecture designed to navigate such challenges with unparalleled efficacy [33]. Fig. 5 demonstrates a flowchart of the proposed BiLSTM network for impulsive urban sound detection.

At its core, the Long Short-Term Memory (LSTM) is a form of Recurrent Neural Network (RNN) that addresses the vanishing gradient problem inherent in traditional RNNs. LSTMs are equipped with memory cells that can maintain information in memory for long periods, making them especially adept at tasks that require the understanding of long-term dependencies — a feature highly relevant to sound

sequences where past sounds can influence the characterization of present ones.

However, when dealing with dangerous urban sounds, which are often abrupt and embedded within larger, intricate auditory contexts, it becomes essential to understand the sound in relation to both its past and forthcoming sequences. This is where the bidirectional approach of the BiLSTM becomes invaluable. Instead of processing sequences in a unidirectional manner (from past to present), the BiLSTM simultaneously processes the data in both forward and backward directions. This bidirectional processing ensures that the model has access to information from both before and after a particular time step, enabling a more comprehensive understanding of the sound's context.

In the context of dangerous urban sound classification, this means that a sudden loud crash, which could signify a vehicular accident or a structural collapse, is not just evaluated based on preceding sounds, but also by the sounds that follow it. Such a dual-context perspective can be crucial in distinguishing between, say, a harmless crash in a construction site versus a car collision that requires immediate attention.

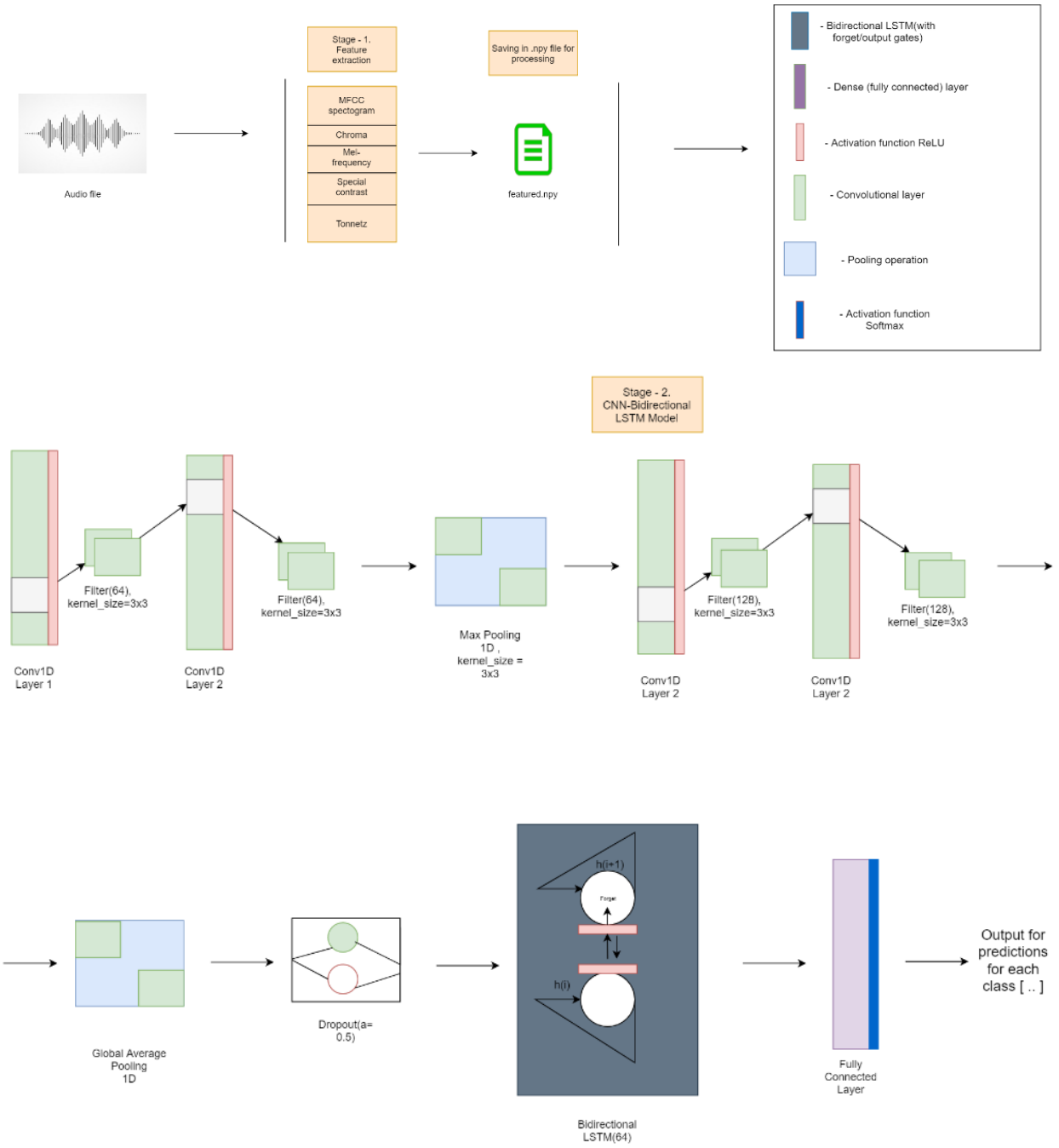


Fig. 5. Proposed Bidirectional LSTM network.

IV. Experimental Results

Furthermore, the inherent structure of the BiLSTM, with its memory gates, allows for meticulous filtering of sound data, ensuring that only relevant information is retained for classification. This selective retention is especially crucial for urban environments where the soundscape is cluttered, and the distinction between dangerous and non-dangerous sounds can be razor-thin.

A. Results of the Proposed Model

The proposed Bidirectional Long Short-Term Memory model offers a groundbreaking approach to dangerous urban sound classification. Its ability to capture intricate temporal relationships from both past and future contexts, coupled with its adeptness at managing long-term dependencies, positions the BiLSTM as a frontrunner in the ongoing quest for creating safer and smarter urban ecosystems. As urban centers across the globe grapple with the challenges of increasing density and complexity, such advanced neural network architectures emerge not just as academic curiosities, but as essential tools for ensuring the well-being and safety of their inhabitants.

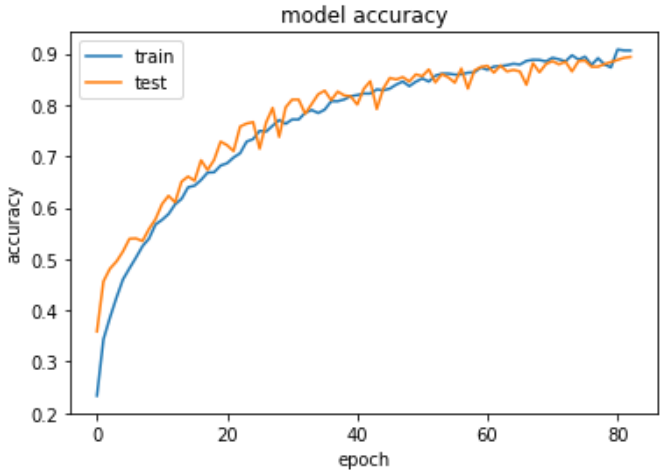


Fig. 6. Model training accuracy.

Further, this is a bidirectional LSTM network that immediately grows in parameters to as many as 352,330. This means that in theory there should be good results. The training time took about 871 seconds or about 14 minutes. Fig. 6 demonstrates the model training and test accuracy for 80 learning epochs.

Fig. 7 demonstrates model training loss in 80 learning epochs. The results show that, the proposed bidirectional LSTM network achieves to 90% accuracy, and 10% training loss, respectively.

Fig. 8 and Fig. 9 demonstrate test accuracy and test loss of the proposed model. Test results show that, the proposed model achieves 90% accuracy in model testing.

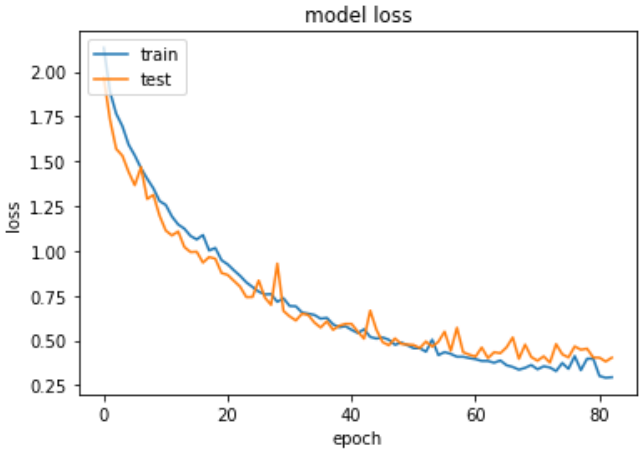


Fig. 7. Model training loss.

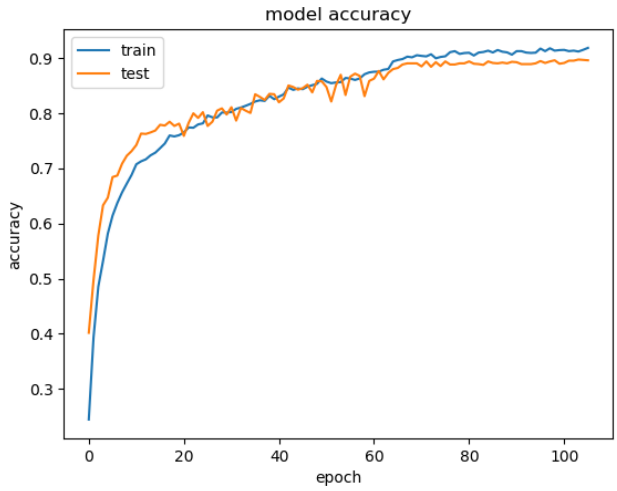


Fig. 8. Model test accuracy.

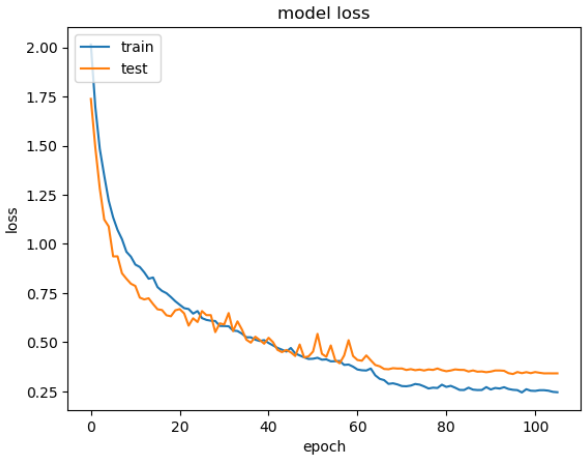


Fig. 9. Model test loss.

V. DISCUSSION

The intricacies and dynamics of urban environments demand a profound, nuanced understanding, especially when delving into the auditory spectrum of these landscapes. Our exploration into the Bidirectional Long Short-Term Memory (BiLSTM) model for classifying dangerous urban sounds opens an array of discussions, both in the realm of neural network architectures and urban acoustics.

One of the most compelling findings from this research is the marked superiority of the BiLSTM model in classifying impulsive, dangerous sounds as compared to traditional sound classification techniques. The bidirectional nature of the model, which processes sound sequences both in forward and reverse temporal orders, demonstrates an inherent advantage in capturing the context of impulsive noises [34]. By concurrently assessing preceding and subsequent sounds, the model offers a panoramic view of the auditory environment, a perspective pivotal in discerning potential dangers in bustling urban soundscapes.

However, while the BiLSTM demonstrates significant promise, it's essential to address its limitations. Training a BiLSTM model, particularly with expansive urban sound datasets, can be computationally demanding [35]. The simultaneous processing of forward and backward sequences necessitates robust computational resources, which may not be readily available in all application scenarios, especially in real-time urban monitoring systems [36]. As cities move toward the vision of smart urbanism, the real-time processing of data becomes crucial. Future research should, therefore, look into optimizing the BiLSTM structure without compromising its classification prowess.

Another aspect worth reflecting upon is the diversity of the urban soundscape dataset employed in this study [37]. While the dataset was expansive, it was primarily curated from a limited number of urban settings. Urban soundscapes can significantly vary based on factors like cultural practices, architectural designs, traffic patterns, and even weather conditions. For the BiLSTM model to be universally applicable, it's imperative to train it with a more globally representative dataset, encompassing the myriad variations of urban environments. This would enhance the model's adaptability, ensuring its efficacy across diverse urban landscapes.

Furthermore, the human auditory system, despite its biological limitations, possesses a remarkable ability to discern sounds based on learned experiences and cultural contexts [38]. The sudden clang of pots in one culture might be dismissed as a benign household activity, while in another, it could be an alert for danger. Incorporating such cultural nuances and learned experiences into the BiLSTM model presents a challenge and an opportunity. The integration of these elements might enhance the model's sensitivity to context-specific dangerous sounds, making it even more aligned with human auditory perception.

Lastly, the ethical considerations of continuous urban sound monitoring need to be highlighted. While the primary intent is safety and rapid response to dangerous situations, the omnipresent nature of sound monitoring systems can raise concerns related to privacy and surveillance. It becomes imperative for urban planners and policymakers to strike a balance, ensuring that the pursuit of safety doesn't infringe upon the privacy rights of city inhabitants.

In summation, this research underscores the transformative potential of the BiLSTM model in the realm of dangerous urban sound classification. The model's bidirectional processing, its adeptness at capturing temporal nuances, and its alignment with the holistic human perception of sounds make it an invaluable tool in the urban auditory toolkit. However, like all pioneering endeavors, this study raises as many questions as it seeks to answer. The computational demands of the model, the need for a more globally diverse dataset, the integration of cultural nuances, and the overarching ethical considerations form a rich tapestry of challenges and opportunities for future research.

As urban centers continue to burgeon and evolve, the imperative to understand, manage, and respond to their auditory landscapes becomes even more pronounced. The Bidirectional Long Short-Term Memory model, with its blend of technological sophistication and auditory acumen, emerges as a beacon in this journey, illuminating the path toward safer, smarter, and more responsive urban ecosystems. This research, albeit a single step, paves the way for a future where cities don't just listen but truly understand.

VI. CONCLUSION

In an era where urban expanses are rapidly growing, manifesting themselves as the epicenters of human civilization, understanding the multifaceted dimensions of these environments is imperative. The auditory realm of cities, teeming with a symphony of sounds both benign and dangerous, necessitates an analytical lens equipped with both precision and depth. This research, centered on the Bidirectional Long Short-Term Memory (BiLSTM) model, underscores this very sentiment, offering a pioneering approach to the classification of dangerous urban sounds.

Our exploration into the BiLSTM model has illuminated its profound potential. By processing sound sequences in both forward and reverse temporal frames, the model imitates the holistic human perception of sounds, transcending the limitations of traditional classification techniques. This bidirectional prowess not only captures the intricate nuances of dangerous sounds but also provides a broader context, pivotal for accurate classification in bustling urban settings.

However, as is characteristic of any academic endeavor, this study also opens avenues for further exploration. While the BiLSTM model is undeniably potent, its computational demands, adaptability across diverse urban landscapes, and the integration of cultural and learned auditory nuances present challenges warranting future research. Moreover, the ethical dimensions of continuous urban sound monitoring, with potential implications for privacy and surveillance, underscore the need for a balanced approach, harmonizing safety with individual rights.

In conclusion, this research signifies a seminal step in the realm of urban sound classification. The BiLSTM model

emerges not merely as a technological marvel but as a testament to the convergence of neural network architectures and urban auditory science. As cities continue their inexorable march towards the future, tools like the BiLSTM will play a pivotal role, ensuring that these urban giants are not just expanses of concrete and steel, but responsive, adaptive, and safe ecosystems for all their inhabitants.

ACKNOWLEDGMENT

This work was supported by the research project — Development of a system for detecting and alerting dangerous events based on the audio analysis and machine learning funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan. Grant No. IRN AP19175674.

REFERENCES

[1] J. Bajzik, J. Prinosil, R. Jarina and J. Mekyska, "Independent channel residual convolutional network for gunshot detection," International Journal of Advanced Computer Science and Applications, vol. 13, no.4, pp. 950-958, 2022.

[2] K. M. Nahar, F. Al-Omari, N. Alhindawi and M. Banikhalaf, "Sounds recognition in the battlefield using convolutional neural network," International Journal of Computing and Digital Systems, vol. 11, no.1, pp. 189-198, 2022.

[3] I. Estévez, F. Oliveira, P. Braga-Fernandes, M. Oliveira, L. Rebouta et al., "Urban objects classification using Mueller matrix polarimetry and machine learning," Optics Express, vol. 30, no.16, pp. 28385-28400, 2022.

[4] Z. Peng, S. Gao, Z. Li, B. Xiao, Y. Qian, "Vehicle safety improvement through deep learning and mobile sensing" IEEE Network, vol. 32, no.4, pp. 28-33, 2018.

[5] Y. Wei, L. Jin, S. Wang, Y. Xu and T. Ding, "Hypoxia detection for confined-space workers: photoplethysmography and machine-learning techniques," SN Computer Science, vol.3, no.4, pp.1-11, 2022.

[6] Omarov, B., Omarov, B., Shekerbekova, S., Gusmanova, F., Oshanova, N., Sarbasova, A., ... & Sultan, D. (2019). Applying face recognition in video surveillance security systems. In Software Technology: Methods and Tools: 51st International Conference, TOOLS 2019, Innopolis, Russia, October 15–17, 2019, Proceedings 51 (pp. 271-280). Springer International Publishing.

[7] K. Pawar, V. Attar, "Deep learning approaches for video-based anomalous activity detection," World Wide Web, vol. 22, no.2, pp.571-601, 2019.

[8] Sultan, D., Omarov, B., Kozhamkulova, Z., Kazbekova, G., Alimzhanova, L., Dautbayeva, A., ... & Abdrakhmanov, R. (2023). A Review of Machine Learning Techniques in Cyberbullying Detection. Computers, Materials & Continua, 74(3).

[9] H. Zogan, I. Razzak, X. Wang, S. Jameel, G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," World Wide Web, vol. 25, no.1, pp. 281-304, 2022.

[10] C. Heipke, F. Rottensteiner, "Deep learning for geometric and semantic tasks in photogrammetry and remote sensing," Geo-spatial Information Science, vol. 23, no.1, pp. 10-19, 2020.

[11] Y. Arslan, H. Canbolat, "Sound based alarming based video surveillance system design," Multimedia Tools and Applications, vol. 81, no.6, pp. 7969-7991, 2022.

[12] A. Rajbanshi, D. Das, V. Udutalapally, R. Mahapatra, "DLeak: an IoT-based gas leak detection framework for smart factory," SN Computer Science, vol. 3. no.4, pp. 1-12, 2022.

[13] Y. Arslan and H. Canbolat, "Sound based alarming based video surveillance system design," Multimedia Tools and Applications, vol. 81, no. 6, pp. 7969-7991, 2022.

[14] R. Sun, Q. Cheng, F. Xie, W. Zhang, T. Lin et. al., "Combining machine learning and dynamic time wrapping for vehicle driving event detection using smartphones," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no.1, pp.194-207, 2019.

[15] G. Chen, F. Wang, S. Qu, K. Chen, J. Yu et. al., "Pseudo-image and sparse points: vehicle detection with 2D LiDAR revisited by deep learning-based methods," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no.12, pp. 7699-7711, 2020.

[16] S. U. Amin, M. S. Hossain, G. Muhammad, M. Alhussein, M. A. Rahman, "Cognitive smart healthcare for pathology detection and monitoring," IEEE Access, vol. 7, no.1, pp. 10745-10753, 2019.

[17] Altayeva, A. B., Omarov, B. S., Aitmagambetov, A. Z., Kendzhaeva, B. B., & Burkitbayeva, M. A. (2014). Modeling and exploring base station characteristics of LTE mobile networks. Life Science Journal, 11(6), 227-233.

[18] I. H. Peng, P. C. Lee, C. K. Tien, J. S. Tong, "Development of a cycling safety services system and its deep learning bicycle crash model," Journal of Communications and Networks, vol. 24, no. 2, pp. 246-263, 2022.

[19] L. Kou, "A review of research on detection and evaluation of the rail surface defects," Acta Polytechnica Hungarica, vol. 19, no.3, pp. 167-186, 2022.

[20] L. M. Bine, A. Boukerche, L. B. Ruiz, A. A. Loureiro, "Leveraging urban computing with the internet of drones," IEEE Internet of Things Magazine, vol. 5, no.1, pp. 160-165, 2022.

[21] S. Khan, L. Alarabi and S. Basalamah, "Toward smart lockdown: a novel approach for COVID-19 hotspots prediction using a deep hybrid neural network," Computers, vol. 9, no. 4, pp. 1-16, 2020.

[22] M. Dua, D. Makhija, P. Manasa and P. Mishra, "A CNN–RNN–LSTM based amalgamation for Alzheimer's disease detection," Journal of Medical and Biological Engineering, vol. 40, no. 5, pp. 688-706, 2020.

[23] H. Gill, O. Khalaf, Y. Alotaibi, S. Alghamdi and F. Alassery, "Multi-model CNN-RNN-LSTM based fruit recognition and classification," Intelligent Automation & Soft Computing, vol. 33, no. 1, pp. 637-650, 2022.

[24] K. Chandriah and R. Naraganahalli, "RNN/LSTM with modified Adam optimizer in deep learning approach for automobile spare parts demand forecasting," Multimedia Tools and Applications, vol. 80, no. 17, pp. 26145-26159, 2021.

[25] S. Hansun and J. Young, "Predicting LQ45 financial sector indices using RNN-LSTM," Journal of Big Data, vol. 8, no. 1, pp. 1-13, 2021.

[26] Y. Xue, P. Shi, F. Jia, H. Huang, "3D reconstruction and automatic leakage defect quantification of metro tunnel based on SfM-Deep learning method," Underground Space, vol. 7, no.3, pp. 311-323, 2022.

[27] L. Zhang, L. Yan, Y. Fang, X. Fang, X. Huang, "A machine learning-based defensive alerting system against reckless driving in vehicular networks," IEEE Transactions on Vehicular Technology, vol. 68, no.12, pp.12227-12238, 2019.

[28] A. M. Youssef, B. Pradhan, A. Dikshit, M. M. Al-Katheri, S. S. Matar et. al., "Landslide susceptibility mapping using CNN-1D and 2D deep learning algorithms: comparison of their performance at Asir Region, KSA," Bulletin of Engineering Geology and the Environment, vol. 81, no.4, pp. 1-22, 2022.

[29] S. Asadianfam, M. Shamsi, A. Rasouli Kenari, "Hadoop Deep Neural Network for offending drivers," Journal of Ambient Intelligence and Humanized Computing, vol. 13, no.1, pp. 659-671, 2022.

[30] Tursynova, A., & Omarov, B. (2021, November). 3D U-Net for brain stroke lesion segmentation on ISLES 2018 dataset. In 2021 16th International Conference on Electronics Computer and Computation (ICECCO) (pp. 1-4). IEEE.

[31] D. K. Dewangan, S. P. Sahu, "Deep learning-based speed bump detection model for intelligent vehicle system using raspberry Pi," IEEE Sensors Journal, vol. 21, no.3, pp. 3570-3578, 2020.

[32] Z. Fang, B. Yin, Z. Du and X. Huang, "Fast environmental sound classification based on resource adaptive convolutional neural network," Scientific Reports, vol. 12, no. 1, pp. 1-18, 2022.

[33] V. Gugnani, R. K. Singh, "Analysis of deep learning approaches for air pollution prediction," Multimedia Tools and Applications, vol. 81, no.4, pp. 6031-6049, 2022.

[34] Omarov, B., Altayeva, A., Turganbayeva, A., Abdulkarimova, G., Gusmanova, F., Sarbasova, A., ... & Omarov, N. (2019). Agent based modeling of smart grids in smart cities. In Electronic Governance and Open Society: Challenges in Eurasia: 5th International Conference, EGOSE 2018, St. Petersburg, Russia, November 14-16, 2018, Revised Selected Papers 5 (pp. 3-13). Springer International Publishing.

[35] H. Kyle, P. Agarwal, J. Zhuang, "Monitoring misinformation on Twitter during crisis events: a machine learning approach," Risk Analysis, vol. 42, no.8, pp. 1728-1748, 2022.

[36] M. Esmail Karar, O. Reyad, A. Abdel-Aty, S. Owyed, M. F. Hassan, "Intelligent iot-aided early sound detection of red palm weevils," Computers, Materials & Continua, vol. 69, no.3, pp. 4095–4111, 2021.

[37] T. Thomas Leonid and R. Jayaparvathy, "Classification of elephant sounds using parallel convolutional neural network," Intelligent Automation & Soft Computing, vol. 32, no.3, pp. 1415–1426, 2022.

[38] Z. Ma, G. Mei, , F. Piccialli, "Machine learning for landslides prevention: a survey," Neural Computing and Applications, vol. 33, no.17, pp. 10881-10907, 2021.

Research on Evaluation Method of Urban Human Settlement Environment Quality Based on Back Propagation Neural Network

Siyuan Zhang[1], Wenbo Song[2]
Jilin University of Architecture and Technology, Changchun, 130114, China[1]
Jilin Provincial Defense Mobilization Office, Changchun, 130051, China[2]

***Abstract*—In order to improve people's living experience, a method for evaluating the quality of urban human settlements based on back propagation neural network is proposed. Firstly, the initial evaluation index system is constructed, the initial evaluation index system is screened, and the final evaluation index system is constructed by using the remaining evaluation indexes. Then, the back propagation neural network is constructed to build an evaluation model, and the evaluation model is trained through the processes of network initialization, hidden layer output calculation and output layer output calculation. Finally, the improved genetic algorithm is used to optimize the back propagation neural network, improve the evaluation performance of the back propagation neural network, and realize the evaluation of human settlements quality. The experimental results show that the accuracy of the evaluation results of urban human settlements quality output by the trained back-propagation neural network model reaches 96.3%, which has a good effect.**

***Keywords*—Back propagation; neural network; urban human settlements; quality evaluation; morbidity index; genetic algorithm**

I. INTRODUCTION

The quality of the human settlement environment has consistently been a fundamental concern in urban research, as it directly impacts human well-being and urban development [1]. With the rapid growth of the global population in the 20th century, the degradation of the ecological environment resulting from industrial development has contributed to the widening regional disparities, and global scholars are increasingly concerned about the sustainable development of human settlements. The urban human settlement environment is a comprehensive regional concept. This refers to the dynamic interaction between the natural environment and human activities [2] and includes all fields of humanity, geography, economy and environment. It is the activity area of urban residents. Theoretically, the meaning of an urban residential environment should be the harmony between man and nature, the unity of material and spirit [3]. However, in civil engineering and geography, they have their own emphasis. In civil engineering, they tend to regard the urban residential environment as a small-scale operation object, which is more reflected in the study of urban architectural planning. In geography, the urban human settlement environment is more regarded as a large-scale geoinformation system [4], and urban structure is more discussed. Assessing urban human settlements holds great significance in terms of the survival and development of the population [5], as it offers essential parameters for the transformation of urban industrial structure and the optimization of human settlements.

Staats et al. [6], in the process of studying the environmental quality evaluation method, the number of parked cars and the number of street trees were taken as the evaluation criteria, and the former was divided into four levels: from no car to full capacity, and the latter was divided into three levels: from no tree to 50% density. The evaluation model is constructed using the neural network in the artificial intelligence method; the evaluation of environmental quality is achieved by analyzing two evaluation indicators. When utilizing this method, it has been observed that the neural network exhibits limited global search capability, making it susceptible to extreme local values and slow convergence. These factors consequently impact the computational efficiency and prediction accuracy of the neural network. Fomina et al. [7] analyzed the key factors and standards that have a significant impact on the quality of the living environment of the population in their research on urban environmental quality assessment methods and considered that ecological environment, urban planning, landscape and social factors are the main evaluation indicators, each set of factors are calculated quantitatively (based on a set of standards and indicators), at the same time, the category coefficient obtained according to the expert survey results is considered in the calculation process. In the application process of this method, some indicators have a weak impact on the evaluation results, which is only because the evaluators empirically think that these indicators are important, while subjective experience may not be objective and reliable, which leads to deviation in the final evaluation results. Longhini et al. [8], in the process of studying the environmental quality evaluation method, the evaluation index for the study is selected using the geochemical multi-index method, and the constructed evaluation model is used to achieve the purpose of environmental quality evaluation. Different indicators must be weighted according to the actual situation. In the process of using this method, the qualitative analysis method is used to obtain the evaluation results, which inevitably leads to a one-sided evaluation and cannot reflect the real situation. Sarkheil et al. [9] employed the fuzzy comprehensive evaluation method to construct the evaluation model. The research area selected for evaluating the environmental quality was the Pass economic energy Zone. It is worth noting that during the

implementation of this method, there was a significant correlation observed among the evaluation indicators utilized; that is, the degree of information overlap between the indicators may be high. If this overlapping information is not processed, it will be emphasized repeatedly in the comprehensive evaluation, thus distorting the comprehensive evaluation results.

Although the above research has made some progress, there are still problems, such as the index system is not scientific and comprehensive, the differences in index selection and weight distribution in evaluation methods, and the lack of unified standards and systems. Therefore, this article studies the evaluation method of urban residential environment quality based on the back-propagation neural network and applies the back-propagation neural network to the evaluation process of urban residential environment quality. The significance of evaluating the quality of urban human settlements lies in evaluating and improving the living environment conditions of residents, providing scientific basis for urban planning, construction and management, and promoting the sustainable development of cities. By evaluating the quality of human settlements, the health of residents can be guaranteed, the influence of harmful factors on human body can be reduced and the quality of life can be improved. The evaluation results can also reveal the advantages and disadvantages of the city, provide decision-making basis for urban planning and construction, and promote the comprehensive development of the city. Evaluating and improving the quality of human settlements can enhance residents' happiness and satisfaction, and increase social stability and social harmony. Therefore, the quality evaluation of urban human settlements is of great significance to both cities and residents. Through the study of this method, it provides a reference for quantifying the quality of urban residential environments and monitoring the livability of cities. The overall structure of the article is as follows:

1) Build the initial evaluation index system, screen the initial evaluation index system, and use the remaining evaluation indexes to build the final evaluation index system.

2) Then, a back propagation neural network is constructed to build an evaluation model, and the evaluation model is trained through network initialization, hidden layer output calculation and output layer output calculation.

3) Using the improved genetic algorithm to optimize the back propagation neural network, improve the evaluation performance of the back propagation neural network, and realize the quality evaluation of human settlements.

II. Quality Assessment of Urban Human Settlements

A. Selection of Evaluation Indicators Based on Pathological Index Cycle Analysis

There are many factors that affect the quality of urban human settlements, and these factors can be used as evaluation indicators of urban human settlements. However, these evaluation indicators are not only different from each other but also closely related. To accurately evaluate the quality of urban human settlements, it is necessary to choose suitable evaluation indicators to build a highly scientific, highly operational, qualitative and quantitative evaluation indicator system. These indicators were selected based on expert opinions and research findings from scholars in relevant fields. Among them, experts' opinions and researchers in related fields play an important role in the quality evaluation of urban human settlements. Experts and researchers in related fields can provide valuable opinions and suggestions to help determine the importance and applicability of evaluation indicators based on in-depth research on urban environment and accumulation of professional knowledge. When selecting the evaluation index of urban human settlements quality, expert opinions can be obtained through expert consultation, expert interview and expert evaluation. The experience and professional knowledge of experts and researchers in related fields are helpful to determine the evaluation index with high authority and credibility. They can evaluate different environmental factors, human settlement needs and socio-economic factors based on their understanding of environmental disciplines and related fields, so as to determine the appropriateness and importance of evaluation indicators.

Considering the potential presence of duplicate evaluation indicators in the initial system for assessing the environmental quality of urban human settlements, where there may be overlapping information among the indicators, it becomes necessary to employ the pathological index cycle analysis method to screen and refine the evaluation indicators in the system.

In order to prevent redundant evaluation indicators from being retained in the selection process of evaluation indicators, a set of urban human settlements environmental quality evaluation indicators $X_1, X_2, \cdots, X_n$, all are the remaining evaluation indicators after removing the indicators with poor criticality. The specific process of screening environmental qualitative assessment of people's urban habitation indicators on the basis of the pathological indicator cycle analysis is as follows:

1) Determine the matrix with Formula (1) X^T characteristic value of $\phi_1, \phi_2, \cdots, \phi_n$:

$$|X^T - \phi_i \times D_n| = 0 \tag{1}$$

In Formula (1) X^T and D_n respectively represent the sample data matrix corresponding to the urban human settlements environmental quality evaluation indicator set transpose matrix and identity matrix of X.

2) C_{I_n} is used as a pathological index, which is determined by Formula (2) C_{I_n}:

$$C_{I_n} = \sqrt{\frac{\phi_1^*}{\phi_n^*}} \times X^T \tag{2}$$

In Formula (2), ϕ_1^* and ϕ_n^* respectively represent the matrix the upper and lower eigenvalues of X^T.

C_{I_n} describes the quality evaluation indicators of urban human settlements $X_1, X_2, \cdots, X_n$. The overall redundancy level is positively proportional to the overall redundancy level of the evaluation index.

3) Determine the indicator of clearing document X_i remaining after $n-1$ a pathological index of human settlements quality evaluation indicators in cities $C_{I_{(n-1)i}}$. Determine the residual according to the process of process (1) and process (2) $n-1$ pathological index of human settlements quality evaluation indicators in cities $C_{I_{(n-1)i}}$.

4) Use Formula (3) to determine the evaluation index of urban human settlement environment quality X_i overall redundancy contribution of C_{i1}:

$$C_{i1} = C_{I_n} - C_{I_{(n-1)i}} \quad (3)$$

C_{i1} describes the removal of urban residential environment quality assessment indicators X_i remaining after $n-1$ a conditioned index of evaluation indicators $C_{I_{(n-1)i}}$, the same as clearing X_i the pathological index of all previous evaluation indicators C_{I_n}. The value is the same as the evaluation index X_i the overall redundancy contribution to all evaluation indicators is positively proportional. The larger the C_{i1} is, the more it should be cleared of X_i.

5) Eliminate the indicators with the largest overall redundancy contribution in all urban human settlements' environmental quality evaluation indicators. If

$$C_{j1} = max\{C_{i1} | 1 \le i \le n\} \quad (4)$$

Indicate the evaluation indicators of the quality of human settlements in n cities $X_1, X_2, \cdots, X_n$ within X_j the largest contribution to the overall redundancy of the evaluation indicator set, X_j needs to be cleared.

For the convenience of description, the above process is defined as the first round of screening of redundant evaluation indicators. Cycle the above process to clear the remaining $n-1$ evaluation index with the largest overall redundancy contribution among the evaluation indexes. Thus, after several iterations, the evaluation indicators with the largest overall redundancy among the remaining evaluation indicators are removed in each iteration until the following termination conditions for the removal of redundant indicators are met.

The termination conditions for the removal of redundant indicators are as follows: if the morbidity index of all remaining urban human settlements' environmental quality assessment indicators is less than or equal to 10, the screening of redundant indicators will be terminated; On the contrary, the redundant evaluation indicators are continuously screened according to the above process until the morbidity index of the remaining evaluation indicators is less than or equal to 10.

Through the above process, the redundant evaluation indicators in the initial urban human settlements' environmental quality evaluation indicator system can be eliminated, and the final urban human settlements' environmental quality evaluation indicator system can be constructed.

B. Construction of Evaluation Index System

Following the implementation of the pathological index cycle analysis method, the evaluation indicators within the initial system for assessing the environmental quality of people's urban habitation have been screened, and the final evaluation index system is constructed, as shown in Table I.

TABLE I. INDEX SYSTEM FOR EVALUATING THE QUALITY OF URBAN RESIDENTIAL ENVIRONMENT

Target layer	Indicator layer
Quality of the urban living environment	Relief
	THI
	Vegetation Index
	Hydrological index
	Traffic accessibility
	Per Capita GDP

The calculation method of each evaluation index in the evaluation index system is as follows:

1) Topographic relief: As the foundation of human survival and development, the terrain changes and geomorphic characteristics on the surface will have a significant impact on the result of the qualitative assessment of people's habitation [10]. Referring to the research results of relevant scholars, the topographic relief is described by Formula (5):

$$R_{DLS} = (max\, H - min\, H) \times \frac{P(A)}{A} \times C_{j1} \quad (5)$$

In Formula (5), $max\, H$ and $min\, H$ are the upper and lower limits of regional altitude, $P(A)$ and A respectively represent the flat area and total area in the region.

2) Temperature humidity index: Climate conditions have an important impact on human activities. The temperature and humidity index describes the degree of mugginess under windless conditions [11], and the formula is as shown below:

$$T_{HI} = t + R_{DLS} \times f \times 1.8t \quad (6)$$

In Formula (6), t stands for the average monthly temperature in Celsius; f stands for the average value of air relative humidity.

3) Vegetation index: The vegetation coverage within the region can not only be used as the main indicator to judge in the qualitative assessment of people's habitation but even as a ecological condition of human life [12]. The normalized vegetation index is calculated as follows:

$$N_{DVI} = (N_{IR} - R)(N_{IR} + R) \times T_{HI} \quad (7)$$

In Formula (7), N_{IR} and R respectively represent reflectance in a near-infrared band and red band.

4) Hydrological index: As the main resource for survival and development in the region [13], the quality of hydrological resources plays a crucial role in determining the quality of urban human settlements [14], and hydrological resources can be described by hydrological index:

$$W_{RI} = (\alpha \times P + \beta \times W_a) \times N_{DVI} \quad (8)$$

In Formula (8), P and α respectively represents the average annual precipitation and its weight in the region, W_a and β respectively represent the water area and its weight.

5) Traffic accessibility: The perfection of the traffic network within the region has a significant role in promoting the improvement of the quality of urban human settlements, so the traffic network has become the main indicator of the adaptability evaluation of regional human settlements, which can be described by the traffic accessibility within the region. The calculation formula is as follows:

$$D = \frac{L_i}{A_i} \times W_{RI} \quad (9)$$

In Formula (9), L_i and A_i respectively represent the total length of roads and the area of the region.

6) GDP per capita: The economic level within the region has a significant impact on the quality of urban human settlements, which can be described by the per capita GDP within the region. The calculation formula is as shown below:

$$P_{CCDP} = \frac{Z_{GDP}}{R_{pop}} \times D \quad (10)$$

In Formula (10), Z_{GDP} and R_{pop} respectively represent the gross domestic product and the total population within the region.

C. Construction of Evaluation Model Based on Back-Propagation Neural Network

1) Evaluation model of back-propagation neural network: An artificial neural network (ANN), referred to as a neural network (NN), is a mathematical model or computing model that imitates the structure and function of the biological neural network [15]. Its unique nonlinear adaptive information processing capability makes it particularly suitable for solving problems with complex internal mechanisms [16], prediction and other fields that have been successfully applied. The back-propagation (BP) neural network is a popular and extensively utilized model in the field of artificial neural networks. It typically consists of an input layer, one or more hidden layers, and an output layer. Fig. 1 shows the topology of a back-propagation neural network with a hidden layer.

The back-propagation neural network algorithm utilizes the gradient descent method as its underlying principle. During the learning process (training) of a neural network, back and forward propagation are utilized. Within the forward engendering stage, intake data is sequentially handled from the intake layer to the covered up one before being transferred to the outcome one. The neurons' state in each one specifically impacts the neurons' state within the subsequent one. Within the occasion that the outcome layer does not abdicate the required outcome, the back-propagation technique is employed. It recursively computes the difference (i.e., error) between the actual input and the expected input layer by layer [17]. The error signal is propagated back through the original connection pathway, and the weights between neurons in each layer are adjusted to minimize the error.

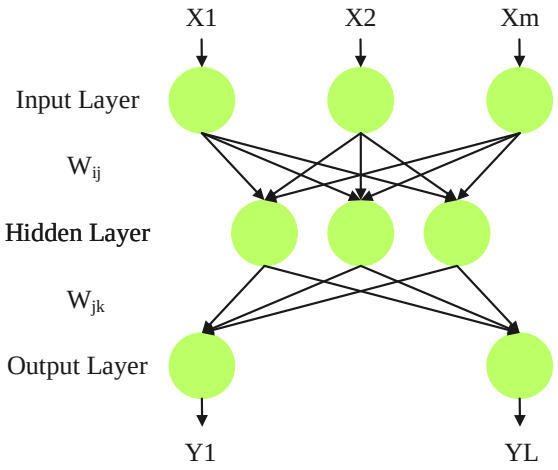


Fig. 1. Topological plan of the back-propagation neural system.

a) Foundation of preparing tests: Normalize the information within the evaluating measures (good quality, ordinary quality, light pollution and heavy pollution) of the six indicators listed in Table I, convert the specific values into [0,1] of the data in the interval, changes the absolute value of the physical system value into a relative value relationship. The urban residential environment quality grading metrics are used as training demos and intake to the network's intake hubs. Through the arbitrary number generation principles, 250 demos of every urban human settlement environmental quality index data were taken, and a total of 1000 samples were trained and modelled. MATLAB R2012b is used to establish the graphical user interface (GUI).

b) Neural network initialization: In the neural network model, we determine the hubs of the intake, hidden and outcome layers. Additionally, initializing the association weights among w_{ij} and w_{jk}; We also assign values to the hidden layer threshold a and output layer threshold b. These parameters are set along with the learning rate and neuron excitation function, select the input sequence (urban living environment quality evaluation index) and output sequence (urban living environment quality grading standard) of the system (X, Y) [18]. The hubs number within the covered up layer is evaluated utilizing an experimental equation.

$$Y_H = \sqrt{\theta_1 + \theta_2} \times P_{CCDP} \quad (11)$$

In Formula (11), θ_1 and θ_2 represent the number of neuron nodes in the input layer and the output layer, respectively.

c) Hidden layer output calculation: Using the input vector X, the association weights between the input layer w_{ij} and the hidden layer, as well as the hidden layer threshold a [19], we can compute the output of the hidden layer H; the calculation is shown in Formula (12), where f is the excitation function, and the excitation function is represented by the well-known Sigmoid function, as depicted in Formula (13).

$$H_j = Y_H \times f\left(\sum_{i=1}^{n} w_{ij} x_i - a_j\right), j = 1,2,\cdots,m \quad (12)$$

$$f(x) = \frac{1}{1+e^{-x}} \quad (13)$$

d) Output layer output calculation: Input according to the hidden layer H, hidden layer and output layer w_{jk} and output layer threshold b, calculate the predicted output of urban residential environment quality O, as shown in Formula (14).

$$O_k = f\left(\sum_{j=1}^{m} H_j \times w_{jk} - b_k\right), k = 1,2,\cdots,l \quad (14)$$

e) Error calculation: Considering the anticipated outcome O and desired outcome Y obtain estimation error e, as shown in Formula (15).

$$e_k = Y_k - O_k \quad (15)$$

f) Weight update: According to the prediction error$e$ to update connection weight w_{ij} and w_{jk}, the calculation formula is shown in Formula (16) and Formula (17). η is the learning rate.

$$w_{ij} = w_{ij} + \eta H_j\left(1 - H_j\right)x_i \sum_{k=1}^{l} w_{jk} e_k \quad (16)$$

$$w_{jk} = w_{jk} + \eta H_j e_k \quad (17)$$

g) Threshold update: According to the prediction error e to update threshold value a, b, the calculation formula is shown in Formula (18) and Formula (19), respectively.

$$a_j = a_j + \eta H_j\left(1 - H_j\right)\sum_{k=1}^{l} w_{jk} e_k \quad (18)$$

$$b_k = b_k + e_k \quad (19)$$

h) Judge whether the algorithm iteration ends: If the iteration is completed, it means that the training process of urban human settlements environmental quality assessment model based on back propagation neural network can be completed, and the show is established; In case the iteration isn't completed, return to the covered up layer yield evaluation part and begin a new preparing alteration procedure [20] until the method iteration is completed.

The evaluation model parameters mainly include the maximum number of training steps, performance parameters, the maximum number of confirmation failures, the number of hidden layer neuron nodes and other parameters [21]. These parameters can be manually modified as needed. At the same time, the initialization weight of the back-propagation neural network model is not unique. The final prediction result is the number of grades that appear most after the decision to repeat the training an odd number of times.

2) Optimization of back-propagation neural network algorithm: The back-propagation neural network exhibits robust capabilities in nonlinear mapping and complex logic operation ability, but the global search ability of the back-propagation neural network is relatively weak and is prone to local extreme values and slow convergence speed, which affects the backpropagation neural system's calculation efficiency and prediction accuracy. To enhance the global search capability of the back-propagation neural network within the assessment model for urban human settlement environmental quality, an improved genetic algorithm is employed for optimizing the network. Genetic algorithm is a search and optimization algorithm that simulates the natural evolution process. By simulating operations such as heredity, mutation and selection, the weight and structure of the network are gradually improved to make it better adapt to the evaluation task. In the process of optimization, the fitness function is defined, that is, the performance and fitness of the network are evaluated. By initializing the initial population composed of a group of genes, new individuals are generated by the operation of genetic algorithm, and they are evaluated according to the fitness function. Through iterative updating, individuals with high adaptability are gradually screened out. The improved genetic algorithm is used to optimize the back propagation neural network, which can improve the accuracy and generalization ability of the network by searching for better network weight and structure combination globally. The evaluation model of urban human settlement environmental quality can evaluate the urban environmental quality more accurately and provide a more scientific basis for urban planning, construction and management.

The conventional genetic algorithm often encounters challenges such as local optimization and slow convergence. To address this, an enhanced genetic algorithm with enhanced global search capability is proposed, building upon the traditional genetic algorithm. By employing an adaptive calculation strategy for crossover probability and mutation probability, the ability of the genetic algorithm to discover the global optimal solution is significantly improved. Compared with other evaluation methods, this improved genetic algorithm can overcome the problems that traditional genetic algorithm is easy to fall into local optimum and slow convergence, and has stronger global search ability. By adopting adaptive crossover probability and mutation probability calculation strategy, the improved genetic algorithm can flexibly adjust the parameters of genetic operation to adapt to the characteristics and difficulties of different problems. It can improve the extensive search ability of the algorithm in the solution space and reduce the dependence on the initial solution, thus effectively avoiding the problem of falling into the local optimal solution and being unable to find the global optimal solution.

a) Design of chromosome coding: The dimension of the chromosome gene vector in real coding is determined by the number of weights and thresholds present in the back-propagation neural network [22], and the formula is as follows:

$$X_i = (w_{11},\cdots,w_{ms},w_{11},\cdots,w_{sn},a_1,\cdots,a_s,b_1,\cdots,b_n) \quad (20)$$

b) Determination of fitness function: In the genetic algorithm, the fitness value of the individual is an important indicator to evaluate the excellent performance of the individual [23]; assuming that the fitness value of the ith individual is F_i, the back-propagation neural network yields a mean square error of $M_{SE}(X_i)$, the fitness function is taken as:

$$F_i = M_{SE}(X_i) \times a_j \times b_k \quad (21)$$

c) Select the design of the operation: Discarding the roulette wheel method in the traditional genetic algorithm, the formula of the probability of each individual being selected is:

$$p_i = \frac{\frac{l}{F_i}}{\frac{\sum_{i=1}^{N} l}{F_i}}, i = 1,2,\cdots,N \quad (22)$$

In Formula (22): $\frac{l}{F_i}$ represents the fitness value of the ith individual, l represents the adjustment factor.

d) Design of cross operation: An adaptive crossover probability is proposed for individuals with poor performance, appropriately increase the crossover probability of the individual to optimize its gene structure; For individuals with good performance, the crossover probability should be appropriately reduced to avoid damaging excellent genes. In addition, in order to ensure population diversity and the algorithm exhibits a fast search speed in the initial stage, followed by enhanced local search capability in the later stage, ultimately leading to convergence and avoiding oscillation at the extreme point [24]; this crossover probability should also be reduced with the iteration of the algorithm.

The calculation formula of individual crossover probability is:

$$p_{ci} = P\left(Pcmin_{cmax} \times \frac{t \times F_{min}}{T \times (F + F_{min}())}\right)_{cmax} \quad (23)$$

In Formula (23): t represents the current iteration number of the algorithm, T represents the total number of iterations of the algorithm, p_{ci} is the probability of the ith individual at the t-th crossing, F_{min} indicates the fitness value of the individual with the best performance of the population, P_{cmax} is the maximum crossing probability, which is 0.6, P_{cmin} represents the minimum crossing probability, which is 0.3.

Chromosome i X_i and the j chromosomes X_j on k and the bit crossing formula is:

$$\begin{cases} X_i^k = (1-\partial)X_i^k + \partial X_j^k \\ X_j^k = (1-\partial)X_j^k + \partial X_i^k \end{cases} \quad (24)$$

In Formula (24): ∂ represents a random number, and $0 \le \partial \le 1$.

e) Design of mutation operation: The mutation operation is implemented to preserve the global search capability in the initial iterations of the algorithm while also ensuring local search capability and stability during the later stages [25]. Consequently, the design assigns equal probabilities for individual mutation and crossover operations. These probabilities are determined based on the individual fitness value and the number of algorithm iterations [26]. The calculation formula is:

$$p_{mi} = P\left(Pmmin_{mmax} \times \frac{t \times F_{min}}{T \times (F_i + F_{min}())}\right)_{mmax} \quad (25)$$

In Formula (25): p_{mi} is the probability of the ith individual at the time of t variation, P_{mmax} represents the maximum probability of variation, 0.005, P_{mmin} represents the minimum probability of variation. Gene j on chromosome i X_i^j and the variation formula of is:

$$X_i^j = \begin{cases} X_i^j + \left(X_i^j - X_{i\max}^j\right) f(t), \lambda > 0.5 \\ X_i^j + \left(X_{i\min}^j - X_i^j\right) f(t), \lambda \le 0.5 \end{cases} \quad (26)$$

In Formula (26): X_{imax}^j express the upper bound of gene X_i^j, X_{imin}^j express the lower bound of gene X_i^j, t represents the current number of iterations, λ represents a random number, and $-1 \le \lambda \le 1$.

f) Determination of population size and iteration number: Since there are many undetermined parameters in the back-propagation neural network, the selected population size is 100 to ensure global optimization; The number of iterations is 500 to ensure the complete convergence of the algorithm. Fig. 2 illustrates the flowchart of the back-propagation neural network model that incorporates an improved genetic algorithm.

The steps to optimize the back-propagation neural network using the improved genetic algorithm are as follows:

Step 1: Pre-process the urban human settlements' environmental quality evaluation index data, determine the network structure of the back-propagation neural network, and determine the coding method;

Step 2: Determine the fitness function and repeatedly select, cross, and mutate the initial population until the fitness value of a chromosome reaches the preset standard;

Step 3: Compute the output of each node in the hidden layer and the output layer sequentially, followed by the calculation of the output error for each node in the output layer.

Step 4: If the output error does not meet the accuracy requirements, update the weights and thresholds of each layer based on the error back-propagation process, and utilize the updated weights and thresholds to calculate the output error. Repeat this process iteratively till the output error satisfies the desired precision criteria, and then the teaching is over.

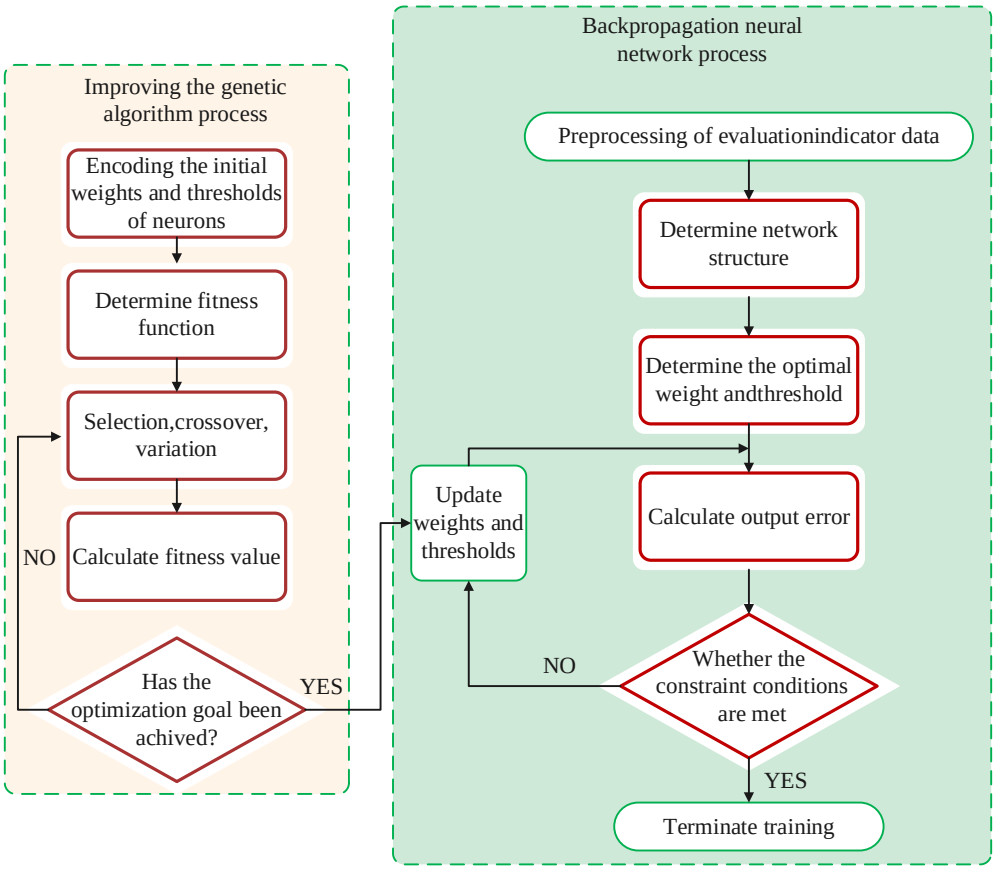


Fig. 2. Flow chart of back-propagation neural network model on the basis of improved genetic algorithm.

III. Experimental Results

This paper studies the urban residential environment quality assessment algorithm on the basis of the back-propagation neural network. In order to verify the application effect of this method in the urban residential environment quality assessment, a region is selected as the research object. The region contains nine cities, which are represented by A-I as shown in Fig. 3. The method in this paper is used to evaluate the urban residential environment quality in the study area. The results are as follows:

A. Data Source and Processing

The data used in the process of qualitative assessment of people's habitation of the research object using this method mainly include environmental data, population data, vector water network distribution data and basic geographic data. In the process of data collection, relevant data are obtained through various channels such as government statistics bureau, environmental monitoring department, professional research institutions and open data platforms. These data sources provide various indicators of urban environmental quality, such as air quality, water quality, noise level and green space coverage rate. In order to ensure the reliability and validity of the data, the data are verified and calibrated. Through data sampling and the integration of multiple data sources, the representativeness and reliability of data are improved, and residents' subjective feelings and evaluation on the quality of urban human settlements are obtained by means of field surveys and questionnaires, so as to increase the comprehensiveness and objectivity of the research. The data sources and corresponding processing processes are shown in Table II.

Based on the processing results of various types in Table II, the evaluation index data used in the process of urban human settlements' environmental quality assessment is obtained.

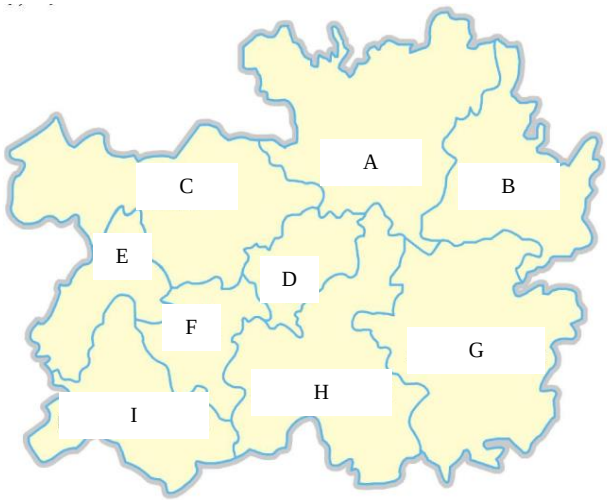


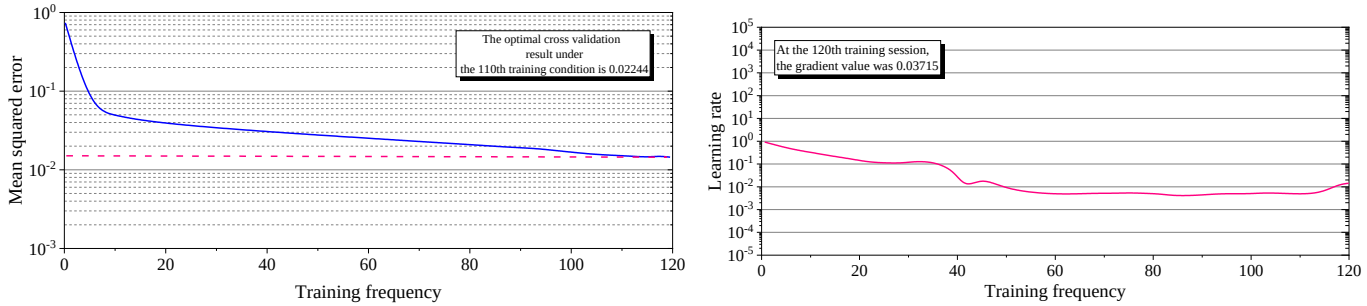
Fig. 3. Overview of the study area.

TABLE II. DATA SOURCES AND CORRESPONDING PROCESSING

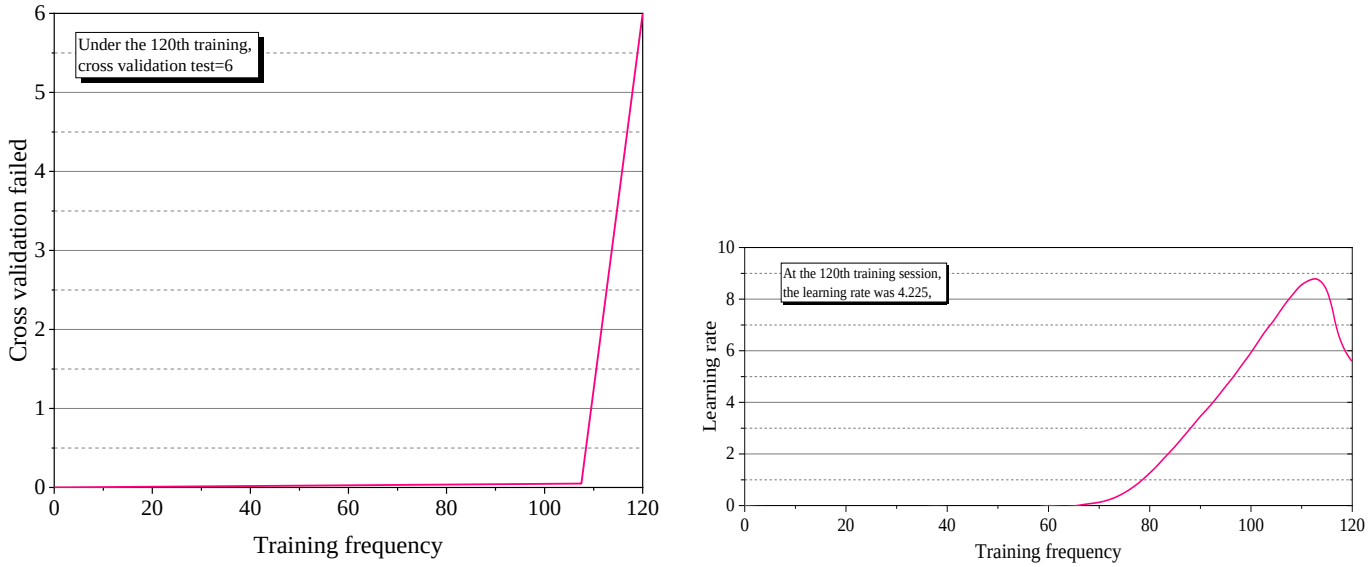
Data type	Relevant content	Main sources	Processing technology
Environmental data	Temperature, relative humidity, precipitation, etc	Regional Meteorological Station Resource Library	For various types of data, the Kriging method, spline method, and inverse gradient distance square method are used for interpolation processing to obtain regional meteorological element layers
	Digital Elevation Model	Global GTO-PO30	Adopting the conic projection of double standard weft lines with positive axis area to obtain 0.5km × Digital elevation model map of 0.5km area
	NDVI	Earth Science Data Sharing Center	-
Demographic data	Demographics	Regional Statistical Yearbook	-
Basic geographic data	Traffic distribution vector data	Data Center of the Chinese Academy of Sciences	Build a 0.5km × 0.5km grid, determine the water network density of the grid through spatial analysis, and convert it to a 0.5 km × Grid scale of 0.5 km
	Distribution data of social settlements	Earth Science Data Sharing Center	On the basis of the data gathered using the sharing centre and combined with the latest map implementation comparison and optimization
Economic data	Various economic data within the region	Regional Statistical Yearbook	-

B. Training and Parameter Adjustment

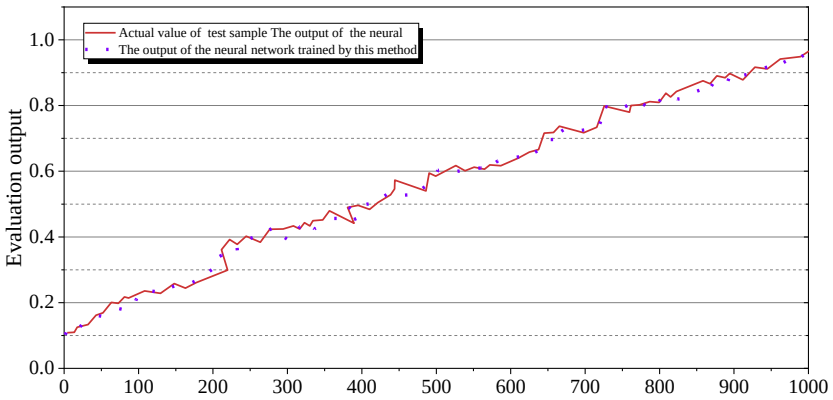
By utilizing the Nprtool toolbox in MATLAB R2012b software, the neural network is trained and evaluated, the data import, data pre-processing, establishment and training of neural networks, and the use of error mean square and confusion matrix are automatically completed to analyze the modelling effect of back-propagation neural networks used in this evaluation method. The back-propagation neural network is constructed with two-layer feedforward architecture. The transmision function employed in the covered up one is the sigmoid, while the entire network is trained using the variable gradient algorithm. 1000 training demos are consumed for training and learning. In which, the training set consists of 700 samples, the verification set comprises 150 samples, and the rest 150 demos are allocated to the set of test. After several iterations, the optimal number of neurons in the hidden layer for the back-propagation neural network is determined as 10, with a training target error set to 0. To mitigate overfitting, the maximum number of training iterations for the neural network is set at 1000. The maximum number of validation set failures is set to 6. The training process and parameters are shown in Fig. 4.



(a) Mean squared error Gradient analysis.



(b) Cross-validation failed learning rate.



(c) Output accuracy.

Fig. 4. Parameter adjustment process of back-propagation neural network training set.

It can be seen from Fig. 4 that the back-propagation neural network used in this method can stop training after 110 times of training, and the error is 0.02244. Among them, it can be seen that the gradient of the back-propagation neural network is declining in the training process, which conforms to the characteristics of back-propagation neural network error back-propagation. The accuracy of the evaluation results obtained by substituting the training data back into the neural network model is 96.3%.

C. Evaluation Results

The trained back-propagation neural system is for evaluating the quality of urban human settlements in nine cities within the study object, and the evaluation results are shown in Table III.

According to the analysis of Table III, the urban residential environment quality of nine cities in the study object was evaluated using the method in this paper, and the results showed that the quality of the urban residential environment in G city was excellent; The qualitative assessment of people's habitations in cities A, D, E and F is average; However, the qualitative assessment of people's habitations in cities B, C, H and I is slightly polluted. At the same time, the obtained results from this approach align perfectly with the actual conditions of each city, indicating the method's capability to complete the qualitative assessment of people's urban habitations accurately.

TABLE III. EVALUATION RESULTS OF THE STUDY OBJECT

City number	Evaluate Results	Actual situation
A	Average quality	Average quality
B	Slightly polluted	Slightly polluted
C	Slightly polluted	Slightly polluted
D	Average quality	Average quality
E	Average quality	Average quality
F	Average quality	Average quality
G	Superior in quality	Superior in quality
H	Slightly polluted	Slightly polluted
I	Slightly polluted	Slightly polluted

D. Analysis of the Improvement Effect of Urban Residential Environment Quality

This method is used to complete the qualitative assessment of people's habitations in all cities within the study object. The assessment outcomes gathered from the algorithm are utilized to optimize the quality of human settlements in various cities. Compare the fluctuation of human settlements' environmental quality assessment grades in different cities before and after the assessment using this method, and the results are shown in Fig. 5.

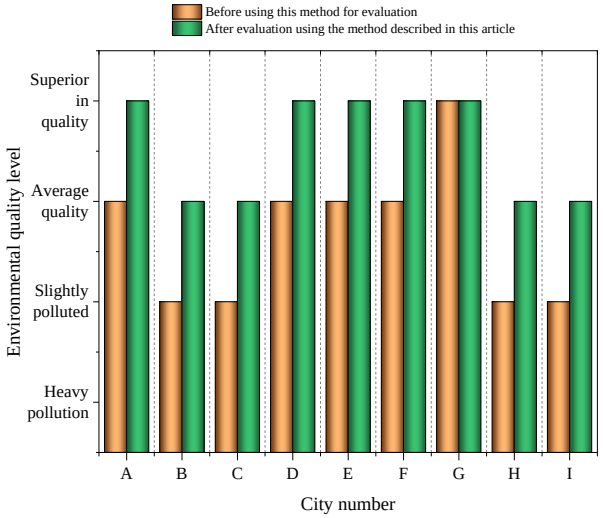


Fig. 5. Fluctuation of the assessment levels of people's habitations quality in different cities.

From the results in Fig. 5, it can be seen that after evaluating the quality of human settlements in different cities using this method, most cities have shown a certain degree of improvement after environmental optimization, further verifying the effectiveness and feasibility of this method in evaluating the quality of urban human settlements. Specifically, except for the relatively high initial evaluation results of City G, which do not require environmental optimization, all eight other cities have achieved further improvement after environmental optimization. This means that using the method presented in this article can accurately

identify areas that need improvement and provide corresponding optimization strategies, thereby improving the quality of urban living environment. This result further demonstrates that the method proposed in this paper has good evaluation performance and provides effective solutions for the quality of human settlements in different cities. It helps urban planners and policy makers to better understand the current state of human settlements in cities and take corresponding measures to improve their quality of life.

IV. Discussion

In the study of the evaluation method of urban human settlements based on back propagation neural network, this method is applied to evaluate and improve the urban environmental quality. By constructing an appropriate index system and training a back propagation neural network evaluation model with good generalization ability, the situation of urban human settlements can be accurately evaluated and scientific basis can be provided for urban planning and management.

Through this study, it is found that the evaluation method of urban human settlements quality based on back propagation neural network has some remarkable advantages. Firstly, this method can integrate multiple environmental factors and socio-economic factors, and comprehensively consider the comprehensiveness and complexity of urban human settlements. Secondly, the back propagation neural network has strong fitting ability and generalization ability, and performs well in dealing with nonlinear relations and predicting future trends. Finally, this method can be flexibly adjusted and optimized according to the actual needs and data, and has high operability and adaptability.

V. Conclusion

This research focuses on the qualitative assessment of people's urban habitations through a backpropagation neural system method. An assessment model is constructed using the back-propagation neural network, and an improved genetic algorithm is introduced to address the issues of local optimization and slow convergence commonly associated with the back-propagation neural network. Additionally, the optimization of the initial weights and thresholds of the network is addressed, ensuring the stability of the algorithm and overcoming the shortcomings of the back-propagation neural network algorithm. Through experiments, the following conclusions are obtained:

1) After 110 times of training, the BP neural network used in this method can stop training, and the error is 0.02244. According to the characteristics of back propagation neural network, the accuracy of evaluation results obtained by substituting training data into neural network model is 96.3%.

2) This method can accurately evaluate the quality of urban human settlements.

3) The method in this paper has good evaluation performance for the quality of human settlements, which can effectively promote the improvement of the quality of urban human settlements.

There are some limitations in the study of this method in practice, such as the lack of availability and completeness of data. Although there are many data sources to choose from, the data of some urban environmental indicators may be lacking or incomplete, which affects the accuracy and reliability of the evaluation model. Future research can solve this problem by improving the coverage and quality of data collection and monitoring systems, or by integrating multiple data sources. At the same time, the construction of evaluation index system is still a challenge. At present, there are still subjectivity and limitations in the selection of evaluation indicators and the distribution of weights, and there is still a lack of unified standards and systems. Further research can explore a more scientific, comprehensive and objective evaluation index system, taking into account social and economic factors, residents' subjective feelings and other factors, so as to evaluate the quality of urban human settlements more comprehensively.

Data Availability

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation."

Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work."

Acknowledgements

This work supported by Jilin Provincial Department of Housing and Construction.

2022 Science and Technology Project Plan (Urban Agglomeration and Regional.

Green development) "Urban Green Human Settlement Space Unit Construction.

Research" project number: 2022-KQ-01.

Competing of Interests

The authors declare no competing of interests.

Authorship Contribution Statement**.** Siyuan Zhang: Writing-Original draft preparation.

Conceptualization, Supervision, Project administration.

Wenbo Song: Language review, Methodology, Software.

References

[1] Z. F. He, J. Yu, Y. Chen, J. H. Wen, and Z. E. Yin, “Potential ecological suitability evaluation of urban construction land: A case study in Baoshan District, Shanghai,” Resources Science, vol. 42, no. 03, pp. 558–569, 2020.

[2] Z. You, Z. M. Feng, Y. Z. Yang, H. Shi, and P. Li, “Evaluation of human settlement environmental suitability in Tibet based on gridded data,” Resour. Sci, vol. 42, pp. 394–406, 2020.

[3] A. Aslam et al., “Mitigation of particulate matters and integrated approach for carbon monoxide remediation in an urban environment,” J Environ Chem Eng, vol. 9, no. 4, p. 105546, 2021.

[4] P. Sestras, Ștefan Bilașco, S. Roșca, B. Dudic, A. Hysa, and V. Spalević, “Geodetic and UAV monitoring in the sustainable management of shallow landslides and erosion of a susceptible urban environment,” Remote Sens (Basel), vol. 13, no. 3, p. 385, 2021.

[5] W. Peng, Y. Sun, C. Liu, and D. Liu, "Study on Urban Land Ecological Security Pattern and Obstacle Factors in the Beijing–Tianjin–Hebei Region," Sustainability, vol. 15, no. 1, p. 43, 2022.

[6] H. Staats and R. Swain, "Cars, trees, and house prices: Evaluation of the residential environment as a function of numbers of cars and trees in the street," Urban For Urban Green, vol. 47, p. 126554, 2020.

[7] N. Fomina and A. Lugovskoy, "Assessment of the quality (comfort) of the living environment in urban conditions," in E3S Web of Conferences, EDP Sciences, 2020, p. 09007.

[8] C. M. Longhini et al., "Environmental quality assessment in a marine coastal area impacted by mining tailing using a geochemical multi-index and physical approach," Science of The Total Environment, vol. 803, p. 149883, 2022.

[9] H. Sarkheil, S. Rahbari, and Y. Azimi, "Fuzzy-Mamdani environmental quality assessment of gas refinery chemical wastewater in the Pars special economic and energy zone," Environmental Challenges, vol. 3, p. 100065, 2021.

[10] K. Chytrý, W. Willner, M. Chytrý, J. Divíšek, and S. Dullinger, "Central European forest–steppe: An ecosystem shaped by climate, topography and disturbances," J Biogeogr, vol. 49, no. 6, pp. 1006–1020, 2022.

[11] M. Ntawubizi, Y. B. Niyonzima, L. Rydhmer, C. D. Hirwa, M. Manzi, and E. Strandberg, "PSX-34 Late-Breaking Abstract: Effect of Genotype and Temperature-Humidity Index (THI) on milk yield of Ankole and its crossbreeds in Rwanda," J Anim Sci, vol. 98, no. Supplement_4, p. 352, 2020.

[12] N. B. Fitzgerald and J. B. Kirkpatrick, "Air temperature lapse rates and cloud cover in a hyper-oceanic climate," Antarct Sci, vol. 32, no. 6, pp. 440–453, 2020.

[13] J. Wang, J. Wu, S. Zhan, and J. Zhou, "Records of hydrological change and environmental disasters in sediments from deep Lake Issyk - Kul," Hydrol Process, vol. 35, no. 4, p. e14136, 2021.

[14] G. Mezger, M. G. del Tánago, and L. De Stefano, "Environmental flows and the mitigation of hydrological alteration downstream from dams: The Spanish case," J Hydrol (Amst), vol. 598, p. 125732, 2021.

[15] F. Li, S. Fang, Y. Shen, and D. Wang, "Research on graphene/silicon pressure sensor array based on backpropagation neural network," Electron Lett, vol. 57, no. 10, pp. 419–421, 2021.

[16] B. Li, Y. Liu, and L. Lai, "Computational model of grid cells based on back - propagation neural network," Electron Lett, vol. 58, no. 3, pp. 93‑96, 2022.

[17] Y.-W. Kao and H.-H. Chen, "Associated Learning: Decomposing End-to-End Backpropagation Based on Autoencoders and Target Propagation," Neural Comput, vol. 33, no. 1, pp. 174–193, 2021.

[18] B. Kenchappa and K. Shivakumar, "An acoustic evaluation of light weight granular porous materials for urban air mobility applications," J Acoust Soc Am, vol. 148, no. 4, p. 2455, 2020.

[19] J. Zhao, C. Zhang, L. Min, N. Li, and Y. Wang, "Surface soil moisture in farmland using multi-source remote sensing data and feature selection with ga-bp neural network," Nongye Gongcheng Xuebao/Transactions of the Chinese Society of Agricultural Engineering, vol. 37, no. 11, pp. 112–120, 2021.

[20] H. T. Li and Z. D. Shao, "Marine Sediment Quality Evaluation Based on IGWO and BP Neural Network [J]," Computer Simulation, vol. 37, no. 8, pp. 344–347, 2020.

[21] S. Wang, T. Wu, K. Wang, and T. Sarkodie-Gyan, "Ferrograph analysis with improved particle segmentation and classification methods," J Comput Inf Sci Eng, vol. 20, no. 2, p. 021001, 2020.

[22] Y. P. V. Acuna and Y. Sun, "An efficiency - improved genetic algorithm and its application on multimodal functions and a 2D common reflection surface stacking problem," Geophys Prospect, vol. 68, no. 4, pp. 1189‑1210, 2020.

[23] Y. Su, S. Jin, X. Zhang, W. Shen, M. R. Eden, and J. Ren, "Stakeholder-oriented multi-objective process optimization based on an improved genetic algorithm," Comput Chem Eng, vol. 132, p. 106618, 2020.

[24] J. Tang, Y. Yang, W. Hao, F. Liu, and Y. Wang, "A data-driven timetable optimization of urban bus line based on multi-objective genetic algorithm," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 4, pp. 2417–2429, 2020.

[25] H. Wang, Z. Fu, J. Zhou, M. Fu, and L. Ruan, "Cooperative collision avoidance for unmanned surface vehicles based on improved genetic algorithm," Ocean Engineering, vol. 222, p. 108612, 2021.

[26] Y. Wang, Z. Wu, G. Guan, K. Li, and S. Chai, "Research on intelligent design method of ship multi-deck compartment layout based on improved taboo search genetic algorithm," Ocean Engineering, vol. 225, p. 108823, 2021.

Basketball Motion Recognition Model Analysis Based on Perspective Invariant Geometric Features in Skeleton Data Extraction

Jiaojiao Lu
Sport Institute, Yantai Institute of Science and Technology, Yantai 265600, China

***Abstract*—The study proposes a recognition method based on skeleton data to address the basketball action recognition, especially those posed by viewpoint changes in videos. The key of this method is to extract geometric features of viewpoint invariance and combine them with spatio-temporal feature fusion techniques. In addition, the study constructs a dynamic topological map of the human skeleton based on long and short-term neural networks to improve the model performance. The experimental results showed that the research method had an average accuracy of 97.85% for Top-5 metrics on the Kinetics dataset and 97.82% for Top-5 metrics on the NTU RGB+D dataset. It is significantly better than the other three state-of-the-art methods. According to the experimental results, it achieves efficient and stable basketball action recognition, which is significantly superior to existing methods. This research not only provides a more efficient method for basketball motion recognition, but also provides valuable references for other sports action recognition fields.**

***Keywords*—Skeleton data; perspective invariance; geometric features; basketball recognition; spatio-temporal feature fusion**

I. Introduction

Basketball, as a global sport, not only attracts countless spectators and enthusiasts, but also becomes the focus of in-depth discussions in academic research and business. With the development of technology, analyzing the importance of basketball matches from a technical perspective is becoming increasingly prominent [1]. Among them, basketball action recognition technology plays an important role in player skill analysis and training assistance [2]. However, traditional visual basketball action recognition methods rely on the analysis of continuous video frames. They are limited by background noise, complex environments, and especially the change of camera viewpoints, which makes the action recognition face great challenges [3]. Skeletal data provide a more abstract and concise representation of movements than traditional methods, and are more effective in resisting the effects of perspective changes [4]. However, due to the poor fusion effect of spatio-temporal features in skeleton data, the accuracy of existing recognition models is not ideal [5]. To address this problem, the study proposes a novel basketball sports recognition model, which is based on viewpoint invariant geometric features and aims to improve the accuracy and robustness of skeleton data recognition. In addition, the study also explores the method of constructing a dynamic topological map of the human skeleton to further improve the model performance. This study consists of five parts. The first part is an overview of this study. The second part is a summary of relevant research. The third part has two sections. Firstly, a dynamic topology map of the human skeleton is constructed. The second section introduces a basketball motion recognition method based on perspective invariant geometric features. In the fourth section, the proposed method is tested on a dataset and in a real environment. The results are analyzed. The fifth section is a summary and outlook for this study. This study develops a new basketball action recognition system. It not only utilizes skeleton data more efficiently, but also shows significant improvement in both recognition accuracy and robustness. It is expected that this study can provide strong technical support for basketball game analysis, teaching assistance and other related applications.

II. Related Works

In the action recognition, feature extraction and fusion are the foundation of accurate recognition. Therefore, a large number of scholars have conducted in-depth research on this field [6]. Zhou W et al. proposed a novel feature fusion network to improve the performance of object detection under low light and uneven lighting conditions. This study applied a cross modal fusion module to fuse the corresponding size features of target detection and thermal modes. Then, the bidirectional reverse fusion module was used to achieve bidirectional fusion of foreground and background information. In the experimental results, the proposed feature fusion network was superior to other advanced methods [7]. Zhang X et al. proposed a network that combined multi-scale hierarchical feature fusion and mixed convolutional attention to solve the single image defogging in image recognition. By fusing multi-scale layered features, the haze level and image structure information were accurately estimated, resulting in the restored image containing less residual haze. According to the experimental result, it exceeded the most advanced defogging algorithm [8]. Choi H et al. proposed a new multimodal image feature fusion module to solve the transmission line inspection. The output of the multi branch feature extraction block was aggregated into an attention vector in the channel attention block. Each input feature was recalibrated. According to the experimental result, it was superior to the single mode input [9].

With the emergence of various optimization techniques, the improved action recognition models are gradually receiving attention. To solve the distinguishable feature extraction in skeleton action recognition model, Song Y F et al.

embedded an advanced separable convolutional layer into the Multiple Input Branches network. The research constructed an efficient Graph Convolutional Network (GCN) baseline. A composite extension strategy was designed to synchronously extend the width and depth. In the NTU 60 dataset testing, the accuracy reached 92.1% [10]. Li M et al. proposed a symbiotic graph neural network for 3D skeleton motion recognition. The network structure included a basic part, an action recognition part, and a motion prediction part. In the backbone network, a multi-scale GCN based on joints and parts was used to extract key features. Compared with the current methods, the method had better performance on all four datasets [11]. Li C et al. proposed a new temporal and spatial recalibration method to address the complex changes in skeletal joints in motion recognition. The research constructed a novel temporal attention mechanism based on residual learning to calibrate the frames of skeleton data. Compared with the most advanced methods, the research method significantly improved performance. It had the best results on six action recognition datasets [12].

In summary, there have been many studies in the action recognition, especially in feature extraction and fusion. However, most of these studies have focused on improving the performance of object detection under specific environmental conditions, specific problems in image processing, or specific applications. These approaches are often not applicable to complex sports scenarios, especially in basketball action recognition. They fail to adequately address the challenges posed by changes in perspective [13]. This research fills this gap by proposing a new basketball motion recognition model. It focuses on extracting viewpoint-invariant geometric features from skeleton data and incorporates spatio-temporal feature fusion techniques. Compared with existing studies, the innovation of the study is follows. It proposes an action recognition method for complex sports scenarios, which specifically addresses the viewpoint invariance in basketball. Further, the study effectively improves the accuracy and robustness of the model by constructing a dynamic topological map of the human skeleton based on long and short-term neural networks. In contrast to existing research in GCNs, symbiotic graph neural networks, and temporal re-space recalibration methods, the study focuses on combining action recognition with perspective invariance. It is an area rarely covered in existing research. As a result, the study not only makes significant improvements in the efficiency and accuracy of basketball action recognition, but also provides a new perspective to explore the action recognition problem. It provides a valuable reference for research in this area.

III. Construction of Basketball Motion Recognition Model Based on Perspective Invariant Geometric Features in Skeleton Data Extraction

To achieve more efficient and stable basketball motion recognition results, the study optimizes the model from two aspects. Firstly, based on short-term and short-term memory neural networks, a dynamic topology map of the human skeleton is constructed to provide more accurate data support for the recognition model. Then, a new spatio-temporal feature fusion method is proposed based on the perspective invariant geometric features extracted from skeleton data. On this basis, the research constructs a basketball motion recognition model.

A. Dynamic Topology Map Construction of Human Skeleton

Basketball action recognition based on human skeleton data aims to identify the types of actions represented by human skeleton time series. Traditional recognition methods have two categories. One is based on manual features, which mainly capture the dynamic relationships of joints through manual design features. The second is based on deep learning. It conducts end-to-end modeling with recurrent neural networks to capture joint information [14]. In practical scenarios, an action consists of multiple video frames. It is difficult to accurately display the dependency relationship between joints and bones by manually creating a topology map. The individual training and parameters for each video frame not only require a large amount of computation, but may also lead to catastrophic forgetting. Therefore, the research adopts a continuous learning method. The Long Short-Term Memory (LSTM) is utilized to dynamically construct the topology map of the human skeleton to improve the recognition performance of the basketball motion recognition model. The human skeleton sequence is composed of continuous human skeleton frames. Each frame is a set of joint coordinates and coordinates confidence. The definition of the human skeleton sequence is shown in Eq. (1).

$$VT = \{VT_1, VT_2, ..., VT_T\} \tag{1}$$

In Eq. (1), T represents the number of human skeleton frames. VT_i represents the i-th human skeleton frame in the human skeleton sequence. VT_i is composed of joint points containing spatio-temporal information, as shown in Eq. (2).

$$VT_i = \{V_1, V_2, ..., V_N\} \quad (1 \le i \le T) \tag{2}$$

In Eq. (2), N represents the joint points in the human skeleton frame. The i-th joint point V_i in the human skeleton frame is shown in Eq. (3).

$$V_i (1 \le i \le N) = (x_i, y_i, score_i) \tag{3}$$

In Eq. (3), (x_i, y_i) represents the position information of the joint point. $score_i$ represents the confidence information of the joint position. Therefore, the dimension of the human skeleton sequence is (T, N, C). C is the position vector dimension. The dynamic topology diagram of the human skeleton is shown in Fig. 1.

In data preprocessing, firstly, normalize the joint feature vectors. Then, convert skeleton sequences from different positions to the same position to promote convergence. A relationship graph is a many to many graph structure that exists between nodes. Convert multiple diagrams into a set of relationship triplets, as shown in Eq. (4) [15].

$$\{(u, r, v)\} \subseteq V \times E \times V \tag{4}$$

In Eq. (4), u and v are entities. r represents a relationship between entities. Therefore, encode multiple human bone sequence datasets into relational triplet sequence

datasets. Based on the multi relationship features of human skeleton data, Eq. (5) defines a sequence of triples.

$$RT = \left\{ (VT_i, r_{ij}, VT_j) \middle| 1 \le i \le T, 1 \le j \le T \right\} \tag{5}$$

After obtaining the relational triplet dataset, the features need to be decoupled. The feature decoupling module extracts entities and feature embedding vectors from a triplet dataset. Then, based on the association between independent joints in each action, the video frames of the triplet sequence are decomposed into multiple features to learn the embedding vectors of the entities. At the same time, encode the action types to learn the embedding vectors of relationship features. Entity feature decouples a single video frame into multiple features. Then the embedded feature vectors of each joint feature are learned separately. Therefore, the entity $VT_i \in V$ represented by each single video frame is transformed into a set of multiple independent joint points, as shown in Eq. (6).

$$VT_L = \left[V_1, V_2, ..., V_N \right] \tag{6}$$

In Eq. (6), $VT_L \in R^d$. d represents the vector dimension. After performing one-hot encoding on the action type, the study uses embedded feature vectors to represent action analogies, thereby constructing a dynamic topology map. The vectors obtained from the feature decoupling module are used to construct skeleton topology maps for different action categories. Firstly, the K-means is applied to cluster the relationship feature vectors representing action categories. Then the dataset is grouped based on the clustering center. Next, based on the action categories encoded by one-hot, combined with the attention mechanism and the partial update strategy, the study constructs a topology diagram for each type of action. Eq. (7) defines the i-th training set.

$$T_i = \{ (u_1^{T_i}, u_1^{T_i}, v_1^{T_i}), ..., (u_m^{T_i}, u_m^{T_i}, v_m^{T_i}) \} \tag{7}$$

In Eq. (7), m represents the instance number of T_i. For a skeletal relationship triplet, the research uses the attention mechanism to extract the relevant joint features of continuous video frames VT_i and their relationship r_{ij}. Triple (VT_i, r_{ij}, VT_j) is assigned N attention weights. Eq. (8) shows the importance α_r^i of the i-th joint point in the current relationship r.

$$\alpha_r^i = \frac{exp(\alpha_r^i)}{\sum_{j=1}^{N} exp(\alpha_r^j)} \tag{8}$$

In Eq. (8), α_r^j represents the importance of the j-th joint point in the current relationship r. Then, the research adopts the video frame VT_i with the highest attention weight and the first i joint points. Different features are used to construct a relationship topology diagram for this action category, as shown in Fig. 2.

In Fig. 2, a single skeleton frame contains 18 embedded feature vectors. The red leg nodes are the first six joint features most relevant to this action. The model experiences forgetting when updating parameters. Therefore, the study introduces a partial update strategy to dynamically adjust the topology map. When connecting a new skeleton relationship triplet, the model first identifies the parts related to the new data in the existing skeleton relationship graph. However, this only updates highly relevant features. Due to the complex connections between nodes, new data may affect multiple existing nodes. Therefore, activate the nodes directly or indirectly connected to it. Based on feature similarity, perform selective updates and further optimize computational efficiency. Finally, perform local updates on joint features that are highly similar to the new data.

B. Basketball Movement Recognition Based on Perspective Invariant Geometric Features

In modern basketball, numerous factors can easily affect target action recognition, such as local occlusion, rapid movement, and noise caused by inherent unstable factors in cameras [16]. These factors can cause the collected joint points to shake, resulting in a large amount of noise. Fig. 3 displays the schematic diagram of human joint shaking.

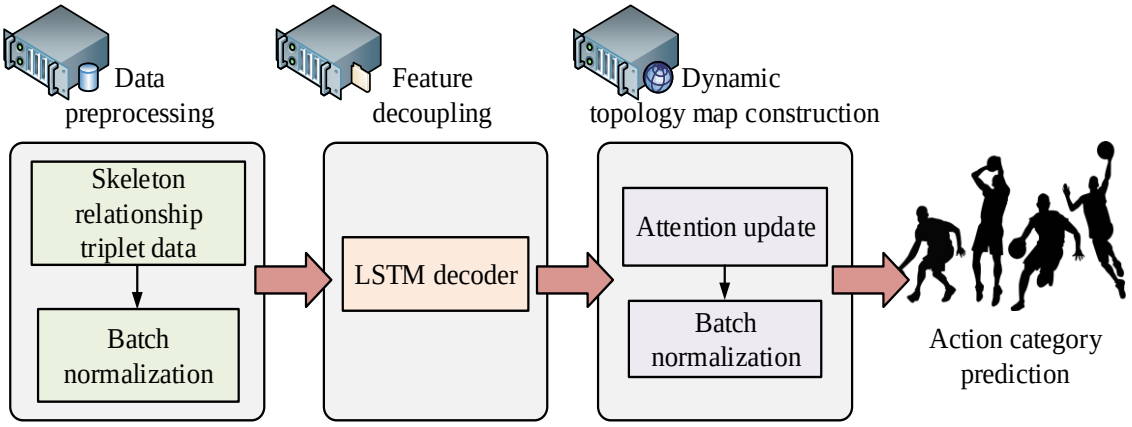


Fig. 1. The Process of human skeleton motion recognition based on dynamic topology graph.

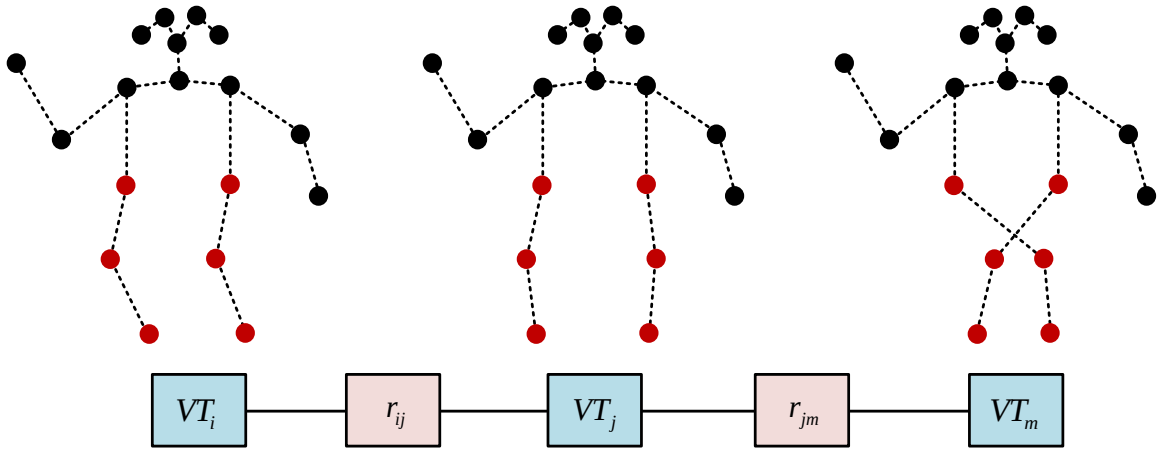


Fig. 2. Schematic diagram of feature encoding.

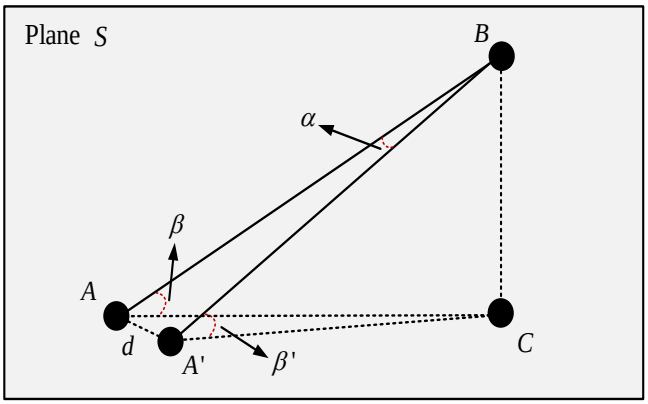


Fig. 3. Schematic diagram of human joint shaking.

In Fig. 3, when joint point A experiences shaking and moves to A', the shaking path is a periodic rotation around the joint BC. The rotation radius is AC, and the angle is α. The joint point A or conventional geometric characteristics used as network input may lead to system stability issues [17]. In addition, in cross perspective testing, apply two Kinect devices to calculate the three-dimensional coordinates of human bones. This calculation uses different projection centers, which may also introduce interference [18]. To address these challenges, a few feature representations that are not sensitive to three-dimensional spatial transformations of skeleton data are proposed. It is named the Planes of 3D Joint Motions Vector (P3DJMV). Cosine similarity is the primary distance comparison method for this research, primarily because it focuses on measuring directional similarity between vectors rather than the size. This property is particularly important for basketball motion recognition because the study focuses more on capturing and comparing features of the motion patterns rather than the absolute size of the action. In addition, cosine similarity typically performs well when dealing with high-dimensional data, which is particularly useful for research application scenarios that analyze complex motion data. Since it is based on directional similarity, cosine similarity naturally ignores differences in the size of the data. When comparing, it is possible to fairly handle different sizes of motion patterns. P3DJMV uses the cosine angle between vector $\underset{AB}{\rightarrow}$ and vector $\underset{AC}{\rightarrow}$ as the feature descriptor for skeleton data. In this way, even if point A shakes to point A', it will not follow the shaking due to point A shaking. In the plane S composed of A, B, and C, the angles β and β' are approximately equal. To explore the spatial variation of joint points, a feature representation based on node momentum (FRNM) is also designed, as shown in Fig. 4.

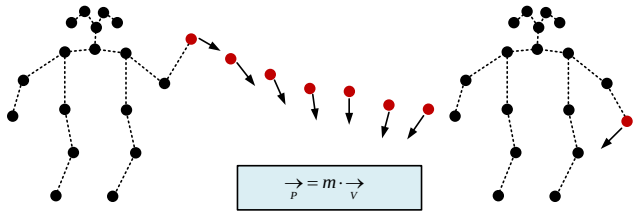


Fig. 4. Schematic diagram of feature FRNM.

Fig. 4 shows the FRNM visualization diagram of a basketball player's catch movement. The end node of the athlete's left hand is the target node, displaying its continuous changes in space. The arrow direction represents the motion direction of the target node. The length roughly reflects the motion speed. P3DJMV aims to simplify the three-dimensional spatial into a one-dimensional angular space, as displayed in Fig. 5.

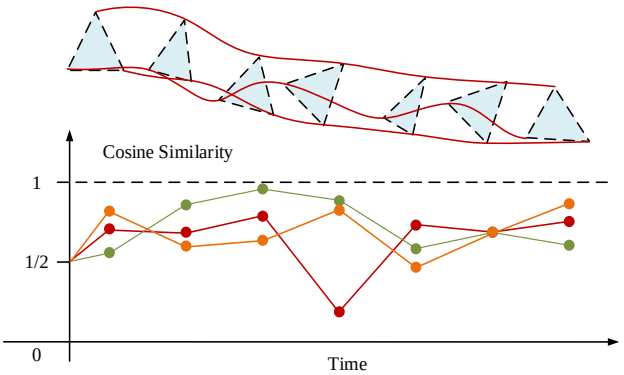


Fig. 5. P3DJMV feature visualization.

Unlike P3DJMV, FRNM essentially treats human joint points as a particle [19]. The particle mass is 1. Momentum represents the trajectory of particles. The human skeleton data contains 25 nodes. It is simplified to 20 nodes. Then, solve for the average and standard deviation of the entire dataset. The data meets the standard normal distribution through standardized processing. Afterwards, FRNM arranges groups from these 20 joints to extract possible planes, with a total of 1140 possibilities. Eq. (9) represents these planes expressions.

$$C_p = \{p_i, p_j, p_k\}, p \in N^+[1,1140]\ i, j, k \in N^+[1,20] \tag{9}$$

In Eq. (9), p_i, p_j, and p_k represent three joint points. p represents a plane. The three points p_i, p_j and p_k can obtain three vectors, as shown in Eq. (10).

$$\begin{cases} V_p(1) = (p_i^x - p_j^x, p_i^y - p_j^y, p_i^z - p_j^z) \\ V_p(2) = (p_i^x - p_k^x, p_i^y - p_k^y, p_i^z - p_k^z) \\ V_p(3) = (p_k^x - p_j^x, p_k^y - p_j^y, p_k^z - p_j^z) \end{cases} \tag{10}$$

In Eq. (10), the cosine value $V_p(1)$, $V_p(1)$, and $V_p(1)$ are shown in Eq. (11).

The P3DJMV is stacked in the tensor form of $F \times H \times W$. F represents the frames. H represents length. W represents the width. The node is abstractly represented as a physical particle. The mass is 1. The particle momentum is $\rho = mv$, and m is 1. If the trajectory of the physical particle is differentiated, then v can be obtained by differentiating the distance of the particle's motion per unit time, as shown in Eq. (12).

$$\begin{cases} A_p(1) = cos\alpha = \dfrac{V_p(1) \cdot V_p(2)}{V_p(1) * V_p(2)} \\ A_p(2) = cos\alpha = \dfrac{V_p(1) \cdot V_p(3)}{V_p(1) * V_p(3)} \\ A_p(3) = cos\alpha = \dfrac{V_p(3) \cdot V_p(2)}{V_p(3) * V_p(2)} \end{cases} \tag{11}$$

$$v = \lim_{\Delta t \to 0} \frac{y_{t+1} - y_{t-1}}{2\Delta t} \tag{12}$$

In Eq. (12), t represents the motion time. The FRNM is stacked as a geometric manifold $F \times H \times W$. F represents the frames. H represents length. W represents the width. Finally, P3DJMV and FRNM are fused and input into the prediction network. The research constructs the model based on 2D ResNet18. Its spatial down sampling remains unchanged. The first layer convolution is used to update the weights of each P3DJMV vector. In the spatio-temporal feature learning stage, each construction layer undergoes batch normalization and activation functions [20]. Finally, the two proposed features are fused after the fully connected layer. The research constructs a basketball motion recognition model based on perspective invariant geometric features extracted from skeleton data, as shown in Fig. 6.

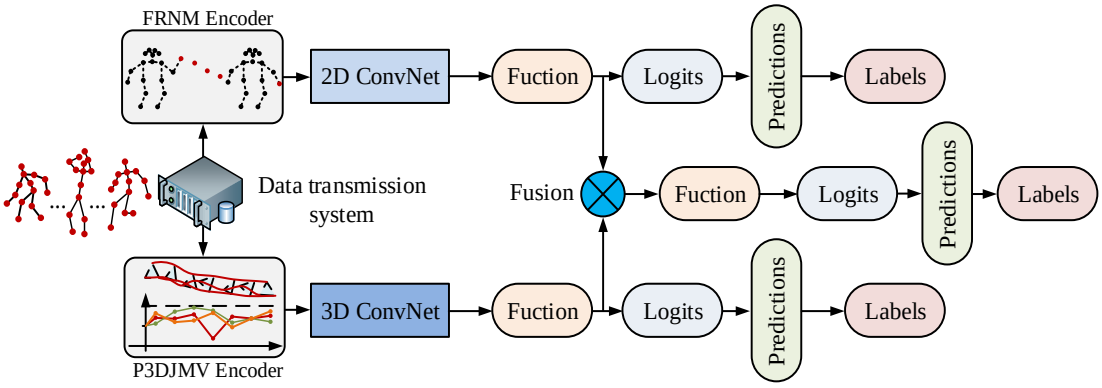


Fig. 6. Basketball movement recognition model based on skeleton data.

IV. Experimental Analysis

To improve the basketball motion recognition performance, the research proposed the basketball motion recognition model based on perspective invariant geometric features extracted from skeleton data. The dynamic topology map of the human skeleton was constructed to provide better output for recognition models. Then, a basketball motion recognition model based on perspective invariant geometric features was designed. To verify the effectiveness, test experiments were designed on datasets and real environments. The experimental results were analyzed.

A. Test Results in the Dataset

The dataset used in the experiment was two publicly available large-scale behavior recognition datasets, Kinetics and NTU RGB+D. Kinetic was a human behavior dataset that contained 300000 edited videos and 400 types of actions. The basketball movement segments were divided into training sets and validation sets. The second generation Kinetics depth sensor collection constituted the NTU RGB+D collection, with a total of 56000 edited action videos. To verify the validity of the research method, the experiments were conducted in the same environment. In the software architecture of the experimental system, this study

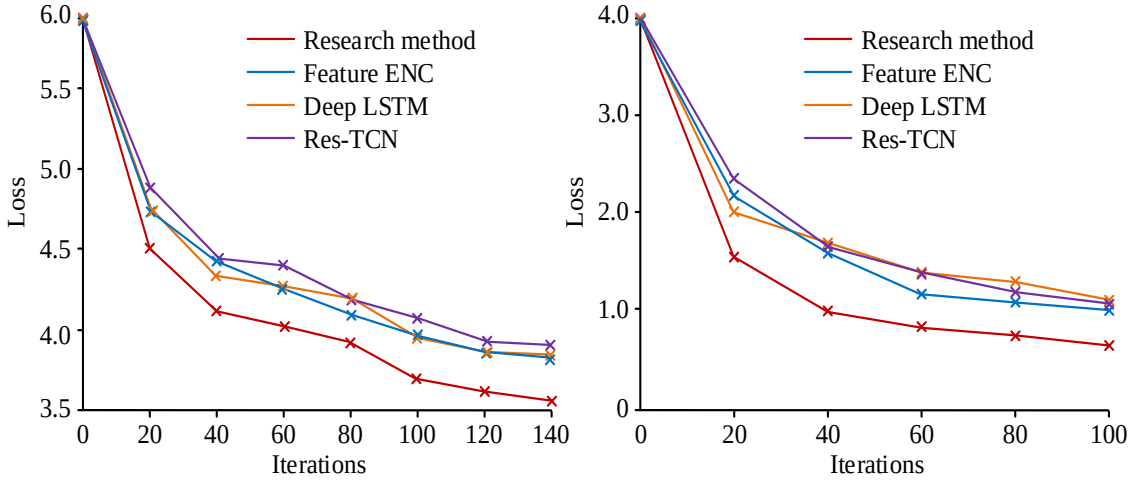
implemented five core modules, which were the Hikvision video capture module, the Noiton motion capture module, the Kinect somatosensory camera module, the Linux-based WiFi data acquisition module, and the synchronization control system used to coordinate these modules. Except for the Linux-based WiFi data acquisition module, all other modules ran on Windows 10 operating system. The research utilized the TCP/IP protocol to communicate with various collection subsystems, achieving data collection and storage operations. After precise testing, the system could ensure that the message synchronization error was within 100 milliseconds. In addition, the each acquisition module was flexible, which could work together on multiple hosts as well as operate independently. It enhanced the adaptability and reliability of the system. The experimental hardware environment was shown in Table I.

TABLE I. INTRODUCTION TO EXPERIMENTAL ENVIRONMENT

Name	Configuration Introduction
CPU	Intel Core i9-10920X CPU@3.50 GHz
GPU	NVIDIA1 GeForce RTX 3080Ti
Running memory	32GB
Operating system	Windows 10
Programming Language	Python
Development environment	Anaconda 3+python 3.6+pytorch 1.9

Top-1 and Top-5 classification accuracy were applied to evaluate recognition performance. The Top-1 indicator represented the probability that the first ranked category in the predicted category score vector was a true category. The Top-5 indicator represented the probability that the top five categories in the predicted category score vector contained the correct category. The proposed methods were compared with advanced methods, including Feature Encoding (Feature ENC) [21], Deep LSTM based on Recurrent Neural Network (RNN) [22], and Residual Temporal Convolutional Network (Res-TCN) based on Convolutional Neural Network (CNN) [23]. Among them, Feature ENC could effectively extract and encode key features, which improved the model's ability to recognize complex action patterns. The method demonstrated excellent performance in dealing with different types of action data, which was crucial for identifying diverse and complex basketball motions. Deep LSTM was suitable for dealing with time-series data. It combined the spatial feature extraction ability of CNN with the advantages of time series data processing. The residual network structure could avoid the gradient vanishing problem in deep network training, which represented the most advanced technology currently used to handle complex action sequences. The performance of different algorithms during training iterations was shown in Fig. 7.

Fig. 7 (a) showed the iterative loss curve on the Kinetics dataset. From the graph, the method proposed in the research showed significant differences from the other three methods after 20 iterations. The proposed method achieved lower loss values with faster iteration speed. When the iterations were 140, the loss of the research method was 3.52. The loss values for Feature ENC, Deep LSTM, and Res-TCN were3.85, 3.89, and 3.98, respectively. Fig. 7 (b) showed the iterative loss curve on the NTU RGB+D dataset. From the graph, the method demonstrated better performance after 20 iterations. When the iterations were 100, the loss value was 0.82. The other three methods were 1.18, 1.32, and 1.26, respectively. The results on the test set demonstrated the effectiveness of this method. The Top-1 indicator results of different algorithms were shown in Fig. 8.



(a) Kinetics dataset (b) NTU RGB+D dataset

Fig. 7. Curve of loss value with number of iterations.

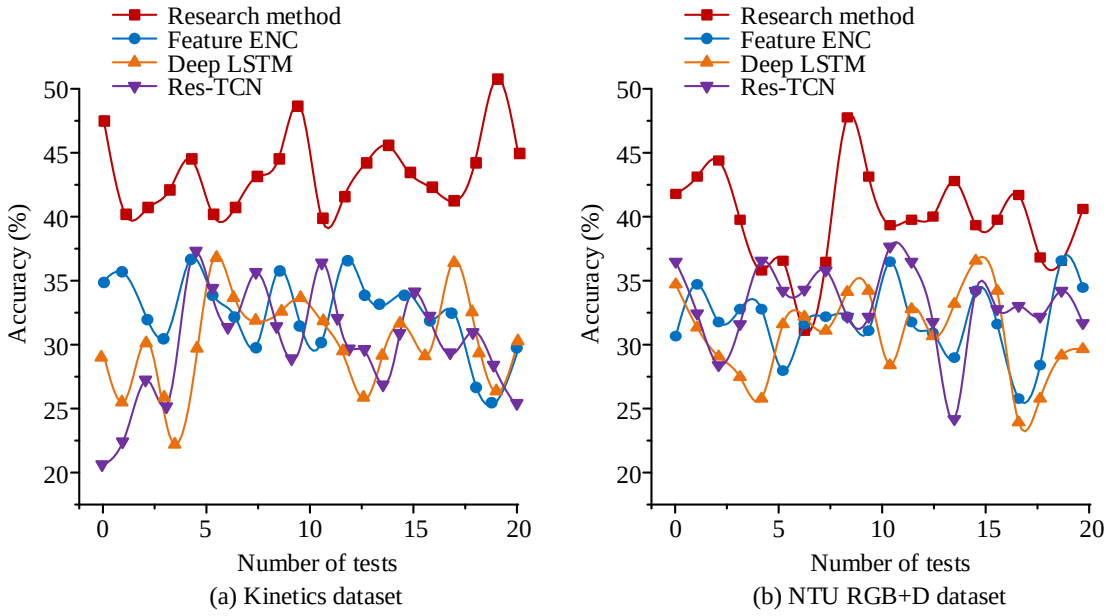


Fig. 8. Comparison results of Top-1 indicators.

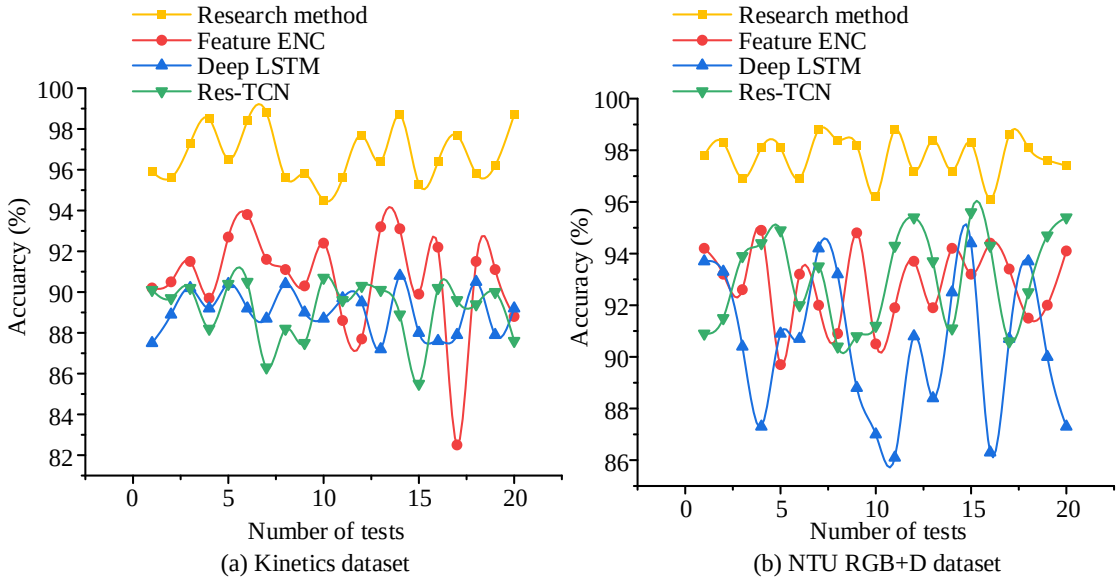


Fig. 9. Comparison results of Top-5 indicators.

Fig. 8 (a) showed the Top 1 standard accuracy. From the graph, the accuracy exceeded the other three methods. The average accuracy of the method in 20 tests was 43.52%, while the other three methods were 30.85%, 29.65%, and 27.34%, respectively. Fig. 8 (b) showed the Top 1 standard accuracy on the NTU RGB+D dataset. From the graph, the accuracy was higher than the other three methods in most tests. The average accuracy of the method in 20 tests was 40.52%. The other three methods were 34.12%, 32.07%, and 31.67%, respectively. From the experimental results, the research method had higher accuracy. The performance on the Kinetics dataset was superior to that on the NTU RGB+D dataset. The Top-5 indicator results of different algorithms were shown in Fig. 9.

Fig. 9 (a) showed the Top-5 standard accuracy on the Kinetics dataset. From the graph, the accuracy exceeded the other three methods. The average accuracy in 20 tests was 97.85%. The other three methods were 90.81%, 89.95%, and 89.27%, respectively. Fig. 9 (b) showed the Top-5 standard accuracy on the NTU RGB+D dataset. The accuracy also exceeded the other three methods. The average accuracy in 20 tests was 97.82%. The other three methods were 92.08%, 90.11%, and 91.73%, respectively. From the experimental results, compared to the current advanced three methods, the proposed method significantly improved the accuracy of the Top-5 evaluation index, verifying the effectiveness of this method.

B. Real Environment Application Analysis

To evaluate the performance of the basketball motion recognition model based on perspective invariant geometric features extracted from skeleton data in real environments, it performed motion recognition experiments on 80 volunteers. This study first used a confusion matrix to evaluate the performance of the basketball motion recognition model. The confusion matrix displayed the correspondence between the recognition results of each category and the actual results. It helped to understand the performance of basketball action recognition models in different categories, such as which categories were accurately identified and which categories were easily confused. By analyzing the confusion matrix, the researcher could identify and improve the weaknesses of the model, such as increasing the recognition rate or reducing misclassification. The research divided basketball motions into six categories, including shooting, dribbling, layups, passing, and grabbing the board. It performed 280 recognition tests for each action category. Fig. 10 displayed the test results.

Fig. 10 (a) showed the recognition results of the research method. The average correct recognition quantity for each category was 269.6. The expected correct recognition quantity was 280. Therefore, the average accuracy of the research method was 96.3%. Fig. 10 (b) displayed the recognition results of Feature ENC. The average correct recognition quantity for each category of Feature ENC was 247.8. The expected correct recognition quantity was 280. Therefore, the average accuracy of Feature ENC was 88.5%. Fig. 10 (c) showed the recognition results of Deep LSTM. The average correct recognition number for each category in Deep LSTM was 249.2. The expected correct recognition number was 280. The average accuracy of Deep LSTM was 89.0%. Fig. 10 (d) showed the recognition results of Res-TCN. The average correct recognition number for each category in Res-TCN was 240.4. The expected correct recognition number was 280. The average accuracy of the research method was 85.86%. From Fig. 10, the correctly identified color bands in research methods had darker colors, while the incorrectly identified color bands had lighter colors. This indicated that it still outperformed the other three advanced methods in real-world testing. To further validate the performance of the proposed model, the study increased the number of experiments to 20 and used recall rate as an indicator. The recall rate was the proportion of positive category samples (e.g., a specific basketball action) correctly recognized by the model to all actual positive category samples. This metric was particularly important for evaluating the recognition ability. The results were shown in Fig. 11.

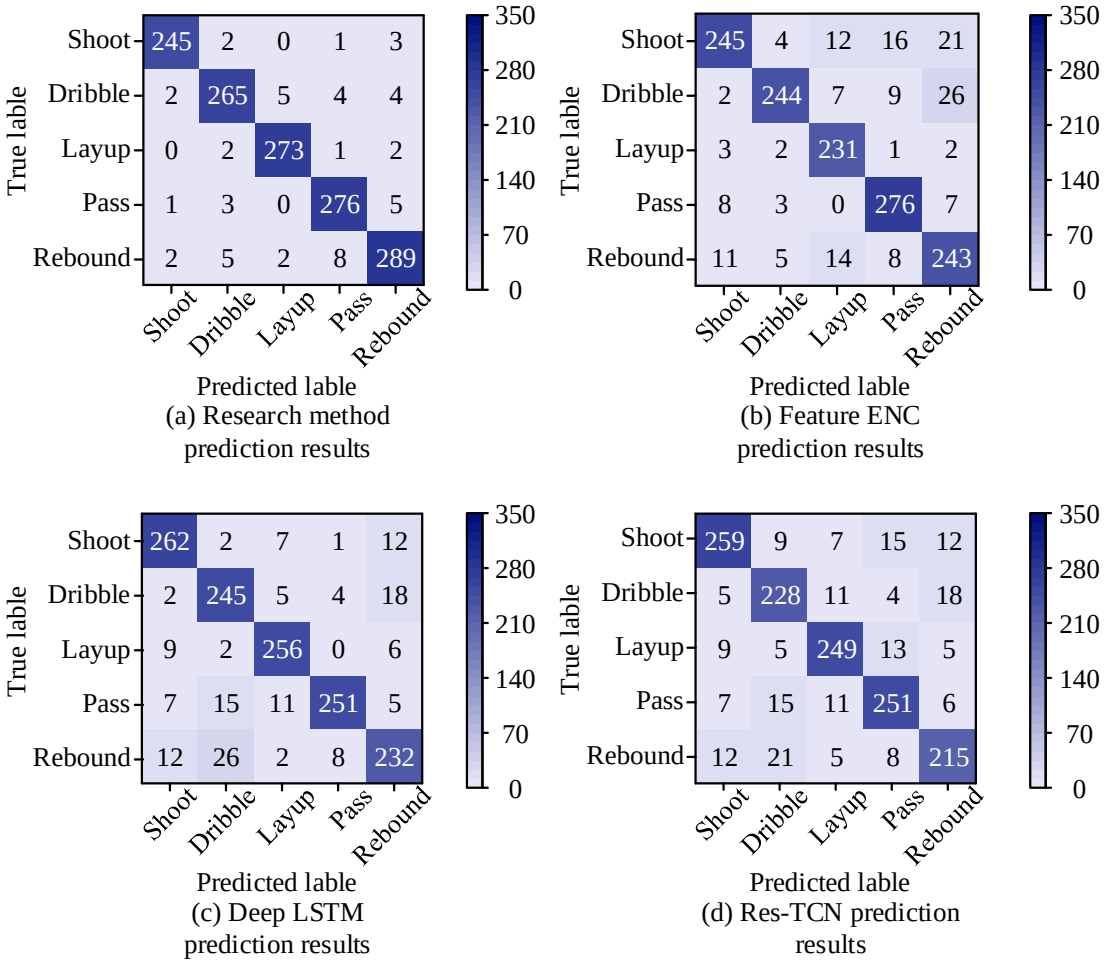


Fig. 10. Results of basketball motion recognition using different methods.

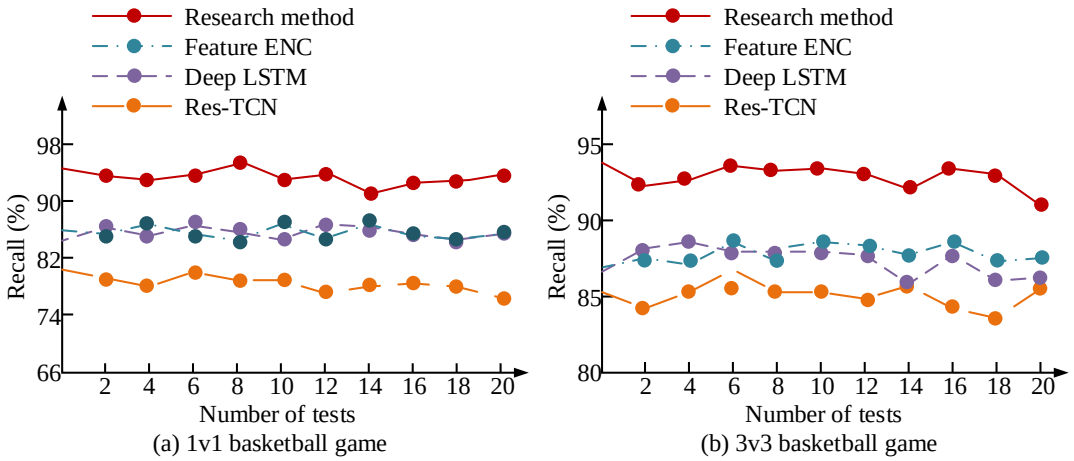


Fig. 11. Comparison of recall of different methods.

Fig. 11(a) showed the action recognition results in 1-to-1 basketball game. The recall of the research method was significantly higher than that of the other three methods. In 20 tests, the average recall rate was 96.35%, while the average recall rate of the other three methods was below 90%. Fig. 11(b) showed the action recognition results in 3-to-3 basketball game. The average recall of the research method reached 94.02% in 20 tests. The average recall of the other three methods was below 90%. From the recall experiments, the recognition performance of the constructed model was significantly better than other methods in different environments. Finally, to verify the computational complexity, the research compared the running times of the four recognition models. The results were shown in Fig. 12.

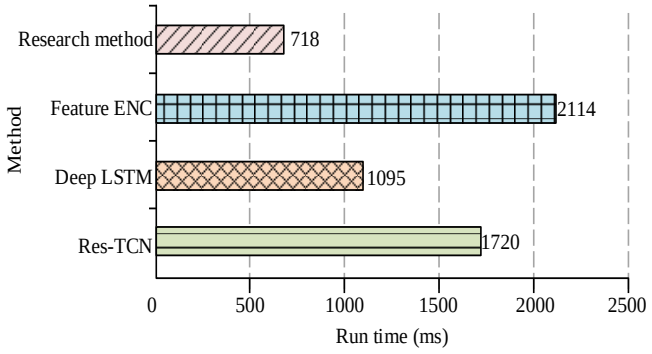


Fig. 12. Comparison of runtime of different recognition methods.

In Fig. 12, the running time of the research method was 718ms. The Feature ENC was 2114ms. TheDeep LSTM was 1095ms. The Res-TCN was 1720ms. Compared to the other methods, the running time of the proposed method was decreased by 66.03%, 34.43%, and 58.26%. According to the results, the basketball motion recognition model based on perspective invariant geometric features extracted from skeleton data had faster recognition speed.

V. Conclusion

To improve the utilization of human bone data and build a more accurate and efficient basketball motion recognition model, the study first constructs a dynamic topology map of human skeleton. Then it serves as input to the recognition model, thereby removing interference from environmental information. Then, based on the perspective invariance in skeleton data extraction, the study proposes two feature representation methods, namely the 3D joint motion vector plane and the feature representation method based on node momentum. By fusing two feature representations, the spatio-temporal feature fusion of skeleton data is achieved, thus constructing a new basketball motion recognition model. In real-world testing, the proposed method had an average accuracy of 96.3% for each category. The average accuracy of Feature ENC was 88.5%. The average accuracy of Deep LSTM was 89.0%. The average accuracy of Res-TCN was 85.86%. The results validated the effectiveness of this research. It demonstrated excellent performance in experiments. However, this research experiment uses single-mode data. Therefore, the recognition effect of multimodal data has not been verified yet. In the future, this method will be further optimized. Combined with other advanced machine learning technologies, this method will be improved to achieve better performance in a wider range of application scenarios.

Reference

[1] ZHAO J, SHE Q, MENG M, CHEN Y. Skeleton Action Recognition Based on Multi-Stream Spatial Attention Graph Convolutional SRU Network. ACTA ELECTONICA SINICA, 2022, 50(7): 1579-1585.

[2] Qin X, Li H, Liu Y, Yu J, He C, Zhang X. Multi-stage part-aware graph convolutional network for skeleton-based action recognition. IET Image Processing, 2022, 16(8): 2063-2074.

[3] Zhang P, Zhang J, Elsabbagh A. Lower limb motion intention recognition based on sEMG fusion features. IEEE Sensors Journal, 2022, 22(7): 7005-7014.

[4] Shu X, Yang J, Yan R, Song Y. Expansion-squeeze-excitation fusion network for elderly activity recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 5281-5292.

[5] Gao S, Yun J, Zhao Y, Liu L. Gait-D: Skeleton-based gait feature decomposition for gait recognition. IET Computer Vision, 2022, 16(2): 111-125.

[6] Song Z, Zhang Y, Liu Y, Yang K, Sun M. MSFYOLO: Feature fusion-based detection for small objects. IEEE Latin America Transactions, 2022, 20(5): 823-830.

[7] Zhou W, Guo Q, Lei J, Yu L, Hwang J N. ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection.

IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(3): 1224-1235.

[8] Zhang X, Wang J, Wang T, Jiang R. Hierarchical feature fusion with mixed convolution attention for single image dehazing. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(2): 510-522.

[9] Choi H, Yun J P, Kim B J, Jang H, Kin S W. Attention-based multimodal image feature fusion module for transmission line detection. IEEE Transactions on Industrial Informatics, 2022, 18(11): 7686-7695.

[10] Song Y F, Zhang Z, Shan C, Wang L. Constructing stronger and faster baselines for skeleton-based action recognition. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(2): 1474-1488.

[11] Li M, Chen S, Chen X, Zhang Y, Wang Y, Tian Q. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(6): 3316-3333.

[12] Li C, Xie C, Zhang B, Han J, Zhen X, Chen J. Memory attention networks for skeleton-based action recognition. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(9): 4800-4814.

[13] Zhang C, Liang J, Li X, Xia Y, Di L, Hou Z, Huan Z. Human action recognition based on enhanced data guidance and key node spatial temporal graph convolution. Multimedia Tools and Applications, 2022, 81(6): 8349-8366.

[14] Zhang Z, Wang S, Liu C, Xie R, Hu W, Zhou P. All-in-one two-dimensional retinomorphic hardware device for motion detection and recognition. Nature Nanotechnology, 2022, 17(1): 27-32.

[15] Oslund S, Washington C, So A, Chen T, Ji H. Multiview Robust Adversarial Stickers for Arbitrary Objects in the Physical World. Journal of Computational and Cognitive Engineering, 2022, 1(4): 152-158.

[16] Choudhuri S, Adeniye S, Sen A. Distribution Alignment Using Complement Entropy Objective and Adaptive Consensus-Based Label Refinement for Partial Domain Adaptation. Artificial Intelligence and Applications. 2023, 1(1): 43-51.

[17] Suneetha M, Prasad M V D, Kishore P V V. Sharable and unshareable within class multi view deep metric latent feature learning for video-based sign language recognition. Multimedia Tools and Applications, 2022, 81(19): 27247-27273.

[18] Zhang K, Li Y, Wang J, Cambria E, Li X. Real-time video emotion recognition based on reinforcement learning and domain knowledge. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(3): 1034-1047.

[19] Giannakeris P, Petrantonakis P C, Avgerinakis K, Vrochidis S, Kompatsiaris I. First-person activity recognition from micro-action representations using convolutional neural networks and object flow histograms. Multimedia Tools and Applications, 2021, 80(15): 22487-22507.

[20] Van Amsterdam B, Funke I, Edwards E, Speidel S, Collins J, Sridhar A, Stoyanov D. Gesture recognition in robotic surgery with multimodal attention. IEEE Transactions on Medical Imaging, 2022, 41(7): 1677-1687.

[21] Zhou Z, Dong X, Li Z, Yu K, Ding C Yang Y. Spatio-temporal feature encoding for traffic accident detection in VANET environment. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10): 19772-19781.

[22] Torres J F, Martínez-Álvarez F, Troncoso A. A deep LSTM network for the Spanish electricity consumption forecasting. Neural Computing and Applications, 2022, 34(13): 10533-10545.

[23] Shang Z, Liu H, Zhang B, Feng Z, Li W. Multi-view feature fusion fault diagnosis method based on an improved temporal convolutional network. Insight-Non-Destructive Testing and Condition Monitoring, 2023, 65(10): 559-569.

Application of Data Mining Technology with Improved Clustering Algorithm in Library Personalized Book Recommendation System

Xiao Lin*, Wenjuan Guan, Ying Zhang
Library, Minjiang University, Fuzhou 350108, China

***Abstract*—The information construction work of university libraries is becoming increasingly perfect. However, the massive amount of data poses significant challenges to the personalized recommendation of books. Cluster analysis has always been an important research topic in data mining technology, and it has a wide range of application fields. Clustering algorithm is a fundamental operation in big data processing, and it also has good application value in personalized recommendation of library books. To improve the personalized service quality of libraries, this study proposes a clustering algorithm based on density noise application spatial clustering. This study introduced a distance optimization strategy and Warhill algorithm to the proposed algorithm, to improve the difficulties in selecting initial parameter neighborhoods and density thresholds in traditional models, as well as computational complexity. Afterwards, this study will integrate the improved algorithm with the density peak algorithm to further improve the operational efficiency of the model. The performance verification of the model demonstrated superior clustering performance. The average accuracy of the proposed model's recommendation is 98.97%, indicating superiority. The practical application results have confirmed that there is a significant similarity between the books read by the readers and the books read by the target readers, and the effectiveness and feasibility of the proposed model have been verified. Therefore, the proposed model can contribute to the personalized recommendation function of libraries and has certain practical significance.**

***Keywords*—Peak density; distance optimization; warhill algorithm; collaborative filtering; book recommendations**

I. Introduction

At present, the informatization work of university libraries has made great progress, but there are still problems such as too long time spent in retrieving information and documents [1]. There are a lot of library resources in university library, and there are a lot of data, which is easy to cause college students to blindly choose books in university library or do not know which books are suitable for them to read. Therefore, more and more attention is paid to the personalized recommendation of university library. At the same time, the development of artificial intelligence has brought new opportunities to the topic. Among them, the traditional intelligent data analysis methods mainly collect and analyze users' browsing records and information, excavate and analyze the collection of documents and resources, and realize the utilization of library resources [2]. However, large amounts of data often have more complex structures and types, which makes it difficult for traditional data analysis methods to meet these needs [3]. Cluster analysis is the most commonly used method in data mining [4]. Through cluster analysis, data with high similarity can be classified into the same category, so that the difference between similar data is small, and the difference between different data is large. In addition, the existing book recommendation algorithms generally have the disadvantages of low recommendation accuracy and poor applicability in university libraries [5]. Therefore, in order to help college students more accurately find their own suitable books in college libraries, this study is based on cluster analysis in data mining. This paper proposes a Density-Based Spatial Clustering of Applications with Noise, The Clustering model of DBSCAN and Density Peak Clustering Algorithm (DPC) is proposed to optimize the personalized recommendation performance of books in university libraries. The innovations of this study lie in: (1) The distance optimization strategy was proposed, which improved DBSCAN and improved the difficulty in selecting its initial parameters. (2) Warhill was introduced to reduce the computational complexity of the model. (3) The improved DBSCAN was integrated with DPC to further improve the operational efficiency of the model. (4) The constructed model was applied in the personalized book recommendation service of library. This study consists of four parts: (1) Firstly, a review was conducted on the current development status of technologies used in the article. (2) The construction process of the model was discussed in detail. (3) The model performance and practical application effects were verified. (4) The full text was summarized and prospects were made for the future.

II. Related Works

The application of CA is quite extensive and has always been a hot topic of discussion among scholars. DPC is unable to identify cluster centers and non-center point error allocation, which can cause a chain reaction. Therefore, DingS et al. proposed an improved method. This method can reduce the density difference of non-uniformly distributed datasets and use low density points as boundaries on the foundation of cluster center and surrounding points' similarity density. The proposed model can make algorithm's clustering accuracy improved while controlling running time [6]. CuiZ and other researchers proposed an improved subspace clustering model to address the shortcomings of subspace clustering methods that cannot balance the sparsity and connectivity of coefficient matrices in high-dimensional image data. This model can preserve the coefficients between the sample and its neighbors,

and prune the spatial connections within the subspace. The proposed model can effectively handle noise data in the Internet of Things and has good clustering accuracy [7]. KarimM and other scholars have conducted a detailed discussion on deep learning based clustering algorithms and pointed out different clustering quality indicators. This study utilizes the clustering methods discussed for text mining in bioimaging, cancer genomics, and biomedical texts. The final conclusion helps to provide new solutions for emerging bioinformatics problems [8]. Incomplete multi view CA methods can lead to intensive computation and complex storage. In response, researchers such as LiuX have proposed an efficient incomplete multi view method. This method utilizes consensus clustering matrix to interpolate the generated incomplete base matrix to optimize the clustering performance of the model. This study validates the performance of the proposed model in terms of clustering accuracy, evolution of consensus clustering matrix, and convergence [9]. Based on the application of CA in data mining, ZouH elaborated in detail on the basic concept, classic algorithms, and implementation process of CA through literature comparison and analysis methods. The study ultimately conducted numerical simulations, confirming the strong universality of the proposed method. It can be applied to data analysis in multiple fields and has strong theoretical value [10].

Personalized recommendation services have received increasing attention in recent years, and many scholars have contributed to this. Researchers such as ArabiH have found that contextual information such as emotions, location, and time can help improve service recommendations. Therefore, they proposed a context aware recommendation system. The system implements personalized settings based on multiple user characteristics and product functions. It utilizes users' personality traits, demographic details, geographical location, and comment emotions to generate personalized recommendations. This algorithm was ultimately proven to greatly optimize recommendation performance [11]. Scholars such as HuixiangX use K-means and utilize the relationships between users, tags, and books for group recommendations. The proposed strategy first clusters users and books, and calculates the cosine similarity between the two groups. Afterwards, it sorted and clustered the books, and tested the personalized recommendation effect. The proposed model improves the recommendation effect of books and provides better book resources for the target group [12]. SarmaD et al. constructed an effective online book recommendation system to address the irrationality of existing recommendation system rating techniques. The system uses the K-means cosine distance function to measure distance to find similarity between book clusters and grade books. The experiment used 10 datasets to validate the proposed model, indicating that the constructed recommendation system has high accuracy [13]. ZhouY has designed an information recommendation book management system based on an improved Apriori data mining algorithm, which integrates borrower and book data information. After cleaning, transforming, and integrating the relevant data, the Apriori algorithm is used to generate an association rule database. The implementation process of association matching is carried out by the personalized recommendation sub module based on the borrower and the books selected in the association rule database. The proposed model has good recommendation performance [14]. KwakW and NohY analyzed their domestic and international trends, policies, and cases based on the development background of artificial intelligence, and proposed the future direction of artificial intelligence services for libraries. This study suggests that in the future, libraries will further optimize artificial intelligence and provide personalized book recommendations to users based on their usage records [15].

To sum up, cluster analysis plays an important role in personalized recommendation. Although cluster-based algorithms have always been favored and improved by scholars, there are still some problems that need to be continuously studied and perfected by scholars. In addition, the existing recommendation services widely exist in e-commerce, short video and other platforms, and there is still a large research space for personalized recommendation of college books. In addition, the existing book recommendation algorithms also have the disadvantages of low recommendation accuracy and poor applicability in university libraries. Therefore, this study proposes a clustering algorithm combining DBSCAN and DPC, and applies it to the book recommendation system, aiming at contributing to the personalized recommendation service of libraries.

III. A Book Recommendation Model Based on Improved DBSCAN and DPC

Traditional DBSCAN has many limitations. Therefore, this section first optimizes and improves it. Then it is integrated with DPC to better achieve the clustering performance and book recommendation effect.

A. Construction and Improvement Strategy of DBSCAN

DBSCAN is the most classic density based CA. DBSCAN adopts the concept of neighborhood and utilizes the spatial distribution characteristics of point sets to cluster the dataset [16]. The basic idea is to determine the number of neighboring data points under a certain threshold, centered around a single data point. Using sparse points as the boundary, low density data points are used as the classification boundary, and high density points are used as another class [17]. Compared with other clustering algorithms, DBSCAN can find clusters of different sizes and shapes under conditions of noise interference. The full name of DBSCAN is "Density-Based Spatial Clustering of Applications with Noise", which is characterized by its ability to effectively process noisy data. Compared with conventional CA, DBSCAN not only handles non convex data well, but also fits convex datasets well. DBSCAN includes elements such as neighborhood, density threshold, core points, boundary points, and outliers. It is assumed that the sample dataset is $S=\{x_1,x_2,\cdots,x_n\}$, $x_i \in S$, and the neighborhood is δ. All sample points in the neighborhood form a subsample set and meet the conditions shown in Eq. (1).

$$N_\delta(x_j)=\left\{x_j \in S \middle| dis\tan ce(x_i,x_j)\le\delta\right\} \quad (1)$$

Eq. (1) represents the conditions that need to be met for the

sub sample set in the hypersphere region with a radius of δ. The density threshold in DBSCAN can be manually set. According to the relevant elements of DBSCAN, there are three relationships between data points in the overall algorithm: density direction, density reachability, and density connection [18-19]. Among them, the meaning of density direction is: for two samples that exist in the domain, if one is the core object, the other sample is called density direct access by that sample. The meaning of density reachability is: for x_i and x_j, if there is a sample sequence $\left\{p_t \middle| t=1,2,\ldots,T\right\}$ that satisfies $p_1=x_i, p_T=x_j$, and p_{t+1} is directly reachable by p_t density, then x_j is said to be reachable by x_i density. At this point, the samples in the sequence are all core points, which means that the density can satisfy transitivity but not symmetry. The meaning of density connection is: for x_i and x_j, if there is a core point x_k, so that both x_i and x_j can be reached by x_k density, then x_i and x_j are called density connections. Fig. 1 shows the schematic diagram of DBSCAN. x_1 and x_2 are the core points. x_3 and x_4 are boundary points. x_2 is directly accessible through x_1 density. x_3 is directly reached by x_2 density. x_4 can be achieved by x_1 density. x_4 and x_3 are density connected.

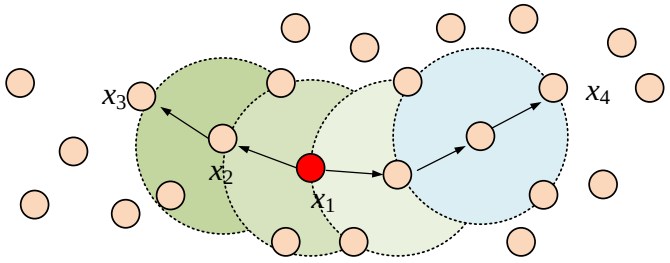


Fig. 1. DBSCAN clustering process.

Compared with other clustering methods, DBSCAN has the following advantages. Firstly, it can handle dense data of any shape. Due to its ability to find and remove noise throughout the entire search process, the clustering process is not affected by sample set noise. Its clustering does not require specifying the number of clusters in advance, and the clustering results are unbiased. However, the classic DBSCAN also has significant drawbacks that cannot be ignored. In DBSCAN, it is difficult to choose the initial parameter neighborhood and density threshold. DBSCAN is not suitable for samples with uneven distribution and large distances, in which case the clustering results are not ideal. For high-dimensional sample sets, the convergence speed of clustering algorithms is slow and cannot achieve accurate clustering results [20-21]. Therefore, this study first proposes optimization for the parameter selection process of DBSCAN, namely a DBSCAN model based on distance optimization (D-DBSCAN). D-DBSCAN can automatically select neighborhood values based on the characteristics of initial density threshold and data distribution density. Assuming the sample set is $S=\left\{x_1,x_2,\cdots,x_n\right\}$, there is Eq. (2).

$$N\left(x_i\right)=\left\{x_j\in S \middle| 0<d\left(x_i,x_j\right)<\delta\right\} \tag{2}$$

Eq. (2) indicates that for $x_i\in S$, the density of x_i is the number of data points owned within the domain δ. Eq. (3) is the distance coefficient.

$$\theta=N\left(x_j\right)/N\left(x_i\right) \tag{3}$$

In Eq. (3), θ is the distance coefficient, which specifically means that for the sample point x_j within the neighborhood of core point x_i, it is called the distance coefficient of x_j to x_i. Eq. (4) is the distance matrix between samples.

$$D=\left\{D_{ij} \middle| i,j\in R,\ i\neq j\right\} \tag{4}$$

In Eq. (4), D_{ij} represents the distance between samples x_i and x_j. The average distance of the points with the closest density threshold to x_i was calculated and added to the distance set U to calculate the overall average $\overline{U}$ of U. The neighborhood radius was set as the average distance $\overline{U}$, and all core points were identified and added to the core object set Ω. Subsequently, a core point x_i was randomly selected as the clustering center to form a new cluster C_i. All sample points in the δ neighborhood near x_i were identified, and the neighborhood radius δ was adjusted by calculating distance coefficient θ between the sample and core points. By repeating the above operation to find all sample points with achievable density and adding them to C_i, the final result was obtained in Eq. (5).

$$C=\left\{C_1,C_2,\cdots,C_n\right\} \tag{5}$$

Thus, D-DBSCAN can effectively improve the selection of initial parameters. To further reduce the complexity of D-DBSCAN calculation, Warhill was introduced to construct W-D-DBSCAN. Warhill can calculate reachable matrices to reduce the complexity of the model. It is assumed that Eq. (6) is a directed graph.

$$G=\left\langle V,E\right\rangle \tag{6}$$

In Eq. (6), V represents the node set. E refers to an adjacent points set. The node set is $V=\left\langle v_1,v_2,\cdots,v_n\right\rangle$. The matrix is $A=\left(a_{ij}\right)_{m\times n}$.

$$a_{ij}=\begin{cases}1, & v_i\ adjoin\ v_j\\ 0, & v_i\ is\ not\ adjacent\ to\ v_j\end{cases} \tag{7}$$

A directed graph can directly reflect two elements' connection. An adjacency matrix refers to nodes' connection in the directed graph. The two nodes that can be directly reached are regarded as 1. Otherwise, they are regarded as 0. It is

assumed that $G=\langle V,E\rangle$ is a simple directed graph. The node set is $V=\langle v_1,v_2,\cdots,v_n\rangle$. The matrix is $F=\left(f_{ij}\right)_{n\times n}$.

$$f_{ij}=\begin{cases}1, \text{There is a non-zero directed path from } v_i \text{ to } v_j\\ 0, \ others\end{cases} \quad (8)$$

If there is a reachable path between two elements, they are connected and reachable, marked as 1, otherwise marked as 0. In a directed graph, the connectivity between nodes can be directly reflected by a line with an arrow. However, it is difficult to determine the connectivity between two independent nodes, and a matrix of direct connectivity must be used. Usually, Warshall is used to convert adjacency matrices into reachable matrices. $dis[i,j]$ stands for the distance. A matrix $A_{n\times n}$ was established for dataset D and $dis[i,j]$ between data objects i,j was calculated in Eq. (9).

$$A[i,j]=\begin{cases}1, dis[i,j]\le Eps\\ 0, dis[i,j]>Eps\end{cases} \quad (9)$$

In Eq. (9), Eps represents the initial threshold. The adjacency matrix's reachable matrix obtained by Warhill is regarded as transitive closure or density connected sets which are the maximum. Fig. 2 shows the flowchart of W-D-DBSCAN. The density connected set obtained by Warhill is to achieve clustering. Compared with traditional density clustering, its clustering process is simpler.

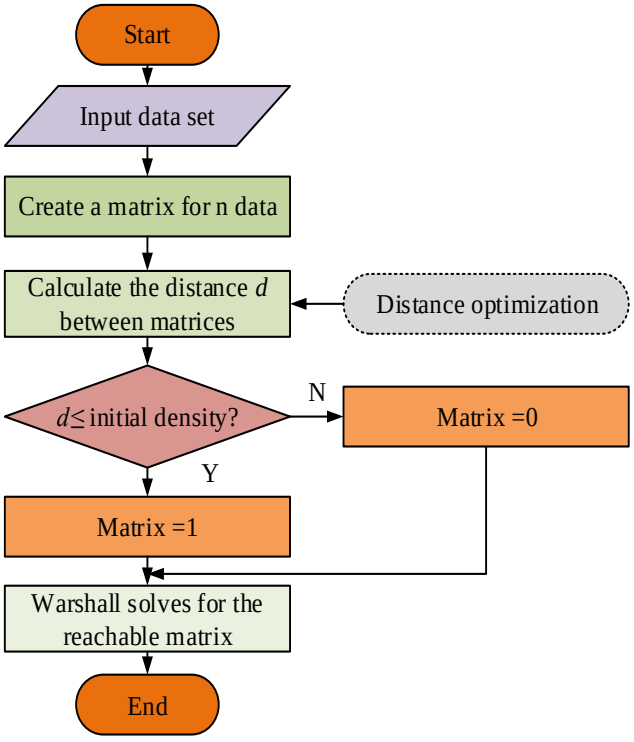


Fig. 2. Flow of the W-D-DBSCAN algorithm.

B. A Book Recommendation Model Integrating D-W-DBSCAN and DPC

DPC is a data clustering method based on maximum connectivity, which not only effectively distinguishes noise, but also has strong robustness for various convex and non-convex data [22-23]. DPC assumes that the centroid of data targets set is surrounded by a low-density data, and this low-density data is assigned to the nearest, less dense centroid data. DPC improves the efficiency of the algorithm and has good clustering performance by manually drawing decision maps and selecting data objects with high relative distance and density for clustering. On this basis, a new method of clustering quality centers based on DPC is proposed. This study adopts a combination of W-D-DBSCAN and DPC, namely the W-D-DBSCAN DPC model, to solve the problem that the centroid selected by DPC on the decision graph cannot be applicable to all data. This method first uses the maximum region density value as the centroid. Secondly, W-D-DBSCAN was used for clustering. Then, samples with high local density were searched out from the remaining samples, and the samples were clustered using W-D-DBSCAN as the center. The above methods were used to search for centroids and cluster them until all data were classified or local noise appeared, indicating the completion of clustering. This method can effectively solve the manual intervention problem in traditional DPC methods, while also reducing the computational complexity of W-D-DBSCAN.

Assuming any data object is i, the local density was calculated for it in Eq. (10).

$$\rho_i=\sum_j \chi\left(d_{ij}-d_c\right) \quad (10)$$

In Eq. (10), ρ_i represents local density. d_c represents truncation distance. d_{ij} represents relative distance. The relative distance between other data and data i was calculated using Euclidean distance in Eq. (11).

$$\sigma_i=\min_{J:\rho_j>\rho_i}\left(d_{ij}-d_c\right) \quad (11)$$

In Eq. (11), d_{ij} represents i and j's Euclidean distance. Through Eq. (11), the relative distance of data i refers to: if i is the largest data object of ρ_i, then σ_i represents the $\max_j\left(d_{ij}\right)$ between other data and data i. If i has no the highest local density, its relative distance is in the data with lower local density than the data, and it is closest to i. Fig. 3 shows the clustering process of W-D-DBSCAN-DPC [24].

In Fig. 3, the comprehensive model first calculates the local density. Then, they are compared to get the maximum local density, with data 7 of this maximum local density as the centroid. Starting from 7, W-D-DBSCAN was used to divide 1, 2, 3, 4, 5, and 6 into centroid data, completing the classification of the first type of cluster. Then, the local density of the remaining dataset was continued to be calculated. And the maximum local density was obtained through comparison again. Using the highest local density value of 10 as the centroid, W-D-DBSCAN was used to

cluster samples 8, 9, 11, and 12, and data 8, 9, 11, and 12 were divided into centroids 10. This process continues until the remaining data does not belong to a category, marked as noise, and clustering ends [25].

This study applies W-D-DBSCAN-DPC to the clustering process of library readers. By categorizing different categories of readers, the first category in the database is the book that the reader is most interested in. Therefore, recommending the first category of books to readers is a desirable approach. The recommendation algorithm of W-D-DBSCAN-DPC effectively solves the traditional "cold start". This method will help improve the performance of recommendation systems and alleviate the pressure of big data processing. Fig. 4 shows the system diagram of the library book recommendation system. W-D-DBSCAN-DPC is used to classify readers with high similarity. By compressing the existing massive reader data into analyzing and recommending the same type of reader data, the recommendation efficiency was improved. The collaborative filtering algorithm is used to generate the Top-n nearest neighbor set of readers, thereby completing recommendations.

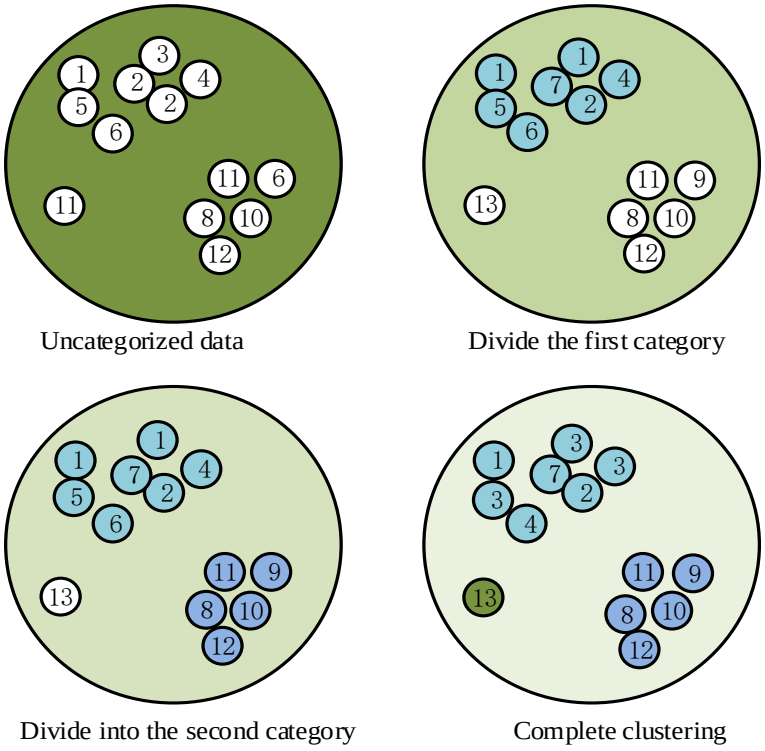


Fig. 3. Clustering process of W-D-DBSCAN-DPC.

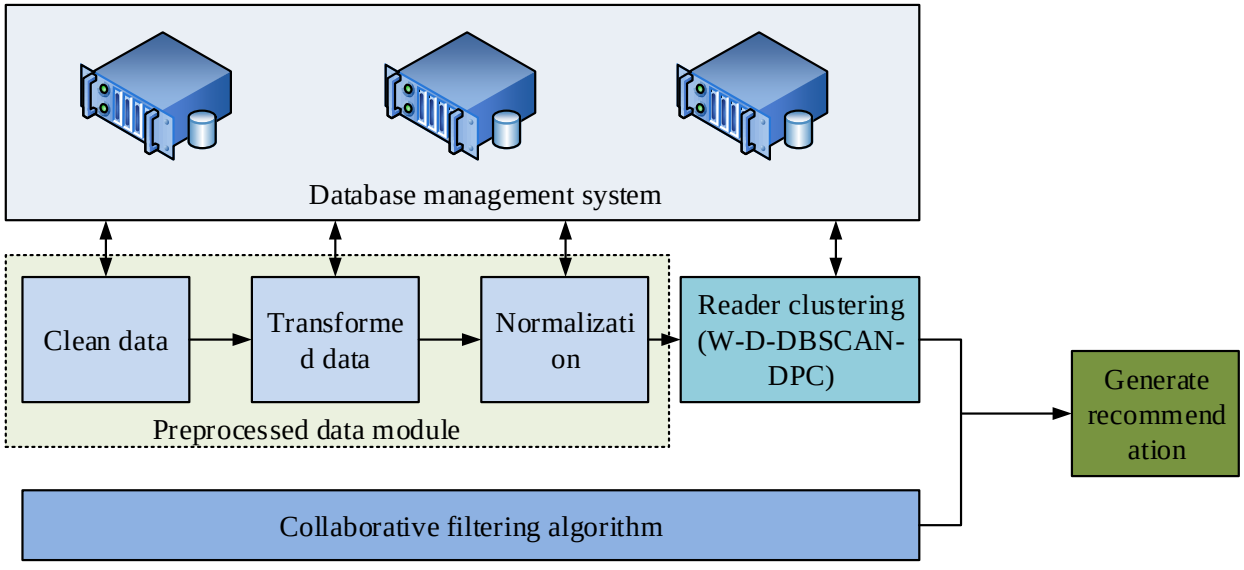


Fig. 4. Library book recommendation system.

This study used W-D-DBSCAN-DPC to cluster readers, identify the most popular books, and recommend them to new readers to solve the "cold start" problem. Adopting this method for users with a reading history reduces the range of data that the recommendation algorithm needs to process and reduces the problems that users may encounter during the reading process. If the target reader has a historical borrowing record, then based on its clustering results, a collaborative filtering algorithm is used to obtain the Top-n neighbor set for other readers of the same type and recommend them. The similarity between the target reader and the reader is calculated by Eq. (12).

$$sim(i,j)=\frac{\sum_{u\in U_{ij}}\left(R_{u,j}-\overline{R_i}\right)\left(R_{u,j}-\overline{R_j}\right)}{\sqrt{\sum_{u\in U_{ij}}\left(R_{u,j}-\overline{R_i}\right)}\sqrt{\sum_{u\in U_{ij}}\left(R_{u,j}-\overline{R_j}\right)}} \quad (12)$$

In Eq. (12), $sim(i,j)$ represents the similarity between readers i,j. U_{ij} represents the book categories that readers i,j are both interested in. $\overline{R_i}$, $\overline{R_j}$ represent the average values of readers i,j's interest in all books. By using Eq. (12), the similarity between the target reader and the reader can be obtained, and the similarity sequence can be obtained by sorting their similarity. Then, the Top-n neighbor set is generated from the high similarity readers. Therefore, the book recommendation model based on W-D-DBSCAN-DPC has been constructed.

To evaluate the clustering effect, Accuracy (ACC), Purity, and contour coefficient were introduced as evaluation indicators in this study. Eq. (13) is the calculation of ACC.

$$ACC=\left(\sum_{i=1}^{n}\delta\left(\hat{C}_i map\left(C_i\right)\right)\right)/n \quad (13)$$

In Eq. (13), C_i represents the category label of the proposed algorithm. $\hat{C}_i$ represents the true label of data object. $map(x)$ represents a mapping function. A high ACC value indicates high clustering quality. Eq. (14) represents the calculation of Purity.

$$purity=\frac{1}{N}\sum_{i=1}^{k}x_i \quad (14)$$

In Eq. (14), N represents the number of datasets. k represents the number of clusters in the dataset. x_i represents the number of correctly clustered data objects. Purity is within 0-1, which is closer to 1, the data clustering accuracy is higher. Eq. (15) is the calculation of contour coefficient.

$$S(X)=\frac{b(x)-a(x)}{\max\left(a(x),b(x)\right)} \quad (15)$$

In Eq. (15), $a(x)$ represents the average distance of other samples within the same cluster of sample x. $b(x)$ represents the average distance between sample x and all sample points within the nearest cluster. The range of contour coefficient values is $[-1,1]$, which is closer to 1, the clustering effect is better.

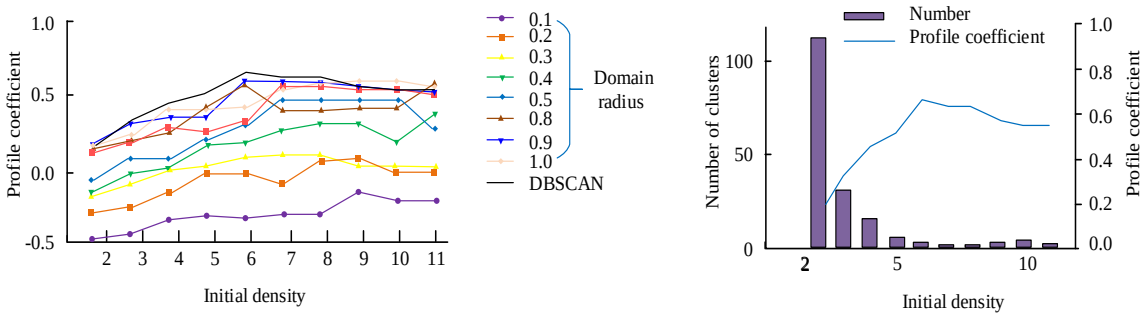
IV. Performance Verification of Book Recommendation Model Application Based on W-D-DBSCAN-DPC

This study first analyzes the clustering performance of W-D-DBSCAN-DPC. For this purpose, sufficient datasets were selected for the study and detailed discussions were conducted. In addition, the actual application effect of book recommendation was verified using a certain university as an example.

A. Performance Verification of W-D-DBSCAN-DPC Model

This study first verifies the clustering performance of the W-D-DBSCAN-DPC model. This study selected three datasets, namely Spiral, Lineblobs, and Aggregation, for validation. Spiral is a set of non-convex spiral datasets. Lineblobs is a set of smiling face datasets. Aggregation includes both spherical and non-spherical datasets. The proposed W-D-DBSCAN-DPC is implemented in C language and visualized using MATLAB. The proposed algorithm was evaluated and analyzed by comparing clustering results at different scales. To evaluate the clustering effect and determine the optimal number of clusters, this study used traditional DBSCAN and D-DBSCAN to cluster the dataset, and obtained the contour coefficients corresponding to different number of clusters in Fig. 5.

In Fig. 5 (a), DBSCAN requires continuous search of two parameters, namely initial value and neighborhood radius, in order to find the optimal solution. D-DBSCAN only needs to find the optimal solution under different initial density conditions, without adjusting the neighborhood radius. Fig. 5 (b) shows the number of class clusters and contour coefficients corresponding to different initial densities. At an initial density of 6, the D-DBSCAN contour coefficient reached its optimal value of 0.668. The contour coefficient of the DBSCAN method is 0.612. Compared with DBSCAN, the clustering performance of D-DBSCAN has improved by 9.17%. These results verify the effectiveness of distance optimization.



(a) DBSCAN profile coefficients at different initial densities and domain parameters

(b) D-DBSCAN Specifies the value of d-dbscan at different initial densities

Fig. 5. D-DBSCAN performance verification results.

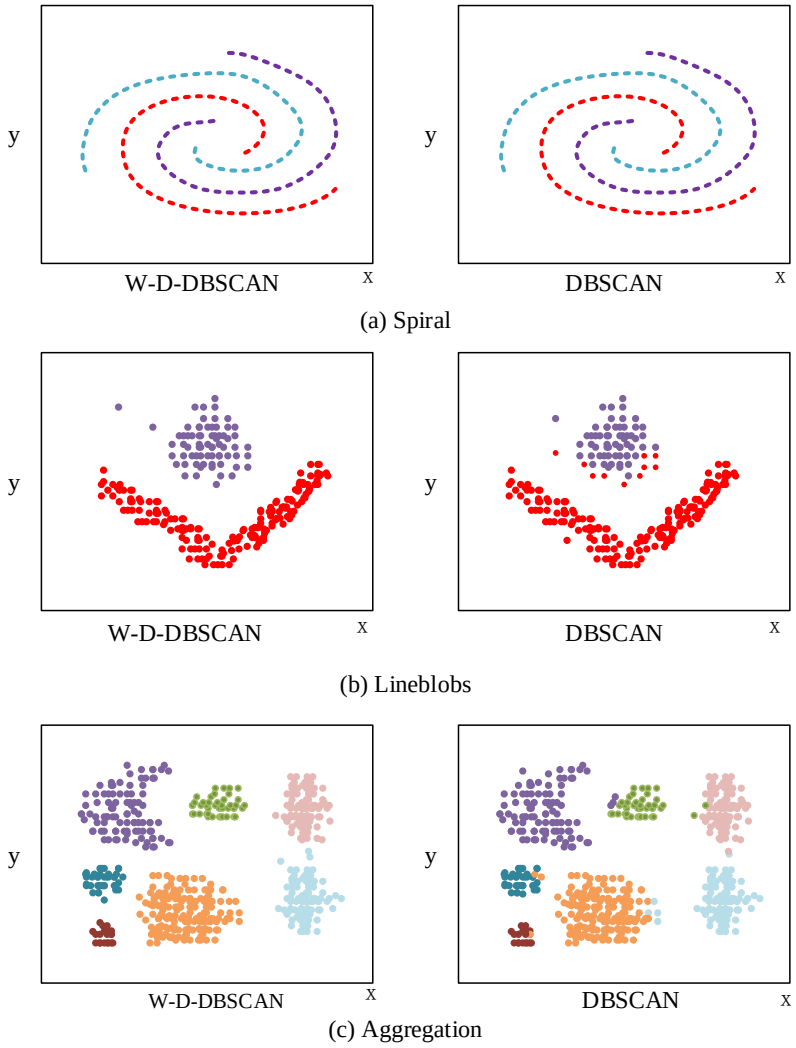


Fig. 6. Clustering results of D-DBSCAN and W-D-DBSCAN.

In Fig. 6, they are the clustering results of three datasets using two methods, D-DBSCAN and W-D-DBSCAN. In the figure, two algorithms' clustering effects on three types of datasets are relatively similar. Because W-D-DBSCAN continues the advantages of D-DBSCAN and can achieve clustering on any dataset.

Fig. 7 shows the comparison results of the runtime between D-DBSCAN and W-D-DBSCAN algorithms on a spiral dataset. The spiral dataset was generated into datasets with different shapes, sizes, and densities. And experiments were conducted on these datasets using D-DBSCAN and W-D-DBSCAN. In Fig. 7, W-D-DBSCAN has a shorter

runtime. As a result, W-D-DBSCAN effectively reduces the complexity of the model and improves its running speed.

To highlight the excellent performance of W-D-DBSCAN-DPC, this study selected four clustering algorithms: DPC, FCM, and K-means for comparative analysis. Fig. 8 shows the clustering performance of four algorithms on three datasets. In Fig. 8(a), K-means cannot cluster non-convex datasets, while the spiral shape of Spiral dataset is non-convex, resulting in clustering errors. In Fig. 8(b), the clustering results of W-D-DBSCAN-DPC and DPC are basically correct because they consider the transfer relationship between data. In Fig. 8(c), the clustering performance of W-D-DBSCAN-DPC is superior to other three algorithms.

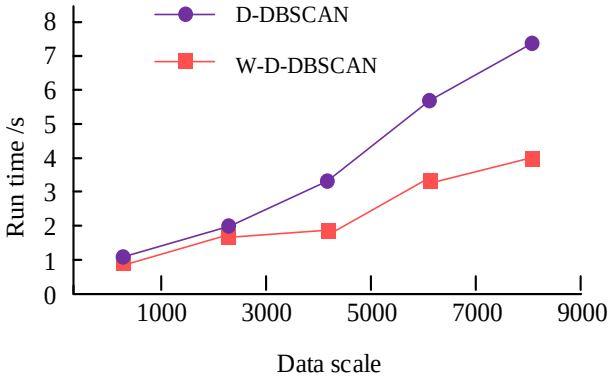


Fig. 7. Comparison of running time of the two algorithms on spiral data set.

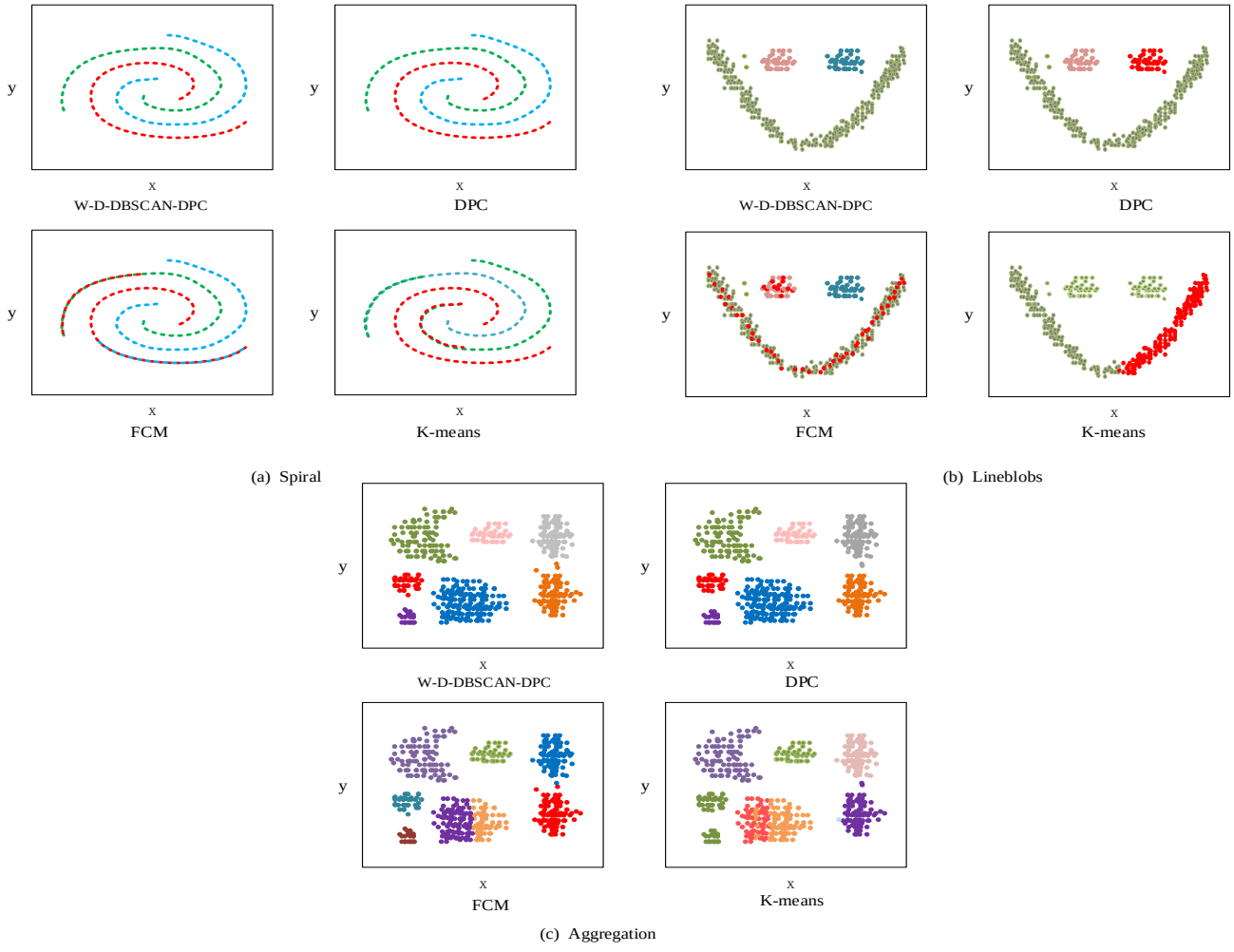


Fig. 8. Clustering effect of four algorithms on three data sets.

TABLE I. PURITY VALUES OF THE FOUR ALGORITHMS

Data set	W-D-DBSCAN-DPC	DPC	FCM	K-means
Iris	0.94	0.91	0.89	0.79
Tae	0.66	0.64	0.55	0.56
Cmc	0.78	0.73	0.63	0.56
Seeds	0.92	0.89	0.88	0.78

To further test the performance of W-D-DBSCAN-DPC, quantitative analysis was conducted on the aforementioned artificial dataset and UCI dataset. Four sets of data were randomly selected from UCI database for experiments on DPC, FCM, and K-means. Table I is four algorithms' purity values. The purity values on four datasets, W-D-DBSCAN-DPC, are the highest, indicating better clustering performance. Its clustering performance in Tae and Cmc datasets is relatively poor because one of these datasets has a large feature value, which affects the clustering results.

To demonstrate the effectiveness of the proposed W-D-DBSCAN-DPC recommendation algorithm in recommendation, the above three algorithms are used as comparative recommendation algorithms for this experiment. Fig. 9 shows the recommendation accuracy obtained by four algorithms on three types of datasets. The average accuracy of W-D-DBSCAN-DPC is 98.97%, while DPC, FCM, and K-means are 95.67%, 93.23%, and 90.34%, respectively. Thus, the superiority of W-D-DBSCAN-DPC was verified.

Fig. 10 shows the comparison results of recall rates obtained by four algorithms on three types of datasets. The average performance of W-D-DBSCAN-DPC is also superior to other algorithms, further verifying its superior performance.

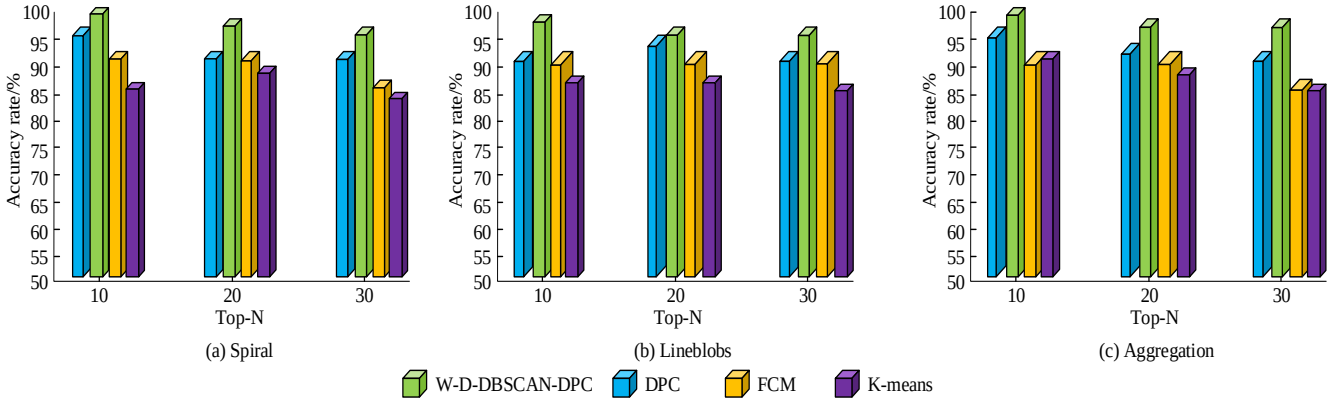


Fig. 9. Recommendation accuracy rates of four algorithms on three types of data sets.

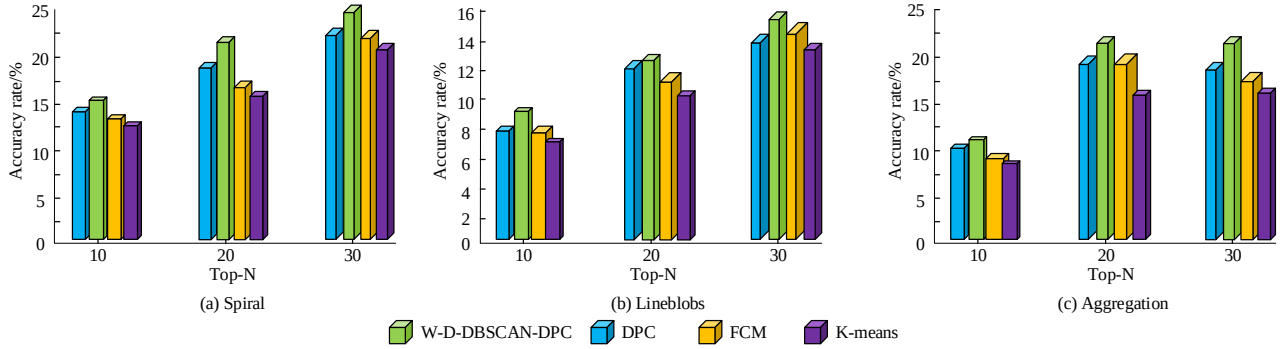
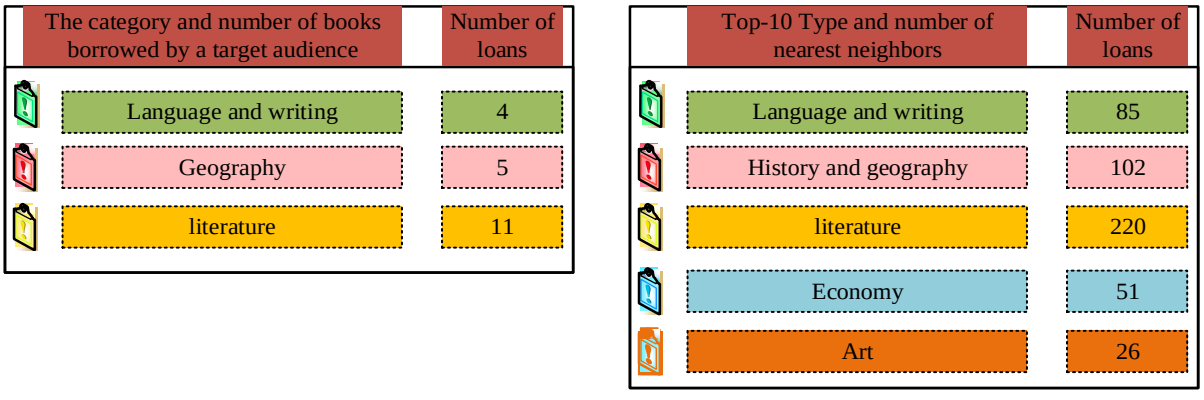


Fig. 10. Comparison results of recall rates of four algorithms on three types of data sets.



(a) The category and number of books borrowed by a target audience

(b) Top-10 Type and number of nearest neighbors

Fig. 11. W-D-DBSCAN-DPC recommended result.

B. Practical Application Effect of Book Recommendation Model Based on W-D-DBSCAN-DPC

This study divides the existing book materials in the library into 22 categories and uses them as reader feature vectors. The experiment extracted borrowing information from the data center of a certain university library for 21st year college students, of which 8011 were borrowing information from the university library. By processing these materials, some reader's interest and preference data can be obtained.

This experiment proved that W-D-DBSCAN-DPC was used for clustering, combined with recommendation algorithms, and finally the Top-n algorithm was used for classification. Fig. 11(a) is a summary of the types of books borrowed and the number of target readers. According to the distribution, when recommending four books to the target audience, two literary books, one local science book, and one language and text book can be recommended. Fig. 11(b) shows the borrowing classification and quantity of readers in the Top-10 neighbor set. These results confirm that among the top 10 neighboring readers of the target audience, in addition to books related to language, geography, literature, etc., there are also books related to history, economy, art, etc. Therefore, according to Fig. 11(b), books on history, economics, art, etc. can be appropriately recommended to target readers. In addition, there is a significant similarity in the types of books read by Top-10 readers and the books read by the target readers, thus verifying the effectiveness and feasibility of this algorithm.

V. Discussion

The problem of "blind selection" or "unable to find suitable books" is common among contemporary college students in the library. Therefore, this topic tries to combine the clustering algorithm with the recommendation system organically, and build a system suitable for university book recommendation. Firstly, the data are collected and classified. On this basis, the user's historical reading records are clustered and classified, and finally the user's Top-n nearest neighbor set is calculated, and books are recommended to it, so as to help students find books suitable for themselves. Density clustering has always been a hot topic in the field of data mining, and the density peak clustering algorithm has pushed the study of density method to a hot trend [26-28]. Starting from cluster analysis and recommendation system, this study introduced the theoretical analysis, research status and common methods of related methods in detail, so as to make sufficient preparation for the follow-up work. This study mainly discusses the density peak and density clustering algorithm. It is found that in the density clustering algorithm, data objects are grouped by density linkage, but the calculation process of density linkage is complicated, and it needs to determine the core point, density reachability, direct density reachability, etc. Density peaks the method of selecting the center of mass on a decision graph is not suitable for all data. The main contents of this research are as follows: (1) Propose distance optimization strategies to improve DBSCAN and improve the difficulty in selecting its initial parameters. (2) Warshall algorithm was introduced to reduce the computational complexity of the model. (3) Merge the improved DBSCAN with DPC to further improve the operating efficiency of the model. (4) Apply the model to the personalized book recommendation service of the library.

Experimental results show that, compared with the traditional DBSCAN algorithm, W-D-DBSCAN-DPC algorithm can find clusters with different shapes and adaptively select appropriate neighborhood radius, which is more suitable for complex student book preferences [29-30]. Finally, according to the results of student book recommendation, different student groups show strong differences in daily borrowing activities. It can be concluded that the reader feature vector based on the borrowing information of college students in the library has a strong correlation with the reading preference of students, which provides more reference and research ideas for college library managers.

VI. Conclusion

The increasing amount of information data in university libraries has brought a lot of inconvenience to the daily lives of teachers and students. This study proposes a W-D-DBSCAN-DPC to enhance the personalized service quality of libraries. The performance of the model was verified: the optimal contour coefficient for D-DBSCAN was 0.668, while for DBSCAN, it was 0.612. Compared with DBSCAN, D-DBSCAN clustering performance improved by 9.17%, verifying the effectiveness of distance optimization. W-D-DBSCAN has a shorter runtime compared to D-DBSCAN, verifying the effectiveness of Warhill. The comparative analysis of four clustering algorithms, W-D-DBSCAN-DPC, DPC, FCM, and K-means, shows that W-D-DBSCAN-DPC's clustering results are basically correct, and its clustering effect is better than other three algorithms. W-D-DBSCAN-DPC has the highest purity values on the four datasets, indicating better clustering performance. The average accuracy recommended by W-D-DBSCAN-DPC is 98.97%, while DPC, FCM, and K-means are 95.67%, 93.23%, and 90.34%, respectively, confirming the superiority of W-D-DBSCAN-DPC. The practical application has confirmed that there is a significant similarity between the books read by Top-10 readers and the books read by the target readers, verifying the effectiveness and feasibility of this algorithm. The drawback is that the threshold in Warhill has not been controlled, resulting in some interference that can be optimized in the future.

Acknowledgment

The research is supported by Fund Project of Steering Committee for Academic Libraries of Fujian Province: Research on Innovative Service of Augmented Reality and Mixed Reality Technology in University Libraries (Project number: FJTGW202228); Major topic of the research project on education and teaching reform of undergraduate colleges and universities of the Education Department of Fujian Province: the construction of smart platform for reading promotion activities of University Library Alliance in Fuzhou University Town (Project number: FBJG 20210319).

REFERENCES

[1] Ez-Zahout A, Gueddah H, Nasry A, Madani R, Omary F. A hybrid big data movies recommendation model based knearest neighbors and matrix factorization. Indonesian Journal of Electrical Engineering and Computer Science, 2022, 26(1): 434-441.

[2] Gupta M, Kumar P. Recommendation generation using personalized weight of meta-paths in heterogeneous information networks. European Journal of Operational Research, 2020, 284(2): 660-674.

[3] Hobbs R. Propaganda in an age of algorithmic personalization: Expanding literacy research and practice. Reading Research Quarterly, 2020, 55(3): 521-533.

[4] Blin K, Shaw S, Kloosterman A, Charlop-Powers Z, Wezel G, Medema M, Weber T. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic acids research, 2021, 49(W1): W29-W35.

[5] Soffer O. Algorithmic personalization and the two-step flow of communication. Communication Theory, 2021, 31(3): 297-315.

[6] Ding S, Du W, Li C, Xu X, Wang L, Ding L. Density peaks clustering algorithm based on improved similarity and allocation strategy. International journal of machine learning and cybernetics, 2023, 14(4):1527-1542.

[7] Cui Z, Jing X, Zhao P, Zhang W, Chen J. A new subspace clustering strategy for AI-based data analysis in IoT system. IEEE Internet of Things Journal, 2021, 8(16): 12540-12549.

[8] Karim M, Beyan O, Zappa A, Costa I, Rebholz-Schuhmann D, Cochez M, Decker S. Deep learning-based clustering approaches for bioinformatics. Briefings in bioinformatics, 2021, 22(1): 393-415.

[9] Liu X, Li M, Tang C, Xia J, Xiong J, Liu L, Kloft M, Zhu E. Efficient and effective regularized incomplete multi-view clustering. IEEE transactions on pattern analysis and machine intelligence, 2020, 43(8): 2634-2646.

[10] Zou H. Clustering algorithm and its application in data mining. Wireless Personal Communications, 2020, 110(1): 21-30.

[11] Arabi H, Balakrishnan V, Shuib N. A Context-Aware Personalized Hybrid Book Recommender System.Journal of Web Engineering (JWE), 2020, 19(3/4):405-428.

[12] Huixiang X, Xiaomin L, Yueyan L. Group Recommendation Based on Attribute Mining of Book Reviews. Data analysis and knowledge discovery, 2020, 4(2/3): 214-222.

[13] Sarma D, Mittra T, Hossain M S. Personalized book recommendation system using machine learning algorithm. International Journal of Advanced Computer Science and Applications, 2021, 12(1):2121-219.

[14] Zhou Y. Design and Implementation of Book Recommendation Management System Based on Improved Apriori Algorithm.Intelligent Information Management, 2020, 12(3):75-87.

[15] Kwak W, Noh Y. A study on the current state of the library's AI service and the service provision plan. Journal of Korean Library and Information Science Society, 2021, 52(1): 155-178.

[16] Bindhu V, Ranganathan G. Hyperspectral image processing in internet of things model using clustering algorithm. Journal of ISMAC, 2021, 3(2): 163-175.

[17] Hu L, Zhang J, Pan X, Luo X, Yuan H. An effective link-based clustering algorithm for detecting overlapping protein complexes in protein-protein interaction networks. IEEE Transactions on Network Science and Engineering, 2021, 8(4): 3275-3289.

[18] Oyewole G J, Thopil G A. Data clustering: Application and trends. Artificial Intelligence Review, 2023, 56(7): 6439-6475.

[19] Liu S, Jin S. 3-D gravity anomaly inversion based on improved guided fuzzy C-means clustering algorithm. Pure and Applied Geophysics, 2020, 177(2): 1005-1027.

[20] Chen D. Automatic vehicle license plate detection using K-means clustering algorithm and CNN. Journal of Electrical Engineering and Automation, 2021, 3(1): 15-23.

[21] Li P, Xie H. Two-stage clustering algorithm based on evolution and propagation patterns. Applied Intelligence, 2022, 52(10): 11555-11568.

[22] Nitu P, Coelho J, Madiraju P. Improvising personalized travel recommendation system with recency effects. Big Data Mining and Analytics, 2021, 4(3): 139-154.

[23] Zhao G, Liu Z, Chao Y, Qian X. CAPER: Context-aware personalized emoji recommendation. IEEE Transactions on Knowledge and Data Engineering, 2020, 33(9): 3160-3172.

[24] Guo Y, Mustafaoglu Z, Koundal D. Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms. Journal of Computational and Cognitive Engineering, 2023, 2(1): 5-9.

[25] Sarkar A, Biswas A, Kundu M. Development of q-Rung Orthopair Trapezoidal Fuzzy Einstein Aggregation Operators and Their Application in MCGDM Problems. Journal of Computational and Cognitive Engineering, 2022, 1(3): 109-121.

[26] Wang M, Zhang Y Y, Min F, Deng L, Gao L. A two-stage density clustering algorithm[J]. Soft Computing, 2020, 24(23): 17797-17819.

[27] Sibille L, Civera M, Zanotti Fragonara L, Ceravolo R. Automated Operational Modal Analysis of a Helicopter Blade with a Density-Based Cluster Algorithm[J]. AIAA Journal, 2023, 61(3): 1411-1427.

[28] Hassan B A, Rashid T A, Mirjalili S. Formal context reduction in deriving concept hierarchies from corpora using adaptive evolutionary clustering algorithm star[J]. Complex & Intelligent Systems, 2021, 7(5): 2383-2398.

[29] Anand S K, Kumar S. Experimental comparisons of clustering approaches for data representation[J]. ACM Computing Surveys (CSUR), 2022, 55(3): 1-33.

[30] Zubaroğlu A, Atalay V. Data stream clustering: a review[J]. Artificial Intelligence Review, 2021, 54(2): 1201-1236.

Workforce Planning for Cleaning Services Operation using Integer Programming

Workforce Modelling in Service Industry

Mandy Lim Man Yee[1], Rosshairy Abd Rahman[2], Nerda Zura Zaibidi[3],
Syariza Abdul-Rahman[4], Norhafiza Mohd Noor[5]
School of Quantitative Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia[1, 3, 5]
Institute of Strategic Industrial Decision Modelling-School of Quantitative Sciences,
Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia[2, 4]

***Abstract*—The cleaning services industry in Malaysia faces significant challenges in effectively managing its workforce. Workforce planning, a critical procedure that aligns employee skills with suitable positions at the right time, is becoming increasingly essential across various organizations, including postal delivery and cleaning services. However, the absence of proper workforce planning from management teams has emerged as a primary concern in this sector. This study identifies an opportunity to improve the workforce planning in the cleaning industry by employing an optimization approach that aims to minimize hiring costs. The main objective of this study is to minimize hiring costs in cleaning services operations at a public university in Malaysia. To achieve this, an optimization model based on integer programming was proposed to represent the current situation. Data collection involved interviews and company reports for the purpose of understanding the current conditions comprehensively. Factors influencing hiring costs were meticulously selected, considering the organization's specific situation. Model evaluation was conducted through what-if analysis, which allowed the evaluation of solutions provided by the modified models in three what-if scenarios. The findings indicated that the proposed modified model could assist organizations in improving the workforce planning by optimizing the allocation of resources, reducing hiring costs, and enhancing cleaner performance. This study offers valuable insights for the management of cleaning services, paving the way for more effective and efficient workforce planning practices in the industry.**

Keywords*—Workforce planning; cleaning services industry; optimization approach; integer programming

I. INTRODUCTION

In a company or an organization, it is important for the management to plan the workforce. Workforce planning is a procedure that assigns employees with the appropriate skills to suitable positions at the right time [1, 2, 3]. The workforce planning issue is pervasive across various industries, such as services, healthcare, education, military, transportation, and many more. These industries are listed in Table I, with the related studies highlighted by different authors.

Previous studies identified and discussed the issues in workforce planning in different industries. For instance, in the services industry, workforce planning challenges were discussed in study [4], where service technicians from various locations must go to the clients' locations, resulting in escalated costs and inefficiencies. These frequent travels not only imposed financial burdens but also hindered service delivery effectiveness due to time constraints. To address these concerns, the researchers introduced a workforce planning model designed to mitigate travel costs for service technicians, concurrently enhancing the overall efficiency of operations. In fact, other industries, such as healthcare, education, military, and transportation, are facing similar workforce issues. Therefore, as these workforce planning issues impact various industries, it will be beneficial to study further especially related to the service industry, since the nature of service is wide and intangible, which will be complex to measure.

TABLE I. RELATED STUDIES ON WORKFORCE PLANNING ISSUES IN DIFFERENT INDUSTRIES

Industry	Workforce Planning Issues	Author
Services	Service technicians located at different locations are required to travel along to the customers' locations.	[4]
Healthcare	Traveling issues happened for the medical staff, including transporting the medicines and medical equipment between warehouses and patients.	[5]
	The systemic delay happened in the hospital consultation service.	[6]
	Scheduling nurses on three shifts per day based on various preferences and constraints.	[7], [8]
Education	Allocating academic staff in quest of optimum knowledge transfer.	[9]
Military	To deploy the right employees in the right place at the right time.	[10]
	To meet the demands of the current force structure under conditions of uncertainty in officer retention.	[11]
Transportation	A pilot can be hired as a first officer or a captain, the company need to determine whether how many persons of pilot need to be hired.	[12]

A cleaning services operation is one of the popular services in the service industry. This service is provided by a contractor to an individual's or an organization's place or property. They typically take on these undesirables but necessary duties, as opposed to a person or employee who would be less than happy about the chance to deal with the dust or waste [13, 14, 15]. The cleaning services industry in Malaysia is facing

numerous challenges when it comes to effectively managing its workforce. One significant challenge is determining the ideal number of cleaners required to complete cleaning tasks within a specified timeframe. In this case, a specific timeframe refers to the required working time of the cleaners assigned by the organization. The specific timeframe ensures that the cleaners complete their tasks within a designated period, allowing for efficient scheduling and resource allocation. However, this often leads to overstaffing, resulting in increased of hiring costs, or understaffing, which can adversely affect the quality of cleaning services provided. Moreover, the existing workforce's performance in the cleaning services operation is unsatisfactory, translating into higher costs for the management team, who must hire additional workers to compensate for the inefficiencies [16]. This will finally lead to general issues in the workforce for a cleaning service operation, i.e. the high turnover rate [17]. The management team's absence of proper workforce planning seems to be the primary cause of this issue. As such, there is an opportunity to improve the workforce planning in the cleaning industry by utilizing a systematic approach to minimize hiring costs and maximize cleaners' performance. Consequently, this study was aimed to minimize the hiring cost at a cleaning services operation responsible for taking care of property-related matters and maintenance in a public university in Malaysia. A campus environment that is hygienic, healthy and pleasant plays a crucial role in shaping the learning experience and the academic performance of students [18]. As a big campus, the university requires cleaning services to manage the whole campus environment well. However, it is all aware that the cleaning service at the university can still be improved in order to achieve a satisfyingly clean and comfortable study area.

The paper is structured into distinct sections to address workforce planning challenges within the cleaning services industry comprehensively. The initial segment outlines the industry's issues and the absence of adequate planning, setting the study's goal: minimizing hiring costs within a university's cleaning services operation. Subsequently, various methodologies and influential factors in workforce planning are explored in the Literature Review which is given in Section II. The Research Methodology in Section III details the study's phases, from problem definition to model evaluation. Finally, the Findings in Section IV present the results of evaluating different scenarios and examining their impact on hiring costs and operational efficiency and lastly Section V concludes the paper. Each section contributes to a comprehensive understanding of optimizing workforce planning in cleaning services.

II. Literature Review

Workforce planning involves analyzing current and future workforce needs as a way of ensuring that it has the right people with the right skills in the right place and at the right time. The significance of workforce planning for organizations has been emphasized in numerous studies. Workforce planning is a continuous process that aids organizations in aligning their workforce with their business objectives and priorities [19]. Additionally, it is closely linked to the broader business planning process. There are three categories of the workforce planning methods, judgmental, mathematical, and a combination of the two [20]. Mathematical methods for the workforce planning fall under the realm of operations research. In a review of the current state of the workforce planning, an imbalance in the field was noted [21].

A. Factors in Workforce Planning

A range of factors with a significant impact must be considered for workforce planning to be effective. Factors affecting workforce planning, such as skills, demand and time windows, have been identified from previous studies and tabulated in Table II.

TABLE II. Factors Affecting Workforce Planning

Factors	Author
Skill requirements	[4], [6], [10], [12], [22],[23], [24], [25], [26], [27]
Workforce demand	[4], [5], [6], [9], [22], [24]
Time windows	[4], [5], [27], [28]
Cost	[4], [8], [29]
Budget	[5], [27]
Experience / Year of services	[10], [30]
Career advancements	[10], [30]
Geographical location	[25]
Physical resources planning	[9]
Language	[25]
Potential for retraining	[25]
Quantity to produce	[28]
Contract type	[5]
Workforce retention rates	[11]
Preference for the working period	[4]
Maximum tasks assigned to the worker	[4]
Workforce holidays	[12]

The table provides a clear overview of the most commonly cited factors in the workforce planning problem by researchers. The results suggest that a significant number of studies prioritize the "skill requirement" as a primary constraint in the workforce planning, indicating the importance of having a skilled workforce to meet organizational goals. Additionally, workforce demand is a critical factor that organizations must consider in their planning. As the second most commonly cited factor, it highlights the importance of accurately estimating the needs of future workforce. Organizations must ensure they have enough resources to meet the required workload while avoiding under or overstaffing, which can lead to an increase in operational costs.

Moreover, time windows are also crucial in the workforce planning, as they help organizations allocate resources efficiently within a specified timeframe. Organizations must ensure that they have enough staff available during peak periods and that they can accommodate fluctuations in demand. Overall, the information presented in the table emphasizes the importance of considering these constraints in the workforce planning. Organizations must prioritize these factors while establishing their workforce planning strategies to ensure they

have the right people with the necessary skills available at the right time to achieve their objectives.

B. Method for Improving Workforce Planning

In the research on workforce planning, the authors proposed several innovative methods to address its associated challenges. The authors' work sheds light on the importance of having a robust workforce planning strategy that considers various factors, such as workforce demand forecasting, skill requirements, time windows, etc. The details of these methods are comprehensively discussed and analysed in Table III.

TABLE III. METHOD FOR IMPROVING WORKFORCE PLANNING

Method	Author
Linear programming	[22]
Mixed integer linear programming	[4], [5], [12]
Integer programming	[31]
Goal programming	[9], [11]
Genetic algorithm	[9], [23]
System dynamics	[8], [23]
Discrete-Event Simulation Model	[11]
Stochastic programming	[30]
Robust optimization	[11]

Optimization is often the preferred method for solving the workforce planning issues in cleaning service operations. This is because it can help minimize costs and increase efficiency while ensuring all cleaning tasks are completed on time. An optimization approach is a mathematical method used in a variety of fields to identify the best option from a list of feasible solutions. It involves working through various solutions to maximize or minimize an objective, frequently under certain constraints. These approaches include methods such as linear and nonlinear programming as well as integer programming. Linear programming is a mathematical technique used to optimize complex problems by determining the best solution to a given set of constraints. It is particularly useful in the workforce planning because it can help organizations allocate their resources efficiently to meet the required workload. The research in [22] applied it in the workforce planning and to reduce manufacturing costs at the same time. Mixed-integer linear programming (MILP) is an extension of linear programming that allows for the inclusion of integer variables in addition to continuous variables. It is useful in the workforce planning because it can help organizations allocate their resources optimally while considering additional constraints, such as the availability of part-time workers or the minimum number of shifts an employee need to work. For instance, MILP is employed in addressing the workforce planning problems that involve task duration as a continuous variable in the constraints, as highlighted by [4, 5]. In another study conducted by study [12] demonstrated the effectiveness of MILP in addressing the workforce planning problems that involve pilot holidays as a continuous variable in the constraints. The study showed that MILP was able to produce a highly effective performance in this regard.

Meanwhile, for a complex problem, Genetic Algorithm (GA) can be used to address the workforce planning problems in a dynamic environment where demand, resources and other constraints are continuously changing [9]. It can also be used to optimize multiple objectives simultaneously such as maximizing productivity while minimizing labour costs. In a study by [23], GA seeks the optimal workforce flow between different ranks. However, it is important to note that GA has limitations, where there is a chance that GA will become trapped in local optima, making it difficult to find the overall optimal solution, which is normally addressed as a near-optimal solution.

Other than optimization, Goal Programming is another popular mathematical programming technique that can help organizations optimize their workforce planning strategies by simultaneously considering multiple objectives or goals. According to [9], the study involves constraints like physical resource planning, for instance, library, laboratory, etc., which must be considered simultaneously. Therefore, it is suitable to use goal programming in solving the problem.

In addition, System dynamics (SD) and Discrete-Event Simulation (DES) are also frequently used in the workforce planning issues. SD is a modeling approach that is useful in workforce planning because it allows for the analysis of complex systems with feedback loops, time delays, and other dynamic factors. The SD model helps simulate the workforce's career progressions from recruitment to interim separation and retirement [6], [23]. Meanwhile, DES is useful in workforce planning because it allows for the analysis of complex systems with a large number of discrete events that occur over time. In workforce planning, DES models can be used to simulate the behavior of the workforce planning system under different scenarios and policies. Although effective for evaluating complicated workforce planning, SD has a limitation when constructing an accurate SD model requires a thorough comprehension of the underlying dynamics of the system, which can be resource-intensive and time-consuming. Similarly, DES has limitations despite being good at capturing discrete events in workforce planning. When simulating complex systems with numerous events, DES models can become computationally difficult and time-consuming, thus restricting their application in real-time decision-making scenarios.

Additionally, as per [30], Stochastic Programming (SP) provides a range of modeling methods to incorporate uncertainty into manpower planning problems. Moreover, Robust Optimization (RO) techniques offer valuable insights as they do not require specific probability distributions for uncertain parameters. These methods might be useful for workforce planning; however, they also bring disadvantages. For instance, SP, while addressing uncertainty in manpower planning, can sometimes lead to complex mathematical formulations that are computationally intensive and challenging to solve, particularly for large-scale workforce scenarios. Including probabilistic components could also need a significant amount of historical data for precise parameter

estimates, providing limits in circumstances when the data is limited or unreliable. Similarly, RO can sometimes produce overly conservative solutions, potentially underutilizing resources or failing to capture the full range of possible outcomes, despite their advantages in handling uncertainty without exact probability distributions. Particularly in situations where finding a balance between robustness and optimization is essential, the practical implementation of RO might require careful evaluation of the trade-off between robustness and optimization performance.

In conclusion, there are numerous methods available for solving workforce planning problems. However, it is essential to select a method that is best suited to meet the constraints and requirements of the specific problem at hand. In this study, the use of integer programming (IP) is considered appropriate. This mathematical optimization approach enables the optimization of the objective function while adhering to constraints in terms of integer values only. The decision variables in the optimization problem can only assume integer values, making IP particularly useful when variables are restricted to discrete values. In the services industry, IP is not commonly used in solving problems. However, IP has been used in other industries such as research by [31], who investigated a hospital contact center that proposed a workforce planning framework using IP to optimize staffing. The decision variables are restricted to be integers in this method, which is the number of agents assigned to each shift. Another research in the education industry was conducted by [32] to allocate students to preferred lecturers for supervision of internships among undergraduate students. Since all the decision variables are integers, IP was the most suitable method to consider. The exact algorithms used in these models help solve the problem efficiently and accurately.

Returning to this study, the decision variable is an integer, i.e. the number of cleaners assigned. Thus, by utilizing IP, the optimal combination of variables satisfies the constraints while achieving the highest possible objective value, which is to minimize hiring costs, can be achieved.

III. Research Methodology

This study aims to develop an IP model for cleaning services operations in order to improve their current workforce planning. This section discusses on how this research is conducted in achieving the objective. The research is designed in a few phases of research activities that help to ensure progress throughout the study. These phases are problem definition, data collection, data analysis, results, and discussion, as shown in Table IV. The detail of each step is briefly discussed in the following subsection.

A. Phase 1: Problem Definition

In the current scenario, cleaners are being assigned by the cleaning services operation based on the location of buildings. This means they will only be placed at one location to concentrate and clean for the duration of the day. These cleaners will work regardless of the type of cleaning task, but only in the assigned building. The cleaners are compensated as full-time workers with work duration of nine hours per day and receive monthly wages. As an initial study, three academic buildings have been considered as the main focus of this study, which involves 25 cleaners. These buildings are the newest buildings in the university and thus appear to be the most organized and systematic. The number of cleaners is quite large when comparing the size of the building to the number of employees, which causes a high turnover rate. Therefore, this cleaning service tries to search for possibilities to reduce the rate and increase profit. Therefore, in order to improve the rate, this research experimented with different scenarios.

TABLE IV. Method for Improving Workforce Planning

Phase	Methods and Techniques
Phase 1: Problem Definition	1. Spotting the real-world problem. 2. Setting the research topic, objectives, and overall project plan. 3. Review of the literature.
Phase 2: Data Collection	1. Gather the potential factors that affect the workforce planning. 2. Interview sessions with the company representative. 3. Reviewing company reports.
Phase 3: Data Analysis	1. Select the suitable factors. 2. Develop an optimization model for the workforce planning. 3. Construct an objective function. 4. Construct mathematical formulation for the constraints.
Phase 4: Model Evaluation	1. Performing model evaluation based on what-if analysis using Microsoft Excel Solver 2. Giving suggestions to the cleaning services operation's management based on the findings.

B. Phase 2: Data Collection

The data collection process for this project involves two main sources: interview sessions with company representatives and analysis of company reports. Interviews provided valuable insights into the current situation of an organizational structure, job requirements, staffing levels, skills, competencies and task types within the cleaning services operation. These insights guide identifying areas for improvement and developing strategies to address the workforce planning challenges. Company reports offer critical data on workforce size, experience levels, pay scales, performance metrics, and other factors influencing the workforce planning. Analyzing this data helps organizations understand their current workforce state, identify skill gaps and make informed decisions for more effective planning while considering the available financial resources. The company report serves as a reference to comprehend company issues and financial performance, aiding in determining available resources for the workforce planning efforts.

C. Phase 3: Data Analysis

In this phase of the study, the focus shifts to selecting critical factors that will play a pivotal role in shaping the objective function of the workforce planning model. With these factors in mind, the study proceeds to construct an IP model that captures the essence of the workforce planning problem. The model integrates the factors seamlessly, facilitating the development of mathematical equations and constraints that reflect the real-world considerations of the cleaning services operation.

1) Factors selection: This study carefully selected several factors to be used as constraints in formulating an objective function. These factors have been chosen based on previous research, as discussed in Section II(A), and how well they align with the specific problem being investigated in this cleaning services operation. Table V shows the factors to be considered in the model formulation of the study.

TABLE V. FACTORS TO BE CONSIDERED

Factors	Definition
Size of the cleaning area	The total size of the area to be cleaned in the campus of the university.
Task duration	The given time frame for the cleaners to complete the task.
Scheduling	To ensure that the given cleaners are available in the time period.
Experience level	To ensure the cleaners efficiently perform the cleaning task within the given time.

Considering the size of the cleaning area as one of the important factors in the workforce planning affects the number of cleaners needed to clean up the area efficiently. A larger facility typically requires more cleaners to handle the workload, while a smaller facility may require fewer cleaners. Task duration is another important factor in the workforce planning for a cleaning service operation because it directly impacts the amount of time that a cleaner will need to spend on a particular task. In this study, the cleaners are given a time frame and are obliged to complete all the tasks allocated to them within the particular time frame.

In addition, effective scheduling helps to optimize the use of available staff and resources, reducing hiring costs and ensuring that the cleaning services are delivered on time and to the required standard. Moreover, experience level is important in the workforce planning for cleaning services operations to ensure that the premises are cleaned and maintained to the desired standard. Cleaning tasks requiring specialised skills or a high level of experience are more likely to be successfully completed by cleaners with sufficient experience within the allotted time frame. This is because it helps cleaners know how to do different cleaning tasks effectively, solve problems, and train new cleaners, resulting in better quality services that meet customer expectations and promote business growth for instance, carpet cleaning, computer lab cleaning, plant care, etc.

2) Model formulation: The problem for this study is now being addressed through an IP model, which incorporates the four factors discussed previously. Subsequently, the mathematical model is developed along with the corresponding constraints.

Indices:

Cleaner: $\{1, \ldots, i, \ldots, I\}$

Task: $\{1, \ldots, j, \ldots, J\}$

Time period: $\{1, \ldots, t, \ldots, T\}$

Parameters:

Cij: The cost of assigning cleaner i to task j at time t

Sj: The size of the area to be cleaned for task j

Ai: The maximum area that cleaner i can clean in one time period

Rj: The required experience level for task j

Ei: The experience level of cleaner i

Variables:

$Xijt = 1$ if cleaner i is assigned to task j at time period t, and $Xijt = 0$ otherwise.

Objectives:

In this model, we consider the objective of minimizing hiring costs, which can be formulated as:

$$\text{Minimize } \sum i\in I \sum j\in J \sum t\in T \; Cijt \; Xijt \quad (1)$$

Constraints:

$$\sum i\in I \sum t\in T \; Xijt \geq 1, \text{ for all } j\in J \quad (2)$$

$$\sum j\in J \; Xijt \leq 1, \text{ for all } i\in I, \; t\in T \quad (3)$$

$$Xijt \geq 0, \text{ for all } i\in I, \; j\in J, \; t\in T \quad (4)$$

This model is formulated to present the current scenario in the cleaning services operation, in which the cleaners are assigned to each building according to the total size of the cleaning area regardless of the skill requirements for the cleaners.

The objective function for this study is shown in Eq. (1). The aim is to minimize the hiring cost of the cleaning services. It is calculated by multiplying the hiring cost by the number of cleaners. Eq. (2) ensures that each task must be assigned to at least one cleaner during its entire duration. Eq. (3) ensures that each cleaner can perform at most one task during each period. Eq. (4) limits all the variables to a non-negative value. This model is then analysed to obtain the solution of the existing scenario.

D. Phase 4: Model Evaluation

In this study, the model evaluation was conducted using what-if analysis. What-if analysis is a process by which we adjust the model and then examine how those modifications will impact the model's results. In a study, there will typically be what-if analysis reflecting several situations. To better solve the problem, it would be good to see how the current results compare to the modified scenario. By conducting what-if analysis, it can gain insights into the potential outcomes and make informed decisions based on the model's sensitivity to different variables. This approach allows us to explore various scenarios and understand the implications of different adjustments on the overall results. This study has three modified scenarios: Model 2, Model 3, and Model 4. The current situation in Model 1 involves the cleaners working nine hours daily for a wage of 1,500 MYR per month. Consequently, the other three scenarios are as follows.

1) Model 2: What if the cleaner is assigned according to the size of the cleaning area?

A book on Setting Household Standards states that it takes four hours to clean 2000 sqft, which is an average of 500 sqft per hour [33]. However, the current assignment in the cleaning services operation is only focused on assigning the cleaners to the building areas without taking any consideration. In this modified scenario, the fewest cleaners possible are used to keep the building clean. Model 2 is the modification of the current Model 1 by adding the new constraint it to ensure the whole building is clean as shown in Eq. (5)

$$\sum j \in J \; Sj \; Xijt \leq Ai, \text{ for all } i \in I, \; t \in T \quad (5)$$

This equation is to ensure that each cleaner should clean at least a given size of area. To ensure that the cleaners can complete the assigned workload, the cleaning area multiplied by the number of cleaners should be larger than the total size of the cleaning area.

2) Model 3: What if the cleaners work in a part-time mode?

Model 3 is an improvement idea modified from the previous model. In this scenario, the cleaners are considered to be part-time workers, and therefore they will work for four hours and receive 30 MYR as their wages per day. The cleaner is assumed to clean 2000 sqft per day. It is possible to decrease the hiring cost due to the number of cleaners needed. Besides, according to the cleaning services operation, the cleaners might need only four hours to clean the building area. Therefore, cutting down the paid wages might help in minimizing the cost.

3) Model 4: What if the cleaner is assigned according to the task type?

In this scenario, similarly, the cleaners are considered to be part-time workers, and they will work for four hours and receive 30 MYR as their wages per day. This scenario was modified in Model 4 to retain all the current cleaners. In this scenario, the cleaners are assigned according to task type without considering the size of the cleaning area.

There are four different task types that make up the cleaning tasks: "General cleaning and maintenance," "Washroom maintenance," "Specialize cleaning," and "Plant care." Table VI shows the cleaning tasks included in the given task types. Each cleaner's performance level was gathered and converted to a range of experience levels. This is due to no available data on the cleaners' experience level. In this study, cleaners with more experience were assumed to perform their jobs more effectively. This assumption allows us to assign suitable cleaners to the given types of task.

Depending on their current experience level, the cleaners will be assigned a task type; for example, if their experience level is below 2, they will be given the task type "General cleaning and maintenance." Then, if they have an experience level of 2, the task will be "Washroom maintenance," and an experience level of 3, "Special cleaning," respectively. However, "Plant care" will only be assigned to cleaners with experience levels above 4, which can call for more patience and competence.

Since every hired cleaner should get one assigned task, therefore a new constraint as shown in Eq. (6), is modified from Eq. (3) in this model.

$$\sum j \in J \; Xijt = 1, \text{for all } i \in I, \; t \in T \quad (6)$$

Eq. (6) ensures that each cleaner must be assigned one task during the given time period. At the same time, a new equation, as shown below, is added to this model in order to ensure that the cleaner manage to fulfill their given task in the given time span as they have relevant skills for their given task.

$$Xijt = 0, \forall i \in I, \forall j \in J, \forall t \in T \text{ where } Rj > Ei \quad (7)$$

TABLE VI. TASK TYPES AND THE CLEANING TASKS

Task Type	Cleaning Task
General cleaning and maintenance	Cleaning desks and chairs, dusting furniture and fixtures, emptying trash bins, sweeping and mopping the floor, cleaning corridors, cleaning stairwells, cleaning windows and mirrors
Washroom maintenance	Cleaning the washroom and refill the supply
Specialized cleaning	Cleaning lift, cleaning computer set, cleaning carpet
Plant care	Trimming and pruning the plants, fertilizing the plants, watering the plant

Eq. (7) enforces that only cleaners with sufficient experience levels can be assigned to the tasks that require specific skills, which ensures that the tasks are completed properly. If the required experience level for task j is higher than the experience level of cleaner *i*, then cleaner *i* cannot be assigned to task *j* during the time period *t*.

IV. FINDINGS

This section discusses the results and findings obtained from the developed models. The model evaluation will be made by comparing the current model with all the modified models.

A. Discussion for Model 1 (Current scenario)

Model 1 was created to reflect the current situation of cleaning services operation. The size of the cleaning area and the task types were not taken into consideration when cleaners were assigned according to buildings. The three academic buildings taken into consideration, named as buildings X, Y and Z, received these cleaners randomly, assigned by the cleaning services operation. These cleaners would do any task types in the assigned building during nine working hours. Their performance level was monitored by the supervisor from the cleaning services operation. The number of cleaners involved in the current existing scenario and their performance are shown in Table VII. These data are provided by the cleaning services operation and will be compared later to the modified scenario.

The number of hired cleaners is multiplied by their monthly wages to determine the hiring cost for this model. In this model, there are 25 cleaners assigned, and their monthly pay is 1,500 MYR. In this situation, the total cost of hiring is 37,500 MYR. Reducing hiring costs is necessary since the high hiring costs strain the cleaning services operation.

B. Model Evaluation

Model evaluation is an important phase in this study, involving the comparison of the current model with modified models. The objective is to evaluate each model's advantages and limitations in order to find a better solution that suits the organization's needs. By assessing the models' performance in different scenarios, management will gain insights into how well the models address the problem. Microsoft Excel's Data Solver is used to run these models, with a computer powered by AMD Ryzen 7 5800X along with 16GB RAM.

TABLE VII. NUMBER OF CLEANERS INVOLVED AND THEIR PERFORMANCE IN THE CURRENT SCENARIO

Cleaner	Performance	Cleaner	Performance
Cleaner 1	80%	Cleaner 14	70%
Cleaner 2	95%	Cleaner 15	90%
Cleaner 3	70%	Cleaner 16	80%
Cleaner 4	75%	Cleaner 17	70%
Cleaner 5	70%	Cleaner 18	70%
Cleaner 6	85%	Cleaner 19	85%
Cleaner 7	80%	Cleaner 20	70%
Cleaner 8	85%	Cleaner 21	70%
Cleaner 9	90%	Cleaner 22	75%
Cleaner 10	70%	Cleaner 23	90%
Cleaner 11	80%	Cleaner 24	70%
Cleaner 12	80%	Cleaner 25	80%
Cleaner 13	85%		

Discussion for Model 2: Due to the high cost of the cleaning services operation to hire cleaners in the current scenario, Model 1 is used as a guideline to modify and come out with a new model in different scenarios. In this scenario, the size of the cleaning area is considered while assigning cleaners. The allocated cleaners must clean at least 14463 sqft for Building X, 12006 sqft for Building Y and 7320 sqft for Building Z. It took about 49 minutes to come up with the optimal solution, as shown in Table VIII.

TABLE VIII. THE ASSIGNMENT OF CLEANER IN MODEL 2

Cleaner	Assigned Location	Cleaner	Assigned Location
Cleaner 1	Building X	Cleaner 14	Not assigned
Cleaner 2	Not assigned	Cleaner 15	Not assigned
Cleaner 3	Not assigned	Cleaner 16	Building Z
Cleaner 4	Not assigned	Cleaner 17	Not assigned
Cleaner 5	Not assigned	Cleaner 18	Not assigned
Cleaner 6	Not assigned	Cleaner 19	Building Y
Cleaner 7	Not assigned	Cleaner 20	Building X
Cleaner 8	Not assigned	Cleaner 21	Building Y
Cleaner 9	Not assigned	Cleaner 22	Building X
Cleaner 10	Not assigned	Cleaner 23	Building Z
Cleaner 11	Not assigned	Cleaner 24	Building X
Cleaner 12	Not assigned	Cleaner 25	Building Y
Cleaner 13	Not assigned		

According to the result, four cleaners were assigned to Building X, three were assigned to Building Y, and two were assigned to Building Z. In a nutshell, total of nine cleaners were assigned to do the cleaning tasks at these buildings. If this model was used, the cleaning services operation would have to pay a monthly hiring cost of 13,500 MYR. Due to the decreased number of cleaners, this value is lower than the existing model. The result is graphically displayed in Fig. 1.

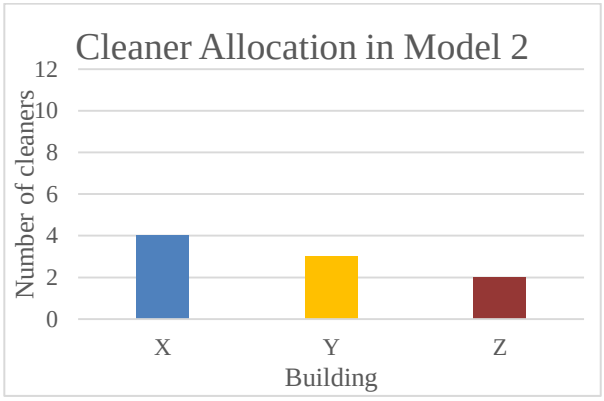


Fig. 1. Cleaner allocation in building based on Model 2.

1) Discussion for Model 3: Based on the interview with the cleaning services operator, the cleaners might only spend four hours cleaning the building despite being obligated to work nine hours every day. Therefore, it would be wise if we could reduce their working hours so that they could increase their level of efficiency. The optimal solution required a long running duration of 19 days. The assignment of cleaners in Model 3 is shown in Table IX.

TABLE IX. THE ASSIGNMENT OF CLEANER IN MODEL 3

Cleaner	Assigned Location	Cleaner	Assigned Location
Cleaner 1	Building X	Cleaner 14	Building X
Cleaner 2	Building Z	Cleaner 15	Building X
Cleaner 3	Building Z	Cleaner 16	Not assigned
Cleaner 4	Building Y	Cleaner 17	Building Y
Cleaner 5	Building Z	Cleaner 18	Not assigned
Cleaner 6	Building X	Cleaner 19	Not assigned
Cleaner 7	Building Y	Cleaner 20	Building Y
Cleaner 8	Building X	Cleaner 21	Not assigned
Cleaner 9	Building X	Cleaner 22	Not assigned
Cleaner 10	Not assigned	Cleaner 23	Building Y
Cleaner 11	Building Z	Cleaner 24	Building X
Cleaner 12	Building Y	Cleaner 25	Building Y
Cleaner 13	Building X		

Based on the result from Model 3, eight cleaners were assigned to the Building X. Besides, seven cleaners were assigned to Building Y while four cleaners were assigned to the Building Z. In this model, a total of 19 cleaners were assigned to carry out the cleaning tasks. The cleaning services operation might spend 570 MYR on the daily hiring cost for these three buildings. Since cleaners would work for 26 days per month, the total hiring cost in this model is 14,820 MYR, which is lower than the current hiring cost. However,

compared to Model 2, the hiring cost would be a little higher because this model might require more cleaners. Fig. 2 presents the cleaner allocation in Model 3.

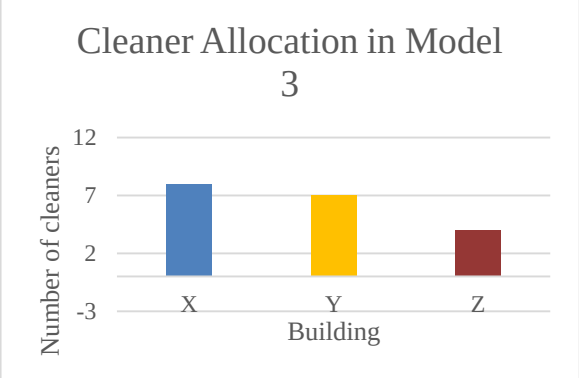


Fig. 2. Cleaner allocation in building based on Model 3.

2) Discussion for Model 4: As in the previous scenario of Model 3, all the cleaners remain considered as part-time workers in this scenario and will work four hours per day for 30 MYR. This scenario aims to enhance the cleaners' performance level, where they are assigned based on specific tasks. Cleaners should perform more effectively when carrying out a task that they are familiar with; thus, cleaners may result in higher performance levels since they are more expert with the task at hand. Therefore, the cleaners are assigned to a given task type based on their experience level in this model. Additionally, the cleaning services operator gets to maintain all the hired cleaners by employing this scenario. The assignment of cleaners in Model 4 are shown in Table X.

Based on the result from Model 4, 11 cleaners were assigned to "General cleaning and maintenance". Besides, 6 cleaners were assigned to "Washroom maintenance" while 4 cleaners were assigned to "Specialized cleaning". At the same time, the 4 cleaners with the highest performance level were assigned to "Plant care". In this scenario, all the hired cleaners would remain with the cleaning services operation and be assigned with cleaning tasks. They might spend 750 MYR on the daily hiring cost for these three buildings. Since cleaners will work for 26 days per month, the total hiring cost in this model is 19,500 MYR, which is lower than the current hiring cost. However, compared to Model 2 and Model 3, the hiring cost will be the highest as it tends to remain all cleaners in the organization. 52 minutes of time is taken to run this model and finding the optimal solution. The cleaner allocation in this Model 4 is clearly presented in Fig. 3.

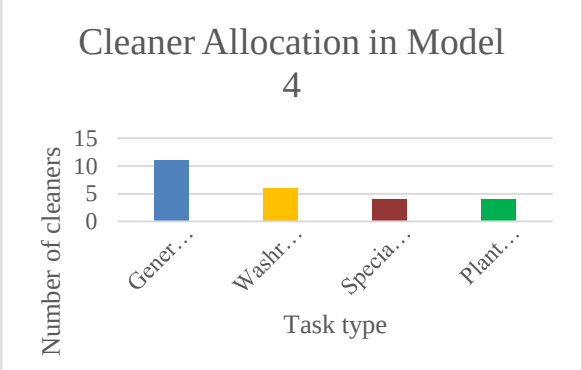


Fig. 3. Cleaner allocation in building based on Model 4.

TABLE X. THE ASSIGNMENT OF CLEANER IN MODEL 4

Cleaner	Experience level	Task type
Cleaner 1	2	Washroom maintenance
Cleaner 2	5	Plant care
Cleaner 3	0	General cleaning and maintenance
Cleaner 4	1	General cleaning and maintenance
Cleaner 5	0	General cleaning and maintenance
Cleaner 6	3	Specialized cleaning
Cleaner 7	2	Washroom maintenance
Cleaner 8	3	Specialized cleaning
Cleaner 9	4	Plant care
Cleaner 10	0	General cleaning and maintenance
Cleaner 11	2	Washroom maintenance
Cleaner 12	2	Washroom maintenance
Cleaner 13	3	Specialized cleaning
Cleaner 14	0	General cleaning and maintenance
Cleaner 15	4	Plant care
Cleaner 16	2	Washroom maintenance
Cleaner 17	0	General cleaning and maintenance
Cleaner 18	0	General cleaning and maintenance
Cleaner 19	3	Specialized cleaning
Cleaner 20	0	General cleaning and maintenance
Cleaner 21	0	General cleaning and maintenance
Cleaner 22	1	General cleaning and maintenance
Cleaner 23	4	Plant care
Cleaner 24	0	General cleaning and maintenance
Cleaner 25	2	Washroom maintenance

3) Model comparison: In this section, the result of all the models is compared and shown in Table XI.

TABLE XII. COMPARISON OF RESULTS

Model	Scenario	Number of Cleaners	Hiring Cost (MYR)
1	Cleaners are assigned randomly to the building.	25	37,500
2	What if the cleaner is assigned according to the size of the cleaning area?	9	13,500
3	What if the cleaner works in a part-time mode?	19	14,820
4	What if the cleaner is assigned according to the task type?	25	19,500

Originally, Model 1 required a high hiring cost of 37,500 MYR with 25 cleaners. Then, the first what-is scenario of Model 2 maintains normal working hours but reduces hiring costs by assigning nine cleaners based on cleaning area size, resulting in a 13,500 MYR cost. Meanwhile, Model 3 aims to enhance performance by minimizing cleaner working hours and employing 19 part-time cleaners for four hours daily at 14,820 MYR/month. On the other hand, Model 4 enhances cleaner performance via task specialization, retaining all cleaners. However, it has the highest cost at 19,500 MYR. It can be seen that these models offer distinct approaches, where Model 2 focuses on area size, Model 3 on efficiency, and Model 4 on specialization.

V. CONCLUSION

This study explores the workforce planning issues in cleaning service operations and provides helpful recommendations. Integer programming (IP) was used to create an optimization model to minimize hiring costs with a focus on three academic buildings in a public university in Malaysia. The current model with three different scenarios was investigated according to several identified factors, such as cleaning area size, task duration, scheduling, and experience level. Each model focused on a different approach and came out with improved hiring costs. Based on the analysis of scenarios, cleaning services operations can choose any model suited to their budget and future planning. However, of all models, Model 3 is recommended due to its lower cost and part-time working hours, which can benefit students and organizational efficiency.

This study contributes significantly to knowledge and practice, though limited by scope and time constraints. As a way to address constraints and improve model evaluation and effectiveness, the research highlights the necessity for heuristic methodologies in future work, especially involving the broader scenarios such as the whole building in a university. In fact, more input factors need to be added to future models for more reliable results.

ACKNOWLEDGMENT

This research was supported by Universiti Utara Malaysia through University Grant (S/O Code 21417). We would like to thank the Research and Innovation Management Centre for facilitating the management of this research work.

REFERENCES

[1] J. Stokker, and G. Hallam, "The right person, in the right job, with the right skills, at the right time: A workforce - planning model that goes beyond metrics," Library Management, vol. 30, no. 8/9, pp. 561-571, 2009, doi: 10.1108/01435120911006520.

[2] P. Yogita, and T. Shruti, "Work Force Planning, Literature Analysis: Digitization compels for a Conceptual Model for Data Driven Decisions," Journal of Business and Management, vol. 19, no.11, pp. 1-11, 2017, doi: 10.9790/487X-1911060111.

[3] A. S. Al Wahshi, "Human resource planning practices in the Omani Public Sector: An exploratory study in the Ministry of Education in the Sultanate of Oman," 2016.

[4] M.P. Doan, J. Fondrevelle, V. Botta-Genoulaz, and J.F.F. Ribeiro, "Studying the impact of different work contract combinations on a multi-objective workforce planning problem," 2019 International Conference on Industrial Engineering and Systems Management (IESM), pp. 1-6, 2019, doi: 10.1109/IESM45758.2019.8948220.

[5] T. Garaix, M. Gondran, P. Lacomme, E. Mura, and N. Tchernev, "Workforce scheduling linear programming formulation," IFAC-PapersOnLine, vol. 51, no. 11, pp. 264-269, 2018, doi: 10.1016/j.ifacol.2018.08.289.

[6] G. Willis, S. Cave, and M. Kunc, "Strategic workforce planning in healthcare: A multi-methodology approach," European Journal of Operational Research, vol. 267, no. 1, pp. 250-263, 2018, doi: 10.1016/j.ejor.2017.11.008.

[7] R. Ramli, S.N.I. Ahmad, S. Abdul-Rahman, and A. Wibowo, " A tabu search approach with embedded nurse preferences for solving nurse rostering problem," International Journal for Simulation and Multidisciplinary Design Optimization, vol. 11, no. 10, pp. 1-10, 2020, doi: 10.1051/smdo/2020002.

[8] R. Ramli, R. Abd Rahman, and N. Rohim, "A hybrid ant colony optimization algorithm for solving a highly constrained nurse rostering problem," Journal of Information and Communication Technology, vol. 18, no. 3, pp. 305-326, 2019, doi: 10.32890/jict2019.18.3.8292.

[9] S. Gupta, and S. Sinha, "Academic Staff planning, allocation and optimization using Genetic Algorithm under the framework of Fuzzy Goal Programming," Procedia Computer Science, vol. 172, pp. 900-905, 2020, doi: 10.1016/j.procs.2020.05.130.

[10] H. H. Turan, S. Elsawah, and M.J. Ryan, "Simulation-based analysis of military workforce planning strategies", In Proceedings of the 2019 International Conference on Management Science and Industrial Engineering, pp. 68-75, 2019, doi: 10.1145/3335550.3335568.

[11] N.D. Bastian, C.B. Fisher, A.O. Hall, and B.J. Lunday, "Solving the army's cyber workforce planning problem using stochastic optimization and discrete-event simulation modeling," In 2019 Winter Simulation Conference (WSC), pp. 738-749, 2019, doi: 10.1109/WSC40007.2019.9004837.

[12] İ. Z. Akyurt, Y. Kuvvetli, M. Deveci, H. Garg, and M. Yuzsever, "A new mathematical model for determining optimal workforce planning of pilots in an airline company," Complex & Intelligent Systems, vol. 8, no. 1, pp. 429-441, 2022, doi: 10.1007/s40747-021-00386-x.

[13] A. M. Galazka, and J. Wallace, "Challenging the 'dirty worker'—'clean client'dichotomy: Conceptualizing worker-client relations in dirty work," International Journal of Management Reviews, vol. 25, no. 4, pp. 707-724, 2023, doi: 10.1111/ijmr.12330.

[14] P. Hamilton, T. Redman, and R. McMurray, "'Lower than a snake's belly': Discursive constructions of dignity and heroism in low-status garbage work," Journal of Business Ethics, vol. 156, no. 4, pp. 889-901, 2019, doi: 10.1007/s10551-017-3618-z.

[15] J. Hughes, R. Simpson, N. Slutskaya, A. Simpson, and K. Hughes, "Beyond the symbolic: A relational approach to dirty work through a study of refuse collectors and street cleaners," Work, employment and society, vol. 31, no. 1, pp. 106-122, 2017, doi: 10.1177/0950017016658438.

[16] M. H. Ishak, and N. N. M. Anasir, "An Assessment of Cleanliness Level from Service Level Agreement and User's Perception in Universiti Tun Hussein Onn Malaysia," Research in Management of Technology and Business, vol.1, no.1, pp. 663-676, 2020, doi: 10.30880/rmtb.2020.01.01.050.

[17] S. D. Gilster, M. Boltz, and J. L. Dalessandro, "Long-term care workforce issues: Practice principles for quality dementia care," The

Gerontologist, vol. 58, no. suppl_1, pp. S103-S113, 2018, doi: 10.1093/geront/gnx174.

[18] Y.J. Utama, Purwanto, and Ambariyanto "Developing Environmentally Friendly Campus at Diponegoro University," Advanced Science Letters, vol. 23, no. 3, pp. 2584-2585, 2017, doi: 10.1166/asl.2017.8712.

[19] S. R. Cave, and G. Willis, "System Dynamics and Workforce Planning," System Dynamics: Theory and Applications, pp. 431-457, 2020, doi: 10.1007/978-1-4939-8790-0_659.

[20] M. Kunc, "Using systems thinking to enhance strategy maps," Management Decision, vol. 46, no. 5, pp. 761-778, 2008, doi: 10.1108/00251740810873752.

[21] P. De Bruecker, J. Van den Bergh, J. Beliën, and E. Demeulemeester, "Workforce planning incorporating skills: State of the art," European Journal of Operational Research, vol. 243, no. 1, pp. 1-16, 2015, doi: 10.1016/j.ejor.2014.10.038.

[22] M. Sivasundari, K. S. Rao, and R. Raju, "Production, capacity and workforce planning: a mathematical model approach," Appl. Math. Inf. Sci, vol. 13, no. 3, pp. 369-382, 2019, doi: 10.18576/amis/130309.

[23] H. H. Turan, S. Elsawah, F. Jalalvand, and M. J. Ryan, "Solving strategic military workforce planning problems with simulation-optimization," In 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1620-1625, 2020, doi: 10.1109/SSCI47803.2020.9308483.

[24] R. Martins, T. Pinto, and C. Alves, "An exact optimization approach for personnel scheduling problems in the call center industry," In International Conference on Computational Science and Its Applications, pp. 407-424, 2023.

[25] Y. Naveh, Y. Richter, Y. Altshuler, D. L. Gresh, and D. P. Connors, "Workforce optimization: Identification and assignment of professional workers using constraint programming," IBM Journal of Research and Development, vol. 51, no. 3/4, pp. 263-279, 2007, doi: 10.1147/rd.513.0263.

[26] A. M. Akl, S. El Sawah, R. K. Chakrabortty, and H. H. Turan, "A joint optimization of strategic workforce planning and preventive maintenance scheduling: a simulation–Optimization approach," Reliability Engineering & System Safety, vol. 54, no. 8, 2022, doi: 10.1016/j.ress.2021.108175.

[27] D. E. Ighravwe, S. A. Oke, D. Aikhuele, and A, Ojo, "An optimisation approach to road sanitation workforce planning using differential evolution," Journal of Urban Management, vol. 9, no. 4, pp. 398-407, 2020, doi: 10.1016/j.jum.2020.06.004.

[28] A. Zakariyya, M. S. Mashina, and Z. Lawal, "Application of linear programming for profit maximization in Shukura Bakery, Zaria, Kaduna State, Nigeria," Dutse Journal of Pure and Applied Sciences (DUJOPAS), vol. 8, no. 1a, pp. 112-116, 2022, doi: 10.4314/dujopas.v8i1a.12.

[29] V. H. Ferreira, P. D. M. Oliveira Filho, E. V. Queiroga, J. M. Silva, E. U. Barboza, T. P. Abud, B.S.M.C. Borba, M.Z. Fortes, B.S. Moreira, and P. H.C. Machado, "Two-phase optimization approach for maintenance workforce planning in power distribution utilities," Electric Power Systems Research, vol. 211, 2021, doi: 10.1016/j.epsr.2022.108236.

[30] N. D. Bastian, B. J. Lunday, C. B. Fisher, and A. O. Hall, "Models and methods for workforce planning under uncertainty: Optimizing US Army cyber branch readiness and manning," Omega, vol. 92, no. 3, 2019, doi: 10.1016/j.omega.2019.102171.

[31] M. Mikaeili, A. Erekat, J. Lee, and M. Khasawneh, "Manpower Planning Framework for a Hospital Contact Center with Service Level Requirements using Integer Programming," Institute of Industrial and Systems Engineers (IISE), vol. IIE Annual Conference, Proceedings, pp. 1187-1192, 2019.

[32] S. Faudzi, S. Abdul-Rahman, R. Abd Rahman, J. Zulkepli, and A. Bargiela, "Optimizing The Preference Of Student-Lecturer Allocation Problem Using Analytical Hierarchy Process And Integer Programming," Journal Of Engineering Science And Technology, vol. 15, no. 1, pp. 261 – 275, 2020.

[33] M. L. Starkey, " Mrs. Starkey's Setting Household Standards - The Key to Successful Service for Employers and Household Managers," United States: Starkey International Institute, 1999.

Tailored Expert Finding Systems for Vietnamese SMEs: A Five-step Framework

Thi Thu Le[1], Xuan Lam Pham[2*], Thanh Huong Nguyen[3]
Department of Research Methodology, Thuongmai University, Hanoi, Vietnam[1]
Department of Information Technology, School of Information Technology in Economics, National Economics University[2, 3]

***Abstract*—This study addresses the underexplored area of EFSs (EFS) tailored for business applications, with a specific focus on supporting Small and Medium Enterprises (SMEs). The principal objective of this research is to develop an EFS designed to cater to the needs of Vietnamese SMEs. The study methodology involves conducting in-depth interviews with Vietnamese SMEs to ascertain their requirements for Vietnamese EFSs. Subsequently, the research proposes an architectural model for the EFS and proceeds to develop the corresponding system. The EFS operates by collecting and analyzing data from diverse online sources to identify Vietnamese experts and individuals of Vietnamese origin who can provide valuable insights and support to enterprises operating in Vietnam. This research framework is guided by five key Husain (2019)'s issues: 1) Expertise evidence selection, 2) Expert representation, 3) Model building, 4) Model evaluation, and 5) Interaction design. By addressing these issues, the study aims to contribute to the development of an effective EFS tailored to the specific needs of Vietnamese SMEs in their quest to find and engage experts for business growth and innovation.**

Keywords*—Expert Finding System (EFS); Small and Medium-sized Enterprises (SMEs); experts; Vietnamese expert resources; business expertise identification

I. INTRODUCTION

SMEs play a key role in Vietnam and Asia's economic development, constituting up to 98% of businesses in Asia and providing about 66% of private-sector jobs from 2007 to 2012 [1]. Howerver, SMEs have to deal with a multitude of challenges, especially the issue of recruiting high-quality labor resources [2]. They employ various personnel sourcing methods, including costly headhunter services and specialized training companies [3]. Alternatively, businesses can utilize recruitment websites and job exchange platforms, but these predominantly cater to common job positions, offering limited information about experts [4, 5]. Another option is independent online searches via platforms like Google and Bing, which yield articles and related information rather than lists of suitable candidates. Specialized social networking sites like LinkedIn and academic networks such as ResearchGate and Academia provide detailed expert profiles. However, expert data is often private and not easily accessible, hindering effective comparisons and selections [4]. Lastly, personal referrals are highly dependent on social networks, varying in effectiveness. To address these issues, one solution is leveraging technology for human resources and production efficiency [6, 7]. Technology enables SMEs to access crucial information quickly for decision-making and strategic management. In Vietnam, the labor force mainly consists of unskilled, low-skilled, and medium-skilled workers, with only 10 million qualified workers, representing 21% of the total labor force [8]. Consequently, this shortage results in intense competition among SMEs to attract and retain qualified experts who can drive innovation and growth [9].

An Expert Finding System (EFS) for business in recruitment is a highly effective solution that leverages advanced technology and data analytics to identify and connect with top talent in a more efficient and targeted manner Husain, et al. [10]. By analyzing candidates' skills, experiences, and qualifications alongside the specific needs and requirements of the job, this system streamlines the recruitment process, reducing time and cost while ensuring that the best-fit candidates are brought to the forefront. This study is a significant attempt to address an often-overlooked aspect of EFS in the domain of business applications, with a specialized focus on supporting SMEs operating in Vietnam. The primary aim of this research is to develop an EFS system matched precisely to the needs of Vietnamese SMEs. This study employed an in-depth interview methodology to gain insights into Vietnamese SMEs' requirements for expert human resources. Subsequently, the study advances with the proposal of an architectural model for the EFS, followed by the actual development of the system. This EFS functions by collating and analyzing data from diverse online sources to pinpoint Vietnamese experts and individuals of Vietnamese origin who possess the expertise necessary to aid enterprises in Vietnam. Guided by the five pivotal issues identified by Husain, et al. [10], including expertise evidence selection, expert representation, model building, model evaluation, and interaction design, this research hopes to contribute significantly to the evolution of an effective EFS, custom-tailored to meet the unique demands of Vietnamese SMEs.

In particular, this research will examine two main research questions:

- Question 1: How do Vietnamese Small and Medium Enterprises (SMEs) perceive their need for expert human resources, and what are the key challenges they face in identifying and engaging experts to support their business growth and innovation and their requirements for Vietnamese EFSs?
- Question 2: How can an Expert Finding System (EFS) be optimized for Vietnamese SMEs by effectively addressing the five key issues outlined by Husain, et al. [10], enabling efficient identification and engagement

of Vietnamese experts and individuals of Vietnamese origin to support SMEs in Vietnam?

The remainder of this study is organized as follows. Section II explores work concerning the Expert Finding System. The Methodology in Section III describes the research design, survey participant information, and system design. The Results in Section IV presents the outcomes of the in-depth interview and the proposed framework of the Expert Finding System for Vietnamese SMEs. Finally, the last section i.e. Section V provides a summary of the results, discusses them, and indicates future work.

II. Expert Finding Systems

The development and deployment of EFS, have enabled users to discover and search for experts and high-quality human resources in various fields for collaboration or knowledge acquisition [11]. Singh in [12] conducted a study highlighting the essential need for organizations to swiftly identify experts in different domains. However, this task presents challenges due to limited and unevenly distributed information about experts. Furthermore, the requirements of those seeking experts are often unclear, and past expert performance lacks sufficient visibility, making it difficult to assess and quantify expertise. The dynamic nature of experts, who may switch jobs and research areas, further complicates expert identification. Additionally, many complex problems require the collective intelligence of diverse experts, underscoring the necessity of developing an expert finding information system to expedite problem-solving and enhance organizational efficiency [12].

Various studies have explored EFSs to address enterprise challenges and global issues [11, 13]. The development of these systems involves the discovery of experts in specialized fields, an examination of existing systems in these domains, and the proposal of models to guide future system design and development decisions [13]. Organizations can also leverage commercial services and solutions to identify experts and evaluate EFSs, taking into account their unique characteristics and the most effective implementation methods [11]. Moreover, In SMEs' pursuit of expert human resources, recent research indicates that social globalization positively influences gender equality in employment opportunities [14].

In Singh [12] study, the essential attributes of an ideal EFS are outlined. This comprehensive framework envisions a system that operates efficiently and effectively. Furthermore, it underscores the significance of categorizing expertise through a subject classification scheme, ensuring that users can easily access and navigate the wealth of knowledge available. Moreover, it strongly emphasizes the quantification of expertise, enabling the system to rank experts, thus aiding users in identifying the most suitable individuals for their specific needs. Ensuring the reliability of information sources is another aspect, ensuring the integrity and accuracy of the system's output. Moreover, the framework envisions fostering expert communities, facilitating collaboration and knowledge sharing among professionals. Lastly, the system is geared towards identifying locally available experts, enhancing accessibility and relevance for users seeking expertise within their geographic vicinity.

Furthermore, apart from research on EFSs for organizations, some studies delve into technical aspects related to collecting and processing expert information and algorithms for accurate expert identification, evaluation, and ranking. Taie, et al. [15] presents methods for classifying expertise, including domain-based classification (enterprise/organization and online community) and technique-based classification (machine learning and graph techniques). However, it is worth noting that combining content-based expertise indicators with social relationships does not completely resolve the issues related to expert identification and ranking [15]. Additionally, Singh [12] and Wang, et al. [16] highlight prominent expert finding methods, such as unstructured data mining, social networking sites, contact management systems, and self-disclosure data from experts.

According to Husain, et al. [10], EFSs have found application across various domains, including academics, enterprises, medicine, online knowledge-sharing communities, online forums, and social networks. In this study, Husain, et al. [10] asserts that defining specific tasks for expert finding systems in certain domains can be challenging. Consequently, there is a research gap in understanding the tasks that expert search systems support for businesses, especially small and medium-sized enterprises (SMEs). This gap will be the focus of the current study.

Moreover, while numerous studies have addressed technical aspects of EFSs, only a limited number of research into the specific needs of Vietnamese SMEs in locating experts in their respective fields to fulfill their human resource requirements effectively have been identified.

Lin, et al. [17] and Husain, et al. [10] have critically examined existing research in the field, highlighting several noteworthy gaps that demand attention and resolution. First and foremost, they emphasize that most studies in the field of EFSs have only focused on the academic domain, and very little is known about EFSs in other domains where expert knowledge is critical. Also, these studies bring up data-related problems that have not been properly dealt with yet, like security issues, data inconsistencies (like having multiple names for the same person or names that sound similar for different people), and the issues surrounding the completeness of expert information. Addressing these issues calls for the development of complex algorithms capable of seamlessly integrating data from diverse sources and tailoring expert-finding models to specific tasks and domains, a task that remains largely unmet. Moreover, Lin, et al. [17] and Husain, et al. [10] stress the imperative of fostering diversity in system development by creating combined datasets from multiple sources, promoting more versatile and robust solutions. Lastly, the identification of expert groups to solve interdisciplinary problems is highlighted as an essential requirement across various applications, underscoring the significance of collaborative and multidisciplinary approaches in expertise seeking. These identified gaps serve as critical pointers for future research, shedding light on the areas in need of further exploration and innovation within the realm of expert finding.

The expert finding task involves five key procedural aspects [10]:

- Expertise evidence selection: This step focuses on extracting data and information relevant to a person's expertise, which is crucial for determining their status as an expert in a particular field.
- Expert representation: EFS aim to provide users with information that aids in not only locating experts but also in selecting the most relevant ones. This requires identifying valuable information for decision-making, considering both documented evidence and contextual factors.
- Model building: Model building encompasses pre-processing, indexing, and modeling. Pre-processing involves handling diverse data sources, recognizing candidate expert identifiers (e.g., names, emails), and addressing challenges like named entity recognition and disambiguation. Modeling and retrieval involve creating models that associate candidate experts with user queries and rank them based on these associations, utilizing various methods such as probabilistic, network-based, and voting models.
- Model evaluation: Evaluating the efficiency of EFS is typically done using test collections (datasets) to assess their performance in retrieving relevant experts in response to user queries.
- Interaction design: Presenting expert search results to users is a critical practical concern. It involves displaying not only ranked lists of experts but also related documents, conferences, journals, and contact details to aid users in assessing the relevance of experts. Additional information like photos, affiliations, publications, and projects can help users gauge an expert's seniority and expertise alignment.

From an in-depth review of relevant EFS literature, we are dedicated to developing a tailored EFS for Vietnamese SMEs. This study will address the five key issues identified by Husain, et al. [10] and overcome the limitations mentioned by Lin, et al. [17]. The proposed EFS will incorporate all essential features from Singh [12] study, including robust data processing, advanced modeling, efficient retrieval, and a user-friendly interface. Our objective is to not only meet the specific requirements of Vietnamese SMEs but also set a global standard in EFS, adhering to best practices in the academic literature. Through this effort, we aim to enhance the expertise-seeking process for Vietnamese businesses, contributing to their growth and success in a dynamic marketplace.

III. METHODOLOGY

A. Research Design

The research process is structured into two distinct stages, each aligned with a research question.

For the first question, firstly, we conducted a review of EFSs to gain an understanding of the research problem and identify gaps in previous studies. Next, primary data was collected from businesses to gain a deeper understanding of the challenges they face when finding and using experts to support their activities. The following steps were taken:

- Determine research goals and questions
- Plan data collection
- Select some survey business
- Conduct interviews
- Analyze the interview results, focusing on identifying the problems that businesses are facing in the process of finding and using experts to support business activities.

For the second question, reliance is placed on the five key issues mentioned by Husain, et al. [10] with proposed methods for handling these issues:

- Identify details for each issue: Gather specific information about each of the 5 key issues highlighted by Husain, et al. [10].
- Collect relevant data: Gather data that is relevant to these issues.
- Develop a system: Create a system or framework to address each of the five key issues.
- Check solution evaluation: Evaluate the effectiveness of the proposed solutions for each of the five key issues. This can be done through data analysis, expert feedback, or user testing.

By following these steps, we can effectively provide practical solutions to the challenges faced in the process of finding experts.

B. Participants

The authors conducted in-depth interviews with 20 representatives of SMEs operating in various fields, including logistics, construction and architecture, information technology, and services (see Table I). Most of these enterprises are concentrated in Hanoi, with a few also located in Ho Chi Minh City or having branches in different areas. The number of permanent employees in each enterprise ranges from 10 to more than 200. The interviews focused on surveying the needs for general human resources and expert human resources in today's businesses. Additionally, the authors surveyed the need for information systems to support searching and linking with customers' human resources in different positions and locations. The interviews were mainly conducted using audio or video calls between December 2021 and January 2022. Most of the interview participants held important positions in the enterprises, such as General Director/Director, Head of department, or Executive Director. The data obtained from the interviews will be coded and synthesized according to groups of indicators, thereby synthesizing into a number of criteria to evaluate the current state of human resource needs in SMEs today. The survey data from the interviews related to how businesses seek sources of experts was used by the research team to propose solutions for building a system that meets the human resource needs of businesses. The system will provide expert information sources for businesses and create an information channel connecting businesses and experts in each field.

TABLE I. INTERVIEW CANDIDATES

#	Industry	Number of employees	Number of enterprises	Participants
1	Logistics	50-200	2	Delivery Director, Branch Head Manager
2	Information Technology	22-210	4	Director
3	Construction	30-600	7	Director, Deputy Director, Head of Department
4	Service	10-150	3	Director
5	Manufacture	23-120	2	Director
6	Architecture	10-30	2	Director

C. System Design

The system is built on a robust foundation that encompasses data collection from diverse sources, data integration, entity resolution, and indexing to ensure the accuracy and accessibility of expert information. The use of NoSQL databases for storing expert data is a practical choice, offering flexibility, scalability, and rapid performance, all essential for accommodating the growing volume of expert data over time. The model building process is systematic, involving input, pre-processing, entity creation, modeling and retrieval, and output generation (see Fig. 1), all aimed at providing users with a seamless experience in finding and connecting with experts. The emphasis on expert ranking is notable, as it helps users efficiently identify experts with substantial expertise and influence in their respective domains. The integration of a learning-to-rank algorithm for expert ranking is a forward-looking approach that promises to improve the quality of expert suggestions.

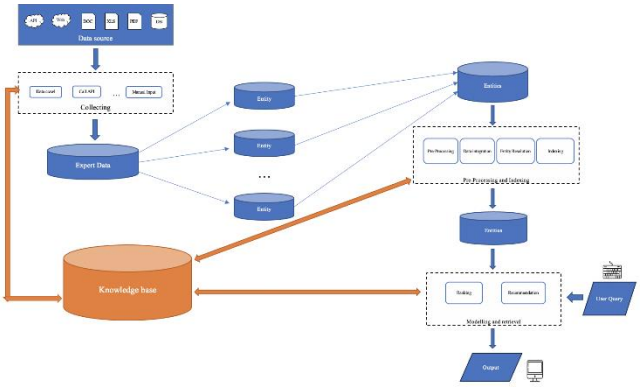


Fig. 1. System architecture model.

IV. RESULTS AND DISCUSSION

A. Understanding the Expert Resource Needs of Vietnamese SMEs

The interview results revealed several issues related to the needs and challenges businesses face when looking for experts, as follows:

1) Concerning the enterprise's available human resources: Assessing the organization's current human resources is the first step in determining the need for additional human resources. Does this human resource meet the business's work requirements during its operations?

According to the results of the interviews, the majority of businesses continue to lack significant human resources in numerous departments. Many positions, including senior managers, middle managers, department heads, team leaders, and employees, require additional personnel. In it, a respondent remarked:

"Positions that we often lack and need to recruit are middle managers and team leaders. However, finding the ideal candidate is not simple."

Most interview participants said that adding human resources is often done depending on the expansion needs of the business.

Typically, the missing human resources consist of both skilled and unskilled labor. All businesses reported a shortage of qualified personnel. One director of business explained:

"It is easier to supplement unskilled labor than highly skilled one. Finding and securing an agreement to collaborate with experts is time-consuming and difficult."

This result was also mentioned in the research of Thang and Nguyen [18], which noted the significant shortage of skilled and qualified human resources within Vietnamese businesses.

Therefore, it is clear that most businesses lack an adequate permanent workforce and must, therefore, consider expanding their human resources, particularly by adding expert personnel, to support their activities.

2) Regarding the importance of finding expert personnel: The majority of interviewees indicated that their organizations need external expert resources. However, companies have shown that the need to search depends on their operational requirements. These needs frequently arise when:

"The enterprise has a new project and requires experts to advise and support phases of the project or fulfill the requirements of fixed, recurring jobs."

Moreover, if a company wishes to expand its operations to a new international market, it will find Vietnamese experts residing in that country. A respondent listed the following as activities for which businesses seek assistance from specialists:

"Marketing, transportation, production, sales, and financial management" as well as "information and communication technology, legal, and risk management" are a few activities that need additional external human resources.

Besides, the European Commission [19] also points out that "*a virtual organization consists of a small core of fulltime employees and outside specialists temporarily hired as needed to work on projects.*" So, acquiring human resources with specialized knowledge from external sources, including foreign experts, is necessary.

3) Regarding the need for consulting and cooperation with experts: All interviewees indicated that they are willing to reach out to experts in the business's area of interest if they require expert advice or consulting services. These individuals also indicated that they had reached out to experts to request collaboration. This demonstrates that businesses desire assistance from individuals who have an in-depth understanding of the problems they are facing and need to solve.

In addition, quite a few instances have occurred when asking in greater detail about the outcomes of contacts with experts. Only a small portion of these got to collaborate. The majority of the remaining suggestions have been rejected to varying degrees.

"When I contact outside experts, they politely decline for a variety of personal reasons. Typically, this is due to an abundance of current work."

This is a significant obstacle for businesses when they want to collaborate with experts.

Thus, beyond the difficulties businesses encounter in finding and reaching out to experts, are there also impediments associated with collaborating with them? Several primary causes are listed, including:

"Not understanding the challenges faced by businesses"

"Unclear understanding of the business operating procedure."

"Unable to agree on conditions for cooperation"

Thus, it can be seen that another challenge businesses face when they want to collaborate with external human resources is how to help them understand the business's operating processes so they can provide the most effective and efficient support.

When interviewees were asked to choose between training existing workers themselves to meet the needs of the business and inviting experts to collaborate in each stage or project, more than half of the responses selected *"Use their employees"* rather than *"Invite experts to cooperate according to needs"* due to the many difficulties mentioned. An analysis of these processes reveals that, in addition to a fixed salary and other benefits, the implementation of the training workers will take a significant amount of time. Otherwise, if enterprises choose to work with experts who are available, they will only have to compensate based on the volume and effectiveness of their work.

Despite understanding the importance of locating and collaborating with experts, it is evident from the results of the interviews that there are numerous difficulties in the search process or contact with them.

4) Regarding the methods and tools used to find experts: According to interviewees, some methods and tools that businesses frequently use to find or contact experts are also critical issues (see Table II). So, they must be carefully considered and evaluated.

Surprisingly, all of the responding managers mentioned the method of finding personnel *"through relationships"* while the other methods were not used much. This leads to the conclusion that the current tools for supporting human resource search activities do not meet the needs of businesses adequately.

TABLE II. METHODS AND TOOLS THAT BUSINESSES FREQUENTLY USE TO FIND EXPERTS

Number of Participant	Methods/tools	Comments
5	Search engines	"Not very efficient. Finding suitable human resources remains a difficult challenge." "Low quality."
8	Job search websites	"At present, there are numerous human resources recruitment channels, but they frequently do not meet the appropriate needs or the quality is inconsistent, slow, or non-responsive." "It is time-consuming."
2	Expert finding systems	"Finding human resources that meet expectations is quite difficult." "The system is quite quick, but its effectiveness cannot be fully evaluated until a face-to-face interview is conducted. However, this is a preliminary initial step."
20	Interpersonal relationships	"This method is quite flexible, high efficiency but waste time and difficult to find number of experts at the same time ."

5) Concerning for finding experts in the field of interest: In light of the current situation that SMEs are facing in the process of searching for experts from outside, the authors have proposed to build an EFS that stores databases in the form of "metadata" about expert profiles for businesses to meet the needs of external personnel search and cooperation. A system with a large quantity of data about human resources who have in-depth knowledge in a variety of business-relevant areas. This system will be developed at no cost to establish a channel of communication between businesses and experts in various fields, and it can also introduce businesses to suitable experts from others if the system does not contain any relevant information.

To implement this idea, the authors conducted interviews with participants regarding system-related issues. The majority of interviewees stated that a specialized expert finding system for SMEs is essential. And the ideal system must provide the following essential features (see Table III):

Businesses have concerned the finding expert, with all of them expressing readiness to evaluate and offer feedback on this system once it's finalized.

D. Addressing Key Issues and Enhancing Efficiency of EFS for Vietnamese SMEs

Based on the insights gathered from the interviews, our objective is to develop an EFS that caters to the demands of locating proficient human resources for Vietnamese businesses. The system's database will be a substantial repository of big data integrated with the Knowledge Base System. The application of data processing techniques extends beyond facilitating information retrieval; it also delves into deeper data comprehension, uncovering interrelationships between data entities, and proffering actionable insights. This serves to motivate and support users in making informed decisions.

TABLE III. SEVERAL FEATURES REQUIRED IN EFS SYSTEMS DESIGNED FOR BUSINESS

#	Features	Requirements
1	Find an expert	Find the expert who most closely aligns with the user's query
2	Booking an Expert	Users submit a request for a specific job. The system will respond a list of potential experts suitable for the job position.
3	Connect or collaborate with experts	Submit a request for an appointment or propose collaboration with a particular expert
4	Introduce/Suggest experts	An expert can suggest another expert they are acquainted with for a customer's requirements.
5	Expert Q&A	Customers ask questions, experts answer.

Throughout the system development process, we gained valuable insights into the global landscape of Expertise Finding System implementations, with a particular emphasis on extensive research articles about EFS. Notably, a study conducted by Husain, et al. [10] has identified challenges within EFS systems that require resolution, including expertise evidence selection, expert representation, model development, model evaluation, and interaction design. These findings serve as a foundational framework for our endeavor to propose methods and techniques to address the aforementioned issues within our EFS system.

1) Expertise evidence selection: Expert information can be collected either manually or automatically, but it can only work against a pre-existing knowledge base (see Fig. 2). The knowlege base consists of a list of titles and positions, jobs by industry or field, high-ranking universities, and companies where many professionals work, certificates. The knowledge data comes from various sources drawn from predefined categories, such as the Yellow Pages directory, the research journals directory, the top-ranked universities directory, or the large-cap companies directory [20]. This dataset will be updated regularly during data analysis and processing.

Moreover, the Knowledge Base also facilitates the handling of Vietnamese data, including information pertaining to 3,000 career types in both Vietnamese and English languages, with synonym support. Additionally, it encompasses the names of nearly 200 countries, complete with geographical coordinates and names in Vietnamese as well as various other languages, along with phone numbers, to accommodate international customers utilizing global phone numbers. The Knowledge Base further includes 200 common Vietnamese names or surnames (unigrams) and almost 10,000 common Vietnamese surname combinations (bigrams) to aid in the identification of Vietnamese names. We have developed a Natural Language Processing (NLP) algorithm specifically designed for detecting Vietnamese names.

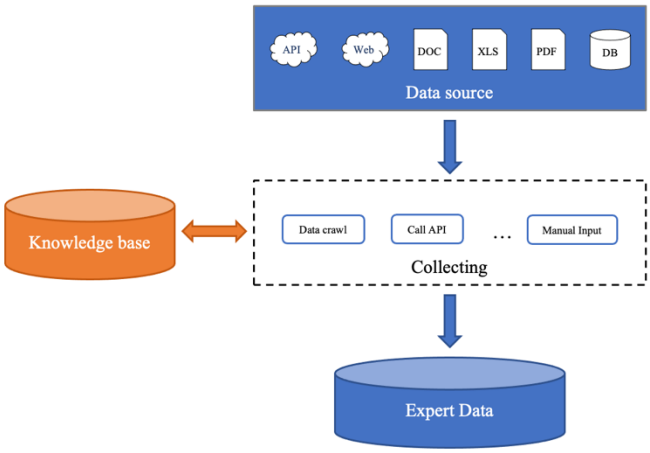


Fig. 2. Expertise selection.

The acquisition of data involves sourcing information from a diverse array of outlets, including the web, textual sources, social networks, APIs, and more. This process of gathering data is executed through a combination of manual and automated techniques. A proficient team will oversee the manual data collection aspect, while an automated data collection program will be deployed across multiple search servers. This automated system scans for pertinent information, extracts the requisite data, and deposits it into an unstructured database to facilitate subsequent analysis. In the context of data collection, the primary objective is to ascertain whether an individual possesses expertise in a specific field. To accomplish this, it is imperative to amass pertinent information pertaining to the individual.

Expert representation: Expert information needs to be effectively organized and stored, making it readily accessible and useful for various applications, such as expert ranking, collaboration, or expertise matching [21]. This information can be divided into the following collections:

a) Expert's Personal information

ID (Primary Key): An auto-incrementing unique identifier for each expert.

Name: The full name of the expert.

Title/Position: The expert's job title or position.

Contact Information:

Email: The expert's email address.

Phone: The expert's phone number.

Education: Details about the expert's educational background. This field can be more complex and structured, with subfields like:

Degree: The type of degree (e.g., Bachelor's, Master's, Ph.D.).

Institution: The name of the educational institution.

Year: The year of graduation.

Skills: A field to list the skills and areas of expertise possessed by the expert. This can be a text field or a list of keywords.

Areas of Expertise: A field to specify the main areas or subjects in which the expert has expertise. This can also be a text field or a list of keywords.

Awards and Honors: Any awards or honors received by the expert, with subfields like:

Award Name: The name of the award.

Year: The year the award was received.

Goal Vision: The expert's goal and vision.

Work Experience: Information about the expert's work history. This field can also be structured, with subfields like:

Position: The job title.

Company: The name of the company.

Start Date: The start date of the job.

End Date: The end date of the job.

Projects: Information about projects that were joined by expert.

Publications: A list of publications authored by the expert. This field can be structured with subfields like:

Title: The title of the publication.

Publication Date: The date the publication was released.

Books/Book Chapters: A list of books or book chapters authored by the expert.

Patents/Inventions: A list of patents or inventions authored by the expert.

b) Relationships

Collaboration, Association and Mentorship: The name of the organization, company or country the expert is affiliated with.

c) Presence

Social Media Profiles: Links or handles to the expert's social media profiles, if applicable.

Personal Website: Links to the expert's personal website.

d) Data Source

A list of data sources where expert information is collected.

Experts can be stored in a data schema that presents in the Fig. 3.

NoSQL databases stand out as an excellent choice for the storage of expert information, thanks to their versatility, scalability, and rapid performance [22]. NoSQL databases, including those employing key-value pairs and document-oriented structures, offer the flexibility to store expert information in a manner that is both adaptable and hierarchical. This facilitates effortless access and updates. Furthermore, these databases are engineered to handle substantial volumes of data and traffic without compromising on performance, rendering them well-suited for the enduring storage of expanding expert data over time. Notably, NoSQL databases excel in accommodating "one-to-many" relationships between entities, such as experts and their publications. As a result, all of an expert's details can be conveniently encapsulated within a single document, simplifying the retrieval and maintenance process.

Fig. 3. Data schema.

2) Model building: An expert finding system's model building process involves 2 steps [10] (see Fig. 4):

a) Pre-processing and Indexing: Cleaning, formatting, and transforming the raw expert data to make it consistent and suitable for analysis. This includes removing duplicates and irrelevant entries, standardizing data formats, handling missing data, tokenizing and normalizing text data, and creating indexes on specific fields or attributes to optimize query performance. Based on the Knowledge Base, the data will be classified, extracted into information fields, and concatenated to form experts.

b) Modelling and Retrieval: Developing a model that can analyze and rank the experts with support from the knowledge base. This model may use various techniques, such as machine learning, natural language processing, or weighted scoring, to evaluate and rank experts based on specific criteria. It takes user queries or criteria as input and searches the knowledge base for relevant experts. The retrieval system uses the model to score and rank experts based on how well they match the query, returning a list of suggested experts that meet the user's requirements.

This process enables users to efficiently discover and engage with experts in a given field or domain, and its effectiveness depends on the quality of data input, the accuracy of the model, and the efficiency of the retrieval mechanism.

Data cleaning and preprocessing are essential for ensuring accurate and reliable expert data. This involves standardizing

data formats, eliminating duplicates, handling outliers, and processing text data by removing special characters, tokenizing, eliminating stop words, and performing stemming/lemmatization. Encoding categorical data, scaling, and normalization are crucial. Finally, data splitting into training, validation, and test sets is necessary for building machine learning models. Data integration for expert information involves consolidating data from various sources to create a comprehensive dataset. This can be achieved through methods like data warehousing, ETL processes, data replication, data virtualization, and data federation. It ensures accessibility and accuracy in expert data for various applications. Entity resolution is vital for ensuring data accuracy in expert information. It involves identifying and consolidating records representing the same entity across diverse sources with variations in attributes. Preprocessing is essential to mitigate these variations. Indexing is used to improve search efficiency by adding indexes to relevant fields, allowing faster data access and reduced search times. It's especially useful in expert information systems for storing expert descriptions, background information, and relationships between experts. Multiple keywords and Boolean operators can be used to refine searches.

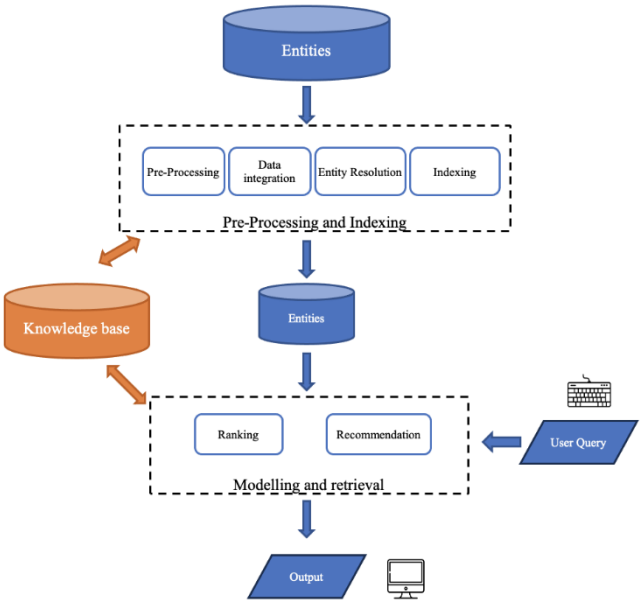


Fig. 4. An expert finding system's model building process.

Managing expert information involves understanding user queries, modeling them, retrieving relevant data, and providing accurate responses. This process begins with interpreting user intent through natural language understanding, query deconstruction, preprocessing, and semantic analysis. The technology also offers recommendation, reducing the need for users to enter complete queries by providing prompt suggestions related to service types, specialties, expert locations, company names, and individual names. The query is then modeled to create a structured representation, and information retrieval is performed, ranking and scoring results. We used Learning-to-rank algorithms to experiment with ranking experts, providing a list of suggested experts for various positions, saving time in the selection process.

3) Model evaluation: Various evaluation methods can be employed, depending on the specific function. For example:

a) Information retrieval: Precision, Recall, F1-Score, Mean Average Precision are used to evaluate the relevance and ranking of search results [23].

b) Recommendation: Precision at K (P@K), Mean Reciprocal Rank (MRR), or click-through rates are used to assess the quality of suggested queries [24].

c) Entity resolution: Measures like Precision, Recall, and F1-Score are used to evaluate the accuracy of entity resolution [25].

d) Ranking: Precision, recall, F1-Score, MAP, NDCG, or other relevant metrics are used to judge the quality of the ordered search results [26].

Moreover, domain experts with expertise in the system's relevant field offer subjective evaluations of its performance, assessing factors like the quality and relevance of search results, the precision of auto-suggestions, the efficiency of entity resolution, and the overall user experience.

4) Interaction design: Interaction design for the system, along with a booking feature, is a pivotal element in delivering a user-friendly and efficient experience. To start, the search functionality should be intuitive and easily accessible, empowering users to input their queries and apply filters to refine their results. The system should efficiently display a list of expert profiles after a search, offering a comprehensive summary of each expert's qualifications and availability (see Fig. 5). A clear and organized layout is essential to help users quickly identify the expert who best matches their needs.

Fig. 5. Advance expert search.

Upon selecting an expert profile, users should gain access to a detailed view of the expert's background, credentials, and reviews from previous clients. Additionally, the interaction design should incorporate easy-to-use booking features (see Fig. 6). Users should be able to select available time slots, input booking details, and confirm appointments with minimal effort. To prevent scheduling conflicts, real-time availability updates are crucial, ensuring a seamless and frustration-free booking process.

Our proposed system's information presentation on a world map offers users an extensive and user-friendly data visualization tool (see Fig. 7). Users have the flexibility to zoom in and out of the map to access statistical data across various geographic scales, from continents to individual towns. This feature not only empowers users to explore data in-depth but also provides a global perspective on the Vietnamese community's reach and activities. The automatic aggregation and real-time display of statistics on the map ensure that users have access to the most up-to-date and comprehensive information. This functionality enhances the system's usability and enables users to make informed decisions or gain a better understanding of the Vietnamese community's global presence.

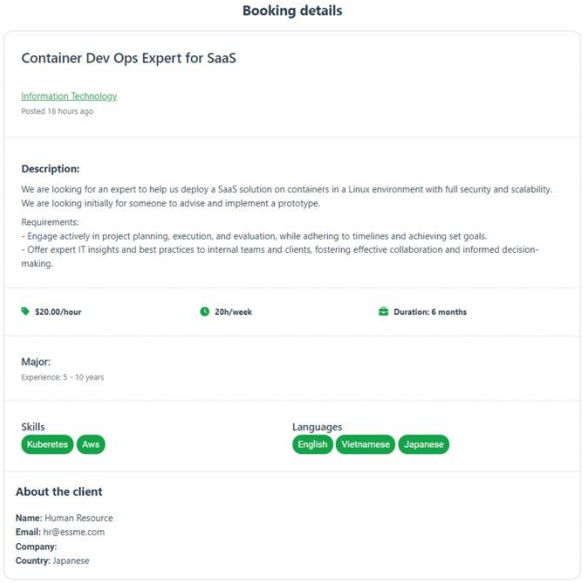


Fig. 6. Expert booking detail.

Fig. 7. Expert map.

V. Conclusion and Future Work

In conclusion, the purpose of the current study was to explore the EFSs customized for business applications, with a particular emphasis on SMEs operating in the Vietnamese context. By conducting in-depth interviews with Vietnamese SMEs, the study provides requirements and challenges these enterprises face when seeking expert resources. Subsequently, the research proposed an architectural model for an EFS for Vietnamese SMEs. This EFS automatically collect data from a multi sources to identify Vietnamese experts. Morover, this study is founded on a well-established framework that addresses five critical issues derived from Husain's seminal work, providing a roadmap for the development of an efficient and tailored EFS. Through the exploration of these key issues, the insights gained from this study may be of assistance to the creation of a specialized EFS capable of meeting the distinct needs of Vietnamese SMEs as they embark on their quest to locate and collaborate with experts in pursuit of business growth and innovation.

Further experimental investigations are needed to develop and refine the proposed EFS system. This should encompass the comprehensive design, construction, and testing of the system to ensure its seamless integration into the operations of Vietnamese SMEs. Continuous data collection and analysis must remain at the core, allowing the system's knowledge base to evolve in real-time and provide the most up-to-date expert information. Enhancement of the EFS model using advanced technologies such as machine learning and natural language processing will significantly improve the accuracy and efficiency of expert recommendations, catering to the unique requirements of Vietnamese SMEs. User feedback and extensive testing will fine-tune the system, making it even more user-friendly and effective for SMEs. In addition, an emphasis should be placed on security measures, scalability, and the potential expansion of the system to serve a broader range of industries and regions. Moreover, it is crucial to conduct long-term assessments to assess the lasting impact of the EFS on Vietnamese SMEs, tracking factors like increased efficiency, cost savings, and tangible business growth.

Acknowledgment

This research is funded by National Economics University, Hanoi, Vietnam.

References

[1] N. Yoshino and F. Taghizadeh Hesary, "Major Challenges Facing Small and Medium-Sized Enterprises in Asia and Solutions for Mitigating Them," Innovation Finance & Accounting eJournal, 2016. [Online]. Available: https://www.adb.org/publications/major-challenges-facing-small-and-medium-sized-enterprises-asia-and-solutions

[2] Q. A. Nguyen, G. Sullivan Mort, and C. D'Souza, "Vietnam in transition: SMEs and the necessitating environment for entrepreneurship development," Entrepreneurship & Regional Development, vol. 27, no. 3-4, pp. 154-180, 2015.

[3] V. K. Tuan, "Analysis of challenges and opportunities for Vietnamese SMEs in the globalization," Journal of Business Management and Economic Research (JOBMER), vol. 4, no. 2, pp. 169-185, 2020.

[4] N. Roulin and A. Bangerter, "Social Networking Websites in Personnel Selection," Journal of Personnel Psychology, 2013, doi: 10.1027/1866-5888/a000094.

[5] A. D. Sterling, "Friendships and Search Behavior in Labor Markets," Management Science, 2014, doi: 10.1287/mnsc.2013.1857.

[6] P. Akpamah and A. Matkó, "Information Technology deployment in Human Resource Management: A case study in deprived regions," (in English), International Review of Applied Sciences and Engineering, vol. 13, no. 1, pp. 1-10, 13 Oct. 2021 2021, doi: 10.1556/1848.2021.00278.

[7] D. Camey, "The Use of Information Technology Applications in Human Resource Management in Organizations," International Journal ofEmerging Research in Management & Technology, vol. 3, no. 3, 2014.

[8] D. T. Y. Nhi and H. C. Phuong, "Training highlevel human resources in the industrial revolution 4.0," in High-quality training program - theoretical and practical issues, Thu Dau Mot University, Binh Duong, 2018.

[9] N. H. Tien and B. Nogalski, "Developing high quality human resource to benefit from CP-TPP and IR 4.0," International Journal of Research in Management, vol. 1, no. 2, pp. 4-6, 01/22 2019.

[10] O. Husain, N. Salim, R. A. Alias, S. Abdelsalam, and A. Hassan, "Expert finding systems: A systematic review," Applied Sciences, vol. 9, no. 20, p. 4250, 2019.

[11] M. T. Maybury, "Expert Finding Systems," in "Technical Report MTR 06B000040," 2006.

[12] H. Singh, Singh, R., Malhotra, A., & Kaur, M. , "Developing a biomedical expert finding system using medical subject headings," Healthcare Informatics Research, 2013.

[13] D. Yimam-Seid and A. Kobsa, "Expert-Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach," Journal of Organizational Computing and Electronic Commerce, vol. 13, no. 1, pp. 1-24, 2003/03/01 2003, doi: 10.1207/S15327744JOCE1301_1.

[14] S. Ghosh, "How trade diversification and economic growth affect gender inequality in female labour market participation? The case of India," Journal of Economics and Development, vol. 24, no. 2, pp. 127-141, 2022, doi: 10.1108/JED-12-2020-0194.

[15] M. Taie, S. Kadry, and A. Obasa, "Understanding expert finding systems: domains and techniques," Social Network Analysis and Mining, vol. 8, no. 1, 08/30 2018, doi: 10.1007/s13278-018-0534-x.

[16] G. A. Wang, J. Jiao, A. S. Abrahams, W. Fan, and Z. Zhang, "ExpertRank: A topic-aware expert finding algorithm for online knowledge communities," Decision Support Systems, vol. 54, no. 3, pp. 1442-1451, 2013/02/01/ 2013, doi: https://doi.org/10.1016/j.dss.2012.12.020.

[17] S. Lin, W. Hong, D. Wang, and T. Li, "A survey on expert finding techniques," Journal of Intelligent Information Systems, vol. 49, no. 2, pp. 255-279, 2017/10/01 2017, doi: 10.1007/s10844-016-0440-5.

[18] N. Thang and H. Nguyen, "Importance of Human Resources in Building Sustainable Enterprises: Cases of Small and Medium Enterprises in Vietnam," 2021, pp. 277-296.

[19] European Commission, "Business management and organization," in "Global innovative leadership module," ERASMUS+, 2015. [Online]. Available: https://ec.europa.eu/programmes/erasmus-plus/project-result-content/9a1c8bee-11f3-48f0-8e25-c86b14cf445a/Business%20Management%20And%20Organization%20Booklet.pdf

[20] M. Brennan, D. Dilenschneider, M. Levin, and J. Robinson, "Finding and Researching Experts and Their Testimony," 2009. [Online]. Available: https://www.lexisnexis.com/documents/pdf/20071211111707_large.pdf

[21] R. Catherine, K. Visweswariah, V. Chenthamarakshan, and N. Kambhatla, PROSPECT: A system for screening candidates for recruitment. 2010, pp. 659-668.

[22] T. T. Le and X. L. Pham, "Towards NoSQL databases: Experiences from actual projects," in 2022 3rd International Conference on Big Data Analytics and Practices (IBDAP), 1-2 Sept. 2022 2022, pp. 15-20, doi: 10.1109/IBDAP55587.2022.9907664.

[23] K. Zuva and T. Zuva, "Evaluation of information retrieval systems," International Journal of Computer Science & Information Technology (IJCSIT) vol. 4, no. 3, pp. 35-43, 2012, doi: 10.5121/ijcsit.2012.4304.

[24] G. Shani and A. Gunawardana, "Evaluating Recommendation Systems," vol. 12, 2011, pp. 257-297.

[25] Z. Bahmani, L. Bertossi, and N. Vasiloglou, ERBlox: Combining Matching Dependencies with Machine Learning for Entity Resolution. 2015.

[26] J. Jiao, J. Yan, H. Zhao, and W. Fan, ExpertRank: An Expert User Ranking Algorithm in Online Communities. 2009, pp. 674-679.

A Zero-Trust Model for Intrusion Detection in Drone Networks

Said OUIAZZANE, Malika ADDOU, Fatimazahra BARRAMOU
ASYR RT, LaGeS Laboratory, Hassania School of Public Works, Morocco

***Abstract*—Today's worldwide introduction of drone fleets in a range of industrial applications has led to numerous network security issues, opening drones up to cyberthreats. In response to these challenges, an innovative approach has been proposed to protect drone fleet networks against potentially dangerous cyberattacks. Indeed, drones are considered as flying computers, and the proposed approach takes into account their complex network structure and communication protocols. The proposed system is designed around a multi-agent architecture, with a hybrid zero-trust detection mechanism against known and emerging cyberthreats. The CICIDS2017 dataset was exploited after performing some essential pre-processing tasks including data cleaning, balancing, binarization and dimension reduction. The proposed approach guaranteed high levels of accuracy and scalability, enabling an effective response to potentially dangerous cyber threat scenarios threatening drone fleets. To evaluate the effectiveness of the proposed system, a test portion of CICIDS2017 was used. The accuracy in recognizing benign network traffic reached 99.99% with a very low false alarm rate, ensuring the system's effectiveness against known and unknown cyber threats. Extensive experimental testing has been carried out on never-before-seen data, highlighting the system's remarkable ability to rapidly recognize cyber threats in real time, thereby enhancing the overall security of drone networks. The contribution of the proposed approach is significant for drone network security, as it introduces a comprehensive model designed to meet the specific security requirements of drone fleets. Finally, the proposed approach offers practical prospects for improving the security of drone applications.**

Keywords*—*Fleet of drones; security; zero trust; intrusions; cybersecurity; zero day; Multi-Agent

I. INTRODUCTION

Over the past few years, the use of drones has grown dramatically, as the Federal Aviation Administration estimates that there are between 2 and 7 million consumer drones in use by 2020 [1]. This major growth in drone usage is reflected in the projected trajectory of the drone industry, which is expected to reach an impressive $30 billion by 2036 [2]. The widespread accessibility and affordability of consumer drones has facilitated their adoption by hobbyists and enthusiasts alike. Consequently, airspace has seen a surge in unmanned aerial vehicles, leading to radical changes in aviation practices and raising important questions about security, privacy and regulatory measures [3].

Drones and drone fleets have become priceless assets for critical and sensitive situation management in a variety of fields. For example, their role in border surveillance is vital for strengthening security and control through the detection of illicit activities [4]. And when it comes to critical infrastructure, drones carry out inspections to identify vulnerabilities in assets such as power lines, bridges and pipelines, thus promoting early detection and preventive maintenance [5]. As for police forces, they use drones to monitor crowds at public events and investigate crime scenes, keeping officers safe. In addition, drones can also contribute to damage assessment, survivor location and rescue operations by providing real-time aerial data during disasters. For environmental monitoring, drones investigate sensitive areas and nature reserves to combat illegal activities, while studying the behavior of flora and fauna [6]. Finally, drones can contribute to healthcare by rapidly delivering medical supplies to inaccessible regions [7].

So, as mentioned, drones are therefore used to deal with more critical situations, as any bypassing of the drone network security can result in material and human damage, and even endanger human lives [8]. Nevertheless, the scientific community focused on communication protocols, battery autonomy and drone network use cases, and ignored the security aspect of this widely used technology [9].

As a result, addressing the security aspect is a crucial point to be taken very seriously by the scientific community, in order to protect drone networks against emerging cyber-attacks. The present work aims to propose a new intrusion detection approach based on a multi-agent architecture with a hybrid detection mechanism respecting the zero-trust principle.

The remainder of this paper is structured as follows: In Section II, we describe the background to the study, defining some concepts relevant to our study. Section II examines the current state of the art regarding the security of UAV networks and several proposed approaches to deal with intrusion detection in UAV networks. Section III presents an in-depth exploration of the proposed IDS architecture, detailing its components and its operational principle. Section IV describes the simulation and testing methodology of the proposed system. Section V delves into results and discussion and Section VI concludes the paper.

A. Drone

A "drone" is any unmanned aircraft, commonly known as an unmanned aerial vehicle (UAV), unmanned aerial system (UAS) or unmanned combat aerial vehicle (UCAV) [10]. For the purposes of this discussion, we focus on consumer/recreational UAVs, as illustrated in Fig. 1.

Fig. 1. Examples of consumer drones.

B. Fleet of Drones

1) Definition of a fleet of drones: A drone fleet is a coordinated and synchronized set of drones that cooperate to accomplish a given mission [11]. UAVs can be configured with various functionalities to satisfy a wide range of user needs. Managing a fleet of drones can involve complex operations, such as centralized control, mission planning, inter-drone communication and real-time data collection [12]. There are a wide variety of possible deployments for drone fleets, from surveillance and inspection to precision farming, logistics and delivery, and disaster relief. [13]

2) Communication modes of a fleet of drones: A drone fleet can be designed based on four possible communication architectures: centralized communication architecture, cellular communication architecture, satellite communication architecture and adhoc communication architecture [14] [15].

a) Satellite communication architecture: This architecture involves establishing communications between drones via satellite connections as shown in Fig. 2. Such links offer global coverage and are therefore particularly well suited to applications requiring wide coverage in remote or extensive areas. [16]

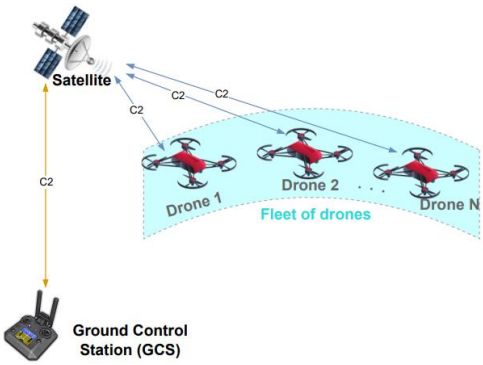


Fig. 2. Satellite communication architecture.

b) Ad hoc communication architecture: This type of communication is based on cooperation between drones in a dynamic, autonomous network, with no dependence on a static infrastructure, as shown in Fig. 3. Nearby drones interconnect autonomously, contributing to decentralized communications and collaborative data exchange, especially in scenarios where there is no conventional infrastructure. [17]

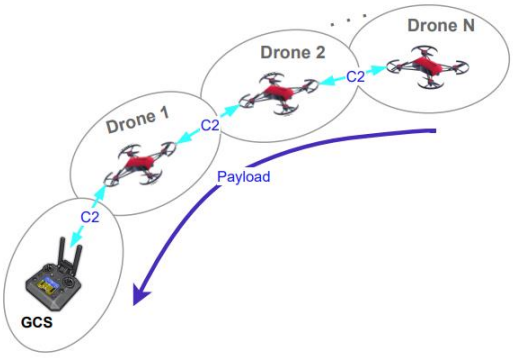


Fig. 3. Ad-hoc communication architecture.

c) Cellular communication architecture: A cellular architecture exploits existing cellular networks to ensure interaction between drones and ground stations as demonstrated by Fig. 4. Acting as roving nodes in the cellular network, drones ensure stable communication over extended distances, especially in urban environments or densely populated areas with established cellular coverage [18].

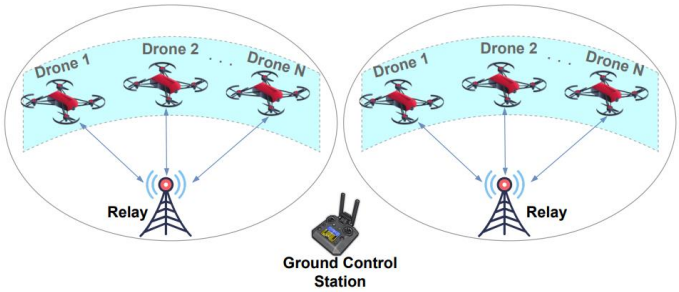


Fig. 4. Cellular communication architecture.

d) Centralized communications architecture: As shown in Fig. 5, this type of architecture relies primarily on a ground control station to manage communications. This configuration enables simplified coordination, efficient data processing and immediate decision-making, for scenarios requiring centralized control and supervision of drone fleets. [19]

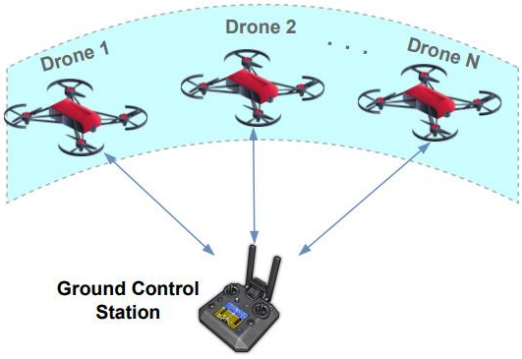


Fig. 5. Centralized communication architecture.

II. State of the Art

A. An Overview of Dangerous Tactics Targeting Drone Networks

Drone networks are vulnerable to a wide range of cyber-attacks. These attacks are all based on well-thought-out strategies, enabling hackers to achieve their malicious objectives, sometimes with life-threatening consequences. These attacks include:

- Communications jamming: The aim of communications jamming attacks is to disrupt or interfere with communications between UAVs and their control stations. Attackers can thus disrupt communication signals and cause loss of control, with serious consequences for ongoing operations. [20]
- DoS and DDoS attacks: This kind of attacks floods the UAV network with a large number of malicious requests, making it unavailable for legitimate communications. These attacks may lead to the interruption of critical business operations or even paralyze the network. [21]
- Data interception: Interception attacks capture sensitive data exchanged between drones and their associated control stations. This can lead to the exposure of confidential data and compromise mission confidentiality. [22]
- Usurpation of control: Spoofing is a serious threat that allows hackers to take control of a drone remotely, bypassing the legitimate control system. This can lead to malicious use of the drone for illegal or dangerous purposes. [23]

B. Attack Surface and Compromising Risk of a Drone Network

Drone networks are designed to cover large geographical areas and rely mainly on WiFi and radio waves as their communications medium [24]. Indeed, various types of communication can be involved in a network of drone fleets (see Fig. 6), including ad-hoc exchanges between drones, interactions with the control station and satellite links for GPS. Consequently, the attack surface of a drone network is considerable, exposing these networks to potential cybersecurity risks [25]. Given that anyone in the vicinity with a WiFi antenna/packet sniffer can potentially attempt to compromise the security principles of drone networks.

Drone manufacturers and researchers have mainly focused on developing communication protocols and improving battery life, while often ignoring the security aspect of drone networks. If the security of a drone network is compromised, numerous security risks can arise [9]. For example, a hacker could take control of the drone and gain access to its payload and the sensitive information it carries [21]. The hacker could also bring down the drone, resulting in property damage and the loss of the aircraft. In addition, the hacker could take control of the drone for malicious purposes, or integrate his own drone into the fleet, thus disrupting delivery and causing a denial-of-service issue [26].

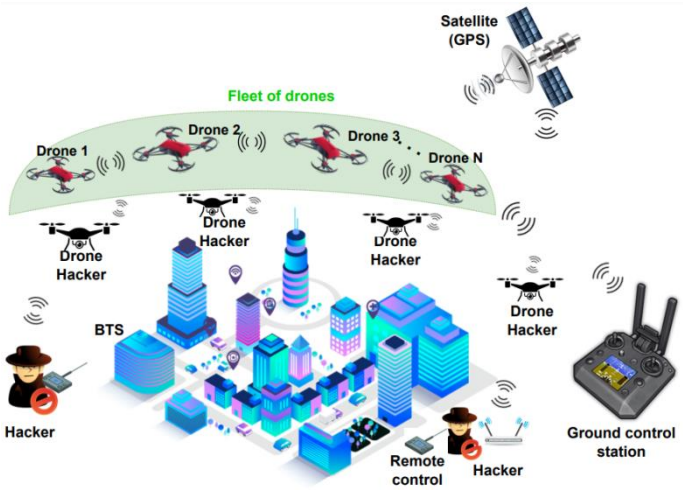


Fig. 6. Overview of a drone network and its attack surface.

C. Related Work

Recent research efforts in the field of intrusion detection systems for UAV fleets have seen the emergence of new, innovative approaches, each addressing distinct security issues and employing diversified methodologies. Fotohi, Abdan and Ghasemi (2022) [27] presented SID-UAV, an innovative system designed to counter the security risks associated with malicious drones, particularly in the context of drone-to-drone communication. In addition, Shrestha et al (2021) [28] proposed a UAV- and satellite-based 5G-network security model that can harness machine learning to effectively detect vulnerabilities and cyberattacks within a drone network. The proposed approach was suitable for both drone and satellite connections. Ouiazzane et al. (2022) [9] proposed a multi-agent intrusion detection system primarily designed according to a multi-agent architecture to counter denial-of-service (DoS) attacks targeting drone networks. The authors demonstrated remarkable accuracy in detecting DoS attacks while minimizing false alarms, underlining the system's effectiveness in protecting drone networks against a spectrum of known and unknown threats. Ihekoronye et al. (2022) [29] introduced a hierarchical intrusion detection system based on anomaly detection, suggesting an effective strategy for securing military UAV networks. They employed random cross-validation and finely tuned hyperparameters, guaranteeing resilience to network delays. The system considers payload constraints, battery limitations and the high mobility that characterizes Internet of Drones (IoD) networks. In a deep learning context, Abu Al-Haija and Al Badawi (2022) [30] proposed an autonomous intrusion detection approach adapted to drone networks. Such an approach makes use of deep convolutional neural networks to discern malicious activities within drone networks, efficiently processing encrypted Wi-Fi data records from commonly used drone brands such as Parrot, DBPower and DJI Spark drones.

Finally, Ouiazzane et al. took up the challenge of intrusion detection in UAV fleet networks in their 2020 study [14], focusing on ad hoc communication architectures. Their multi-agent system, enriched with detection mechanisms based on machine learning, underlines the importance of adaptable, intelligent security measures in the constantly evolving context

of drone fleet networks. These studies have made a significant contribution to the advancement of intrusion detection systems designed to meet the specific security challenges inherent in drone networks.

D. Discussion of Related Work Limitations

The state-of-the-art study carried out on drone security aims to understand the literature surrounding the topic. The security of a drone fleet is rarely addressed, even though these fleets represent the future trend in the use of drones for civilian missions. Most of the work cited in the state of the art has focused on routing protocols, autonomy optimization, and communication architectures while ignoring the security aspect, to which particular attention must be paid given the disastrous damage likely to occur whenever fleet security principles are successfully circumvented.

Little research has been carried out on the problem of intrusion detection in drone fleet networks. In fact, the approaches proposed by the community only address an aspect of the architecture or intrusion detection mechanism, and no complete work tackles all the aspects in question. Furthermore, there are no effective datasets for dealing with the problem of intrusion detection in drone fleets. In addition, even work based on machine learning sometimes uses polluted or obsolete datasets that fail to represent the real network traffic of a drone fleet. These factors lead to more attractive research avenues to further strengthen the security of drone fleet networks.

A drone, like a computer, is made up of a set of fundamental components. Among these, the central processing unit (CPU) and random-access memory (RAM) provide the computing power needed to execute the drone's tasks and missions. In addition, for communication with ground controllers and other aircraft, drones use diverse transmission methods, including WIFI and radio waves. Drones are also equipped with cameras to capture images and record video in their environment. Data storage is essential for drones to retain essential operating systems and files. Furthermore, drones are equipped with built-in sensors, such as GPS, that provide essential information on position and orientation. In the same way that computers depend on power sources, drones rely on batteries for their energy needs. Finally, drones integrate specialized aeronautical hardware enabling them to navigate in the air, while computers employ control peripherals for hands-on operations. To sum up, the analogy between drones and computers underlines the similarity of their components, asserting that the drone is a veritable flying computer.

Cyber security is becoming a major concern for drones and drone fleets [31]. Given their vast attack surface and their reliance on wifi networks and radio waves, these devices are vulnerable to a whole range of cyberattacks.

Like computer networks, drone networks are exposed to potential cybersecurity risks that can jeopardize the security of this widely used technology, essential for carrying out vital missions. In our approach, we see drones as flying computers, and therefore we address their security issues in a similar way to traditional computer network security practices.

III. Proposed Approach

A. Architecture of the Proposed ZT-NIDS System

In this research project, ZT-NIDS (Zero Trust-based Network Intrusion Detection System), a new network intrusion detection system, is introduced. The proposed system is designed with a multi-agent architecture consisting of a set of independent and cooperative agents working together to efficiently detect intrusions in drone. Fig. 7 illustrates the overall architecture of the proposed ZT-NIDS system.

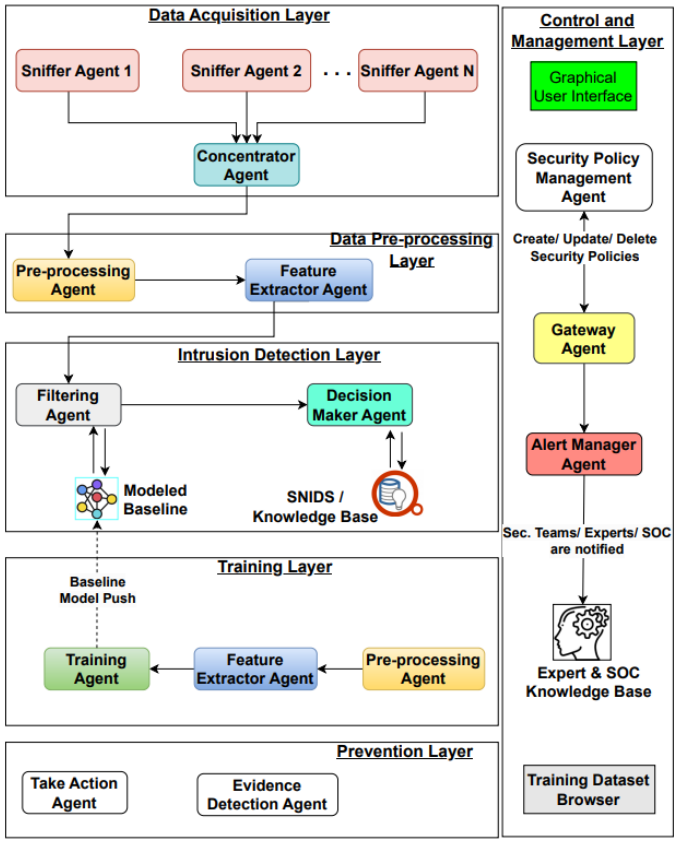


Fig. 7. The architecture of the proposed HNID&PS system.

B. Components of the Proposed Model

The ZT-NIDS system is designed according to a multi-agent architecture and consists of six layers, each with a set of agents.

1) Data Acquisition Layer – DAL: This layer is the input interface to the system, capturing and aggregating traffic from UAV fleet networks. It is made up of two types of agents:

- Sniffer agent: This agent captures drone network traffic in real time. Each sniffer can be positioned to cover a segment of the drone network.
- Concentrator agent: it aggregates all network events originating from all sniffers and timestamps all network events.

2) Data Pre-processing Layer – DPL: The DPL layer pre-processes captured network traffic to clean and process it. In this layer, we consider two types of agents:

- Pre-processing Agent: It cleans, normalizes, and correlates network traffic originating from the concentrator agent to eliminate missing and infinite values.
- Feature Extraction Agent: Its aim is to extract relevant features and reduce the dimensionality of network traffic, while retaining a maximum amount of information for recognizing network packets. To do this, it removes correlated, constant and quasi-constant attributes and passes the extracted features on to the adjacent layer.

3) Intrusion Detection Layer – IDL: The IDL layer receives the extracted attributes and compares them with the baseline and the signature database. It includes a set of two agents:

- Filtering Agent: It receives network packet attributes and checks the match values first against the baseline to recognize normal network traffic, then against the signature database to identify known attacks.
- Decision Making Agent: This agent is responsible for making decisions about the detected intrusions and their severity; it intervenes in accordance with the security policy governing the detection of abnormal activities on drone networks; it generates alerts and sends them to the Alert Manager Agent; it sends orders to the prevention layer, for example, to quarantine network equipment/systems contaminated by the intrusion.

4) Training Layer – TL: The TL layer is responsible for regularly training our system's detection mechanism. It includes three types of agents, the first two already presented: the Pre-processing Agent and the Feature Extraction Agent, plus a third, the Training Agent, described as follows:

- Training Agent: This agent is responsible for receiving pre-processed data sets as input; modeling the network baseline to recognize normal network behavior; generating and updating the modeled baseline module of the intrusion detection layer.

5) Prevention Layer – PL: The PL layer acts in the case of an intrusive attack through a set of two agents:

- Action agent: It is responsible for isolating and eliminating detected intrusions, e.g. by quarantining infected equipment; taking security measures, e.g. revoking access, blocking ports, suspending accounts, etc.
- Evidence Detection Agent: This agent is tasked with identifying the root cause of the security incident based on detailed information about the intrusion; it keeps a history of security incidents and all related documentation.

6) Control and Management Layer – CML: The CML layer enables IT security managers of UAV networks to define security policies and undertake configuration actions using three types of agents:

- Security Policy Management Agent: Managing and controlling the intrusion security policies; Updating machine learning models; Optimizing whitelists and blacklists with information on hacking sources; Managing threshold levels
- Gateway Agent: This agent is responsible for promoting communication and coordination between the various agents; managing data exchanges and communication protocols; coordinating response actions between agents in case of network intrusion; synchronizing agent activities for consistent system functioning.
- Alert Manager Agent: This agent is tasked with correlating alerts generated by the system, so that only relevant alarms are triggered; reducing the false positive rate; notifying security administrators in real time, so that they can anticipate and intervene in case of network intrusion.

C. ZT-NIDS Detection Mechanism / Zero-trust Principle

Since a drone is a sort of flying computer, the network configurations of drone fleets are comparable to those of modern computer networks. Consequently, the proposed intrusion detection system is well suited to overcoming the challenges inherent in cyber threats that could compromise the security of flying aircraft.

The diagram in Fig. 8 illustrates the deployment mode of the proposed system in a drone fleet network. The idea behind it is to monitor the flow of data between the various actors in a fleet of drones, then analyze it to identify known and unknown attacks. The system consists of a group of autonomous entities that work together to successfully accomplish the tasks involved in identifying intrusions into a drone network.

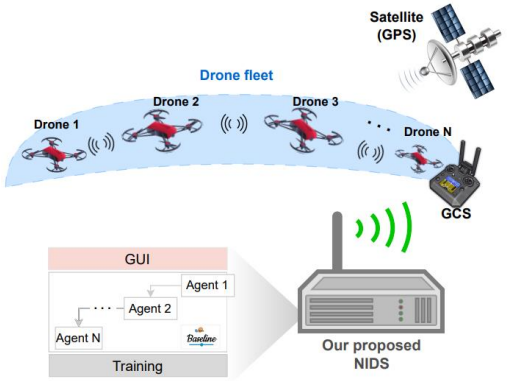


Fig. 8. Deployment mode of the proposed system.

A key element of the proposed system is its adherence to the "zero-trust" principle, which means that the detection mechanism works by considering a minimal level of trust in the processing of network traffic originating from drone fleets networks. To be more precise, the adopted detection mechanism is based on a network baseline as a reference, enabling us to identify any deviation from normal behavior. The diagram in Fig. 9 provides an overview of the system's detection mechanism. Its aim is to detect any attack attempts against drone networks, no matter whether they are already registered or not.

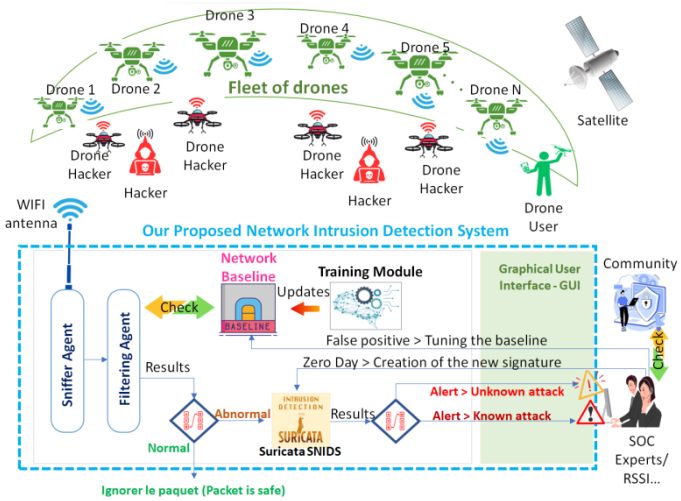


Fig. 9. Zero trust based intrusion detection mechanism.

For this purpose, the system is equipped with a baseline of normal behavior patterns, enabling it to distinguish legitimate transactions on the network. Any deviation from this basic pattern is interpreted as suspicious activity, triggering a notification to security managers. At the same time, the system has a standard signature database containing listed attack patterns. This signature database is continually updated as new attacks are identified through the process of comparing them with the network's reference baseline.

IV. Experimentation

A. Technical Architecture of the Lab used for Simulation

1) Lab environment: The entire laboratory is hosted by VMware on a workstation, whose characteristics are illustrated in Fig. 10.

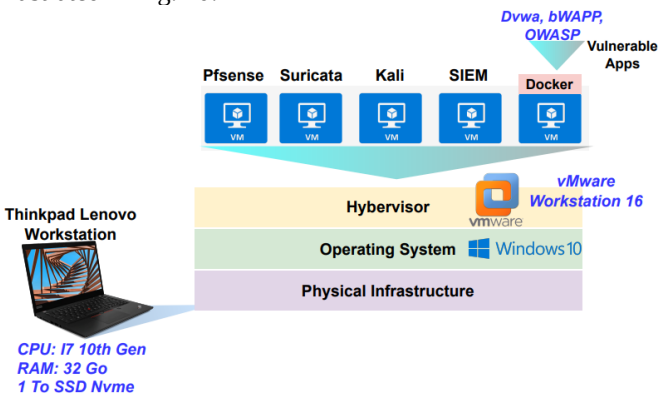


Fig. 10. Technical infrastructure hosting the ZT-NIDS experimental lab.

2) Applicative architecture of the Lab used for simulation: This laboratory aims to demonstrate the practical implementation of the ZT-NIDS system introduced in this research project. Fig. 11 gives a visual presentation of the laboratory's application architecture, highlighting the inputs and outputs of the platform designed to simulate the ZT-NIDS system.

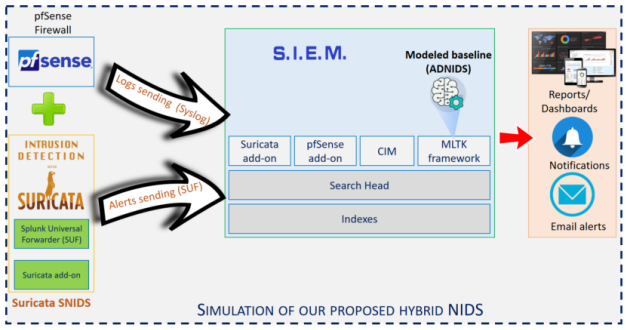


Fig. 11. Applicative architecture of the lab simulating the ZT-NIDS model.

The architecture diagram above shows two modules on the left-hand side: pfSense and Suricata. Both modules play a crucial role in intercepting, collecting and aggregating real-time events originating from drone networks. Suricata functions as a signature-based network intrusion detection system, and is either connected to a SPAN port on the switch, or to a network card operating in Promiscuous mode, enabling it to capture a copy of all ongoing network traffic in real time. The modules Splunk Universal Forwarder (SUF) and Suricata-Add-Ons have been deployed on the Suricata host to facilitate the extraction and recognition of the intercepted network events.

The network events collected are then directed to Splunk's SIEM (Security Information and Event Management) for processing, thanks to a number of additional modules. In particular, the Suricata-Addon, pfsense-Add-On and CIM modules integrated into the SIEM facilitate the normalization, indexing and storage of actionable network events, making it easier to extract attributes describing network packet behaviors.

In the SIEM, we have incorporated the MLTK framework to establish the baseline model for normal network operation. This enables the SIEM to recognize deviations from typical network behavior. With this configuration, Suricata excels at recognizing known attacks, while the Baseline module is specially designed to identify anomalies and previously unidentified zero-day attacks.

The SIEM is configurable to trigger alerts when real attacks are detected, and provides managers with comprehensive security indicators for in-depth analysis. In addition, it provides routine reports to administrators, giving them an overview of the security posture of the monitored networks.

B. Laboratory Components for Simulating the Proposed System

The practical simulation of our ZT-NIDS system relies on the set-up of a laboratory environment, as described in the previous section. This laboratory includes several components and software tools used to emulate the ZT-NIDS model proposed as part of our research. Table I below gives an overview of the components that make up our model and the simulation tools used.

C. Data Collection

Given that a drone is essentially a flying computer, and considering the similarity between traditional computer networks and those used by drones, the CICIDS2017 dataset

was deliberately chosen to evaluate the effectiveness of the system. The main focus was on the benign traffic portion of the data, with the aim of creating a representative baseline of normal network operation. CICIDS2017 was chosen for its consistency and relevance to modern network behavior, which differentiates it from conventional data sets that are often more theoretical than practical.

TABLE I. THE PROPOSED MODEL COMPONENTS VS. SIMULATION TOOLS

Our ZT-NIDS model components	*Simulation tools (Lab)*
Sniffer Agent	Pfsense
Concentrator Agent	Pfsense
Preprocessing Agent	Splunk Universal Forwarder
	Splunk TA for Suricata
	Splunk Common Information Model
	TA-Pfsense
Feature Extraction Agent	Machine Learning ToolKit (MLTK)/ Python/ Splunk Add-ons
Filtering Agent	Splunk Search Head
Decision Maker Agent	Splunk SIEM
Training Agent	Python & Machine Learning ToolKit
Take Action Agent	Ansible
Security Policy Management Agent	Splunk Administration Interface
Gateway Agent	Splunk SIEM
Evidence Detection Agent	Splunk Search App
Alert Manager Agent	Splunk SIEM
Signature-based NIDS	Suricata NIDS
Graphical User Interface	Splunk Dashboard
Benign traffic/ Baseline	CICIDS2017/ benign traffic

1) CICIDS2017 dataset used for evaluating the proposed system: The Canadian Cybersecurity Institute has released the CICIDS2017 dataset to help researchers tackle the challenges of intrusion detection [33]. The CICIDS2017 dataset is available in two distinct formats: CSV files for learning purposes, and PCAP files to enable rigorous evaluation of detection mechanisms proposed by the community.

In this study, we exploited the CICIDS2017 dataset in CSV format to train the ZT-NIDS system in recognizing normal behaviors using machine learning algorithms. This process has created the Baseline module integrated into the architecture of the ZT-NIDS system.

To test the system under real-life conditions, the PCAP format was adopted. Several packets were replayed using tools such as tcpdump, Wireshark and Snort. These tools enable the system to intercept generated network traffic and automatically identify records corresponding to possible signs of attack.

The different data structures used in this case study are shown in Fig. 12. On the one hand, CSV format data sets are used to feed the learning module, enabling the network baseline to be established. On the other hand, PCAP files are used to simulate real network traffic, in order to test the system in a practical, authentic environment.

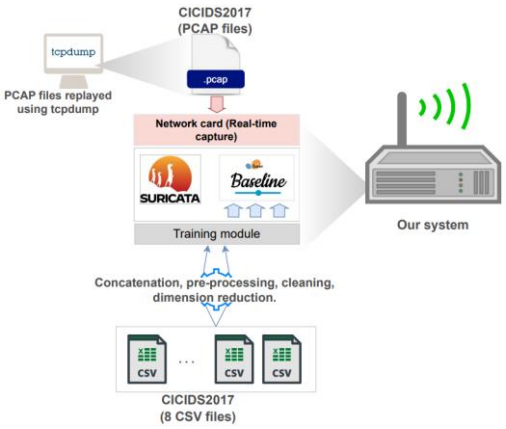


Fig. 12. Dataset used to train and to test the proposed system.

2) CICIDS2017 dataset preprocessing

a) Composition of the intial CICIDS2017 dataset: The CICIDS2017 dataset has been chosen to model the network baseline, as it is reliable, up-to-date and can represent the modern real network traffic [33]. However, it poses certain cleaning, scaling and conversion problems for the use by machine learning algorithms [34]. Accordingly, pre-processing operations need to be undertaken before using the benign class to model the network baseline.

The CICIDS2017 dataset is multi-class in nature, including a "Benign" category reflecting regular network traffic. Additional categories are also included, representing distinct types of known attacks, as shown in Fig. 13.

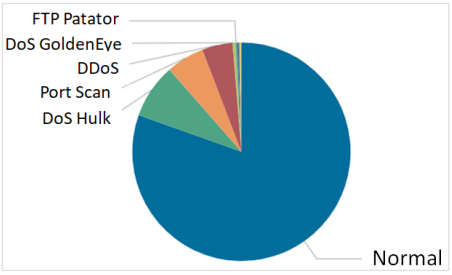


Fig. 13. Traffic classes contained in the intial CICIDS2017 dataset.

b) Cleaning up the CICIDS2017 dataset: The CICIDS2017 dataset suffers from problems of sanitisation, essentially due to the presence of infinite, null or sometimes missing records. Such problems can generally lead to falsified classification results during the machine learning process.

Some transformation processes can produce errors associated with undefined, infinite and oversized values. That's why we used certain Pandas methods like "drop()" to clean up the CICIDS2017 dataset of null, missing and infinite values.

c) Balancing of the CICIDS2017 dataset: The CICIDS2017 dataset is unbalanced regarding normal and abnormal records, as shown in Table II. Indeed, the class labeled "Normal" dominates over that labeled "Abnormal". The abnormal class refers to all the attack classes mentioned in Table I. As a result, over-fitting and under-fitting problems can be generated during the learning phase [35].

TABLE II. COMPOSITION OF THE DATASET BEFORE BALANCING

Class	Records Count	%
NORMAL	1818477	80 %
DoS Hulk	184858	8 %
Port Scan	127144	5.6 %
DDoS	102421	5.02 %
DoS GoldenEye	8234	0.36 %
FTP-Patator	6350	0.28 %
SSH-Patator	4718	0.2 %
DoS slowloris	4637	0.2 %
DoS Slowhttptest	4399	0.19 %
Bot	1573	0.07 %
Web Attack Brute Force	1206	0.05 %
Web Attack XSS	522	0.028 %
Infiltration	29	0.001 %
Web Attack Sql Injection	17	0.0007 %
Heartbleed	9	0.0003 %

In order to overcome the problems of overfitting and underfitting, the dataset has been balanced to ensure a balanced presence of the different classes. We used the SMOTE Python technique to increase the percentage of minority classes. On the other hand, the randomUnderSampler Python technique was used to decrease the number of normal records in order to avoid the performance problems that the laboratory environment might envisage. Accordingly, Table III highlights the dataset composition after its balancing using the two aforementioned python techniques.

TABLE III. COMPOSITION OF THE DATASET AFTER BALANCING

Class	Records Count	%
NORMAL	618000	50.2 %
DoS Hulk	240000	19.49 %
Port Scan	160000	12.99 %
DDoS	150000	12.18 %
DoS GoldenEye	12000	0.97 %
FTP-Patator	7000	0.56 %
DoS Slowhttptest	7000	0.56 %
DoS slowloris	7000	0.56 %
SSH-Patator	5000	0.4 %
Bot	5000	0.4 %
Web Attack Brute Force	5000	0.4 %
Web Attack XSS	5000	0.4 %
Infiltration	5000	0.4 %
Web Attack Sql Injection	5000	0.4 %
Heartbleed	5000	0.4 %

d) CICIDS2017 dataset binarization: The modeling of the network baseline is based on the class tagged "Normal/Benign" in the CICIDS2017 dataset. Therefore, the balanced dataset was transformed into another binary dataset including two main classes of network traffic: Normal and Abnormal.

Python libraries were used and the binary pre-processed dataset comprises 618000 normal traffic records and 618000 abnormal records. Consequently, both classes are present with an equal percentage of 50% each, as shown in Fig. 14.

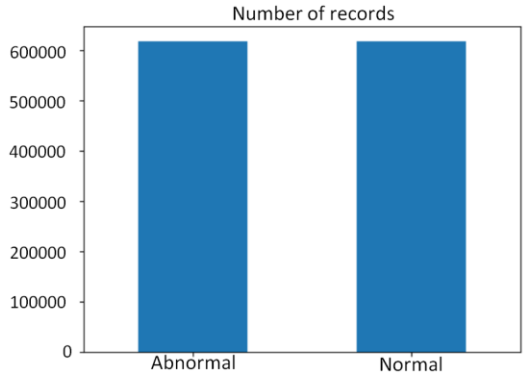


Fig. 14. Proportion distribution after balancing the binarized dataset.

e) Dimension reduction of the training dataset: The binarized balanced CICIDS2017 dataset contains a large number of attributes that are not necessarily relevant (79 attributes). This large number of attributes could cause enormous processing delays and could lead to falsified results when modeling the baseline. It makes more sense to eliminate unnecessary attributes and keep only the most relevant ones.

The StandardScaler imported from the sklearn.preprocessing library was used to downscale the features within the Pandas dataframe before applying the PCA transformation for dimension reduction. 30 principal components with zero cumulative variance were thus retained using the PCA technique, as shown in Fig. 15. This represents a considerable dimension reduction from 79 to 30 components that can describe 99.9% of the information within the standardized dataset.

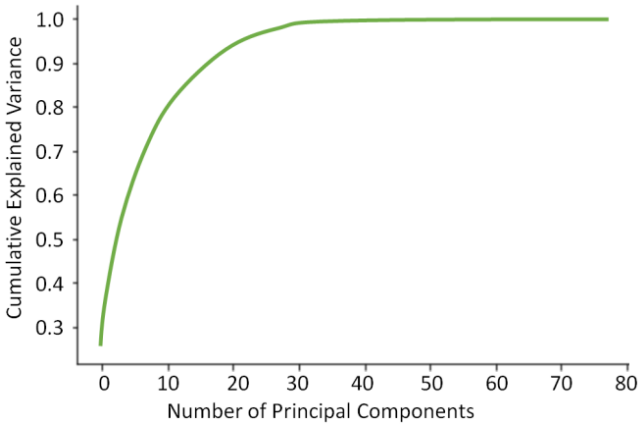


Fig. 15. PCA components used to represent the CICIDS2017 dataset.

D. Baseline Modelling and its Performance Evaluation

1) Network baseline modelling: After preprocessing the CICIDS2017 dataset, we used the "fit()" function in Splunk's MLTK framework to train the system on normal network traffic and generate the baseline model as output.

The MLTK framework is based on the Scikit Learn python library to train machine learning models [36]. This framework is very powerful, fast and reliable, as it copies the events to be processed into memory before launching the pre-processing and training operations.

The sequence of actions undertaken by the "fit()" function is illustrated in Fig. 16 below, in order to generate a model capable of recognizing the network baseline.

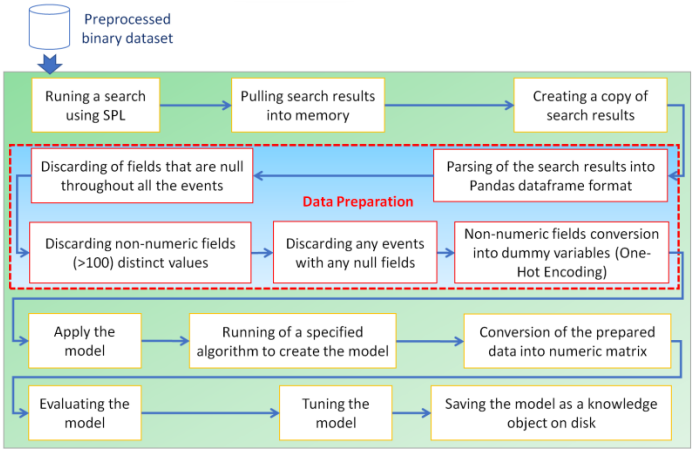


Fig. 16. Workflow of the training process using MLTK Framework applied on the optimized binary CICIDS2017 dataset.

2) Machine learning algorithms used to train the baseline

a) Machine learning algorithms and data sampling: After pre-processing and binarization of the CICIDS2017 dataset, a number of machine learning algorithms were used to build up a model capable of characterizing the network baseline. We used algorithms encompassing a variety of approaches, including Random Forest, Decision Tree, Naïve Bayes and Multi-Layer Perceptron. The key objective was to determine the most efficient and accurate machine learning algorithm that can differentiate between normal and abnormal drone's network traffic.

Having been pre-processed and binarized, the CICIDS2017 dataset was randomly separated into two sets: one for training (comprising 80% of the data) and the other for testing (comprising 20% of the data). In addition, PCAP files were used to evaluate system performance against real-time computer network data. These PCAP files were replayed using the tcpdump utility.

b) Performance evaluation metrics: The evaluation process incorporated key performance measures, such as Accuracy, Precision, Recall, F Score and Confusion Matrix. Importantly, cross-validation was applied with a k-fold value set at 10, guaranteeing a robust and reliable evaluation procedure.

Accuracy measures the number of correct predictions (both true positives and true negatives) that a model makes out of all predictions. It quantifies the overall accuracy of the model's predictions.

Precision measures the exactness of the positive predictions made by a model. It estimates the percentage of true positive predictions out of all positive predictions, and thus the model's ability to avoid false positives.

Recall, also known as sensitivity or true positive rate, measures the model's ability to identify all true positive instances. It refers to the proportion of true-positive predictions to all true-positive cases.

F1 score is the harmonic mean of precision and recall. It is a balanced measure that combines both precision and recall. It is particularly useful in unbalanced data sets where one class clearly outnumbers another.

The confusion matrix is a powerful tool for evaluating the performance of machine learning algorithms. According to [32] and as shown in Fig. 17, the confusion matrix is generally composed of four basic elements:

- True positives (TP): These are cases for which the model was able to correctly predict the positive class.
- True negatives (TN): These are cases for which the model was able to correctly predict the negative class.
- False positives (FP): These are Type I errors, in which the model has incorrectly predicted the positive class when it should have been negative.
- False negatives (FN): Also known as Type II errors, these correspond to cases where the model has incorrectly predicted the negative class when it should have been positive.

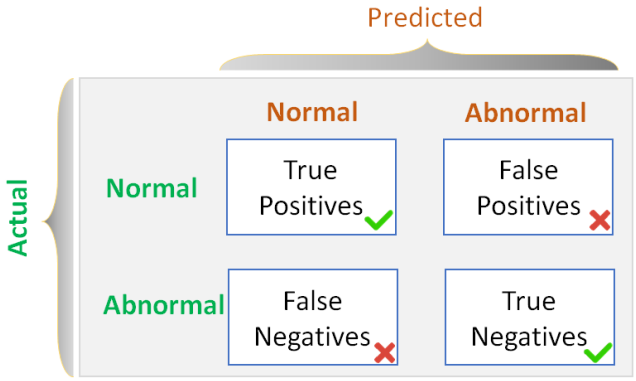


Fig. 17. Components of a confusion matrix (the normal class is our target).

3) Evaluation of the generated network baseline: To develop a highly efficient baseline network model, we tested a number of machine learning algorithms on the pre-processed and binarized CICIDS2017 dataset. In particular, we applied Random Forest, Decision Tree, Naïve Bayes and Multi-Layer Perceptron models.

a) Random Forest – RF: The Random Forest classifier was used to model the baseline of drone network traffic (regular operation of UAV networks). The test confusion matrix is shown in Table IV. The results obtained showed that the model generated was able to accurately recognize benign traffic with an accuracy, F1 score, precision and recall that all reached the exceptional level of 99.99%.

TABLE IV. RANDOM FOREST – TESTING CONFUSION MATRIX

	Predicted NORMAL	Predicted ABNORMAL
NORMAL	247194	6
ABNORMAL	10	247190

b) Decision Tree – DT: The decision tree was tested for modeling the network baseline (benign traffic modeling) and the confusion matrix for the tests carried out is presented in Table V. The generated model was capable of accurately recognizing benign traffic. The measures of accuracy, F1 score, precision and recall all reached the exceptional level of 99.99%.

TABLE V. DECISION TREE – TESTING CONFUSION MATRIX

	Predicted NORMAL	Predicted ABNORMAL
NORMAL	247192	8
ABNORMAL	12	247188

c) Naïve Bayes – NB: Table VI shows the confusion matrix results derived from the evaluation of the Naïve Bayes algorithm. This algorithm was employed to model the network baseline, with the pre-processed and binarized CICIDS2017 dataset. The performance measures obtained include 99.85% in accuracy, 99.84% in F1 score, 99.87% in precision and 99.83% in recall.

TABLE VI. NAÏVE BAYES – TESTING CONFUSION MATRIX

	Predicted NORMAL	Predicted ABNORMAL
NORMAL	246899	301
ABNORMAL	405	246795

d) Multi-Layer Perceptron – MLP: Table VII shows the testing confusion matrix following the use of the Multi-Layer Perceptron algorithm to model the network baseline, focusing on the classification of benign network traffic. The results reveal a precision rate of 99.93%, with F1 score and recall both reaching 99.92%. In addition, the precision measure achieved a high accuracy level of 99.94%.

TABLE VII. MULTI-LAYER PERCEPTRON – TESTING CONFUSION MATRIX

	Predicted NORMAL	Predicted ABNORMAL
NORMAL	247070	130
ABNORMAL	195	247005

4) Baseline modeling – Discussion: The classification results of benign traffic using a set of machine learning algorithms on the CICIDS2017 dataset are extremely promising. For the most part, the algorithms demonstrated an exceptional capacity to recognize normal network traffic with high performance and accuracy. As a result, the model generated can be integrated into the ZT-NIDS system to ensure the detection of deviations from regular network traffic. The positive results obtained are mainly attributable to the pre-processing, balancing, binarization and dimension reduction actions carried out on the CICIDS2017 dataset prior to its use.

In this laboratory, the Decision Tree algorithm was used to model the benign network traffic. This model was then integrated into the ZT-NIDS system, founded on the zero-trust principle, meaning that no network packet should be assumed to be trustworthy. The aim is to distinguish what is considered normal, and any deviation from this basis is automatically considered suspicious.

E. Signature-based Module

As mentioned earlier, the ZT-NIDS mechanism integrates both signature-based and anomaly-based modules enabling it to identify known and unknown cyber threats. Concerning the anomaly-based module, a baseline of benign network traffic has been created. The next focus is on the signature-based module, responsible for recognizing known intrusions.

We have already used Suricata as a signature-based module to detect known intrusions in one of our previous research works [34], and it has demonstrated excellent performance. We highlighted some of its advantages over Snort, such as its multithreading capability and its efficiency in handling known attacks. Consequently, for the ZT-NIDS mechanism, we used Suricata as a SNIDS (Signature-Based Network Intrusion Detection System) module to guarantee detection of known intrusions that deviate from the network baseline.

F. Preparing Attack Scenarios – Going Into Production

To evaluate the effectiveness of detection and simulate network attacks, we used tcpdump to replay pcap records from the CICIDS2017 dataset. Replaying a PCAP portion of CICIDS2017 that was never seen by the system allowed us to measure the performance and accuracy of the proposed approach. The illustration in Fig. 18 gives an overview of the lab architecture and attack scenarios.

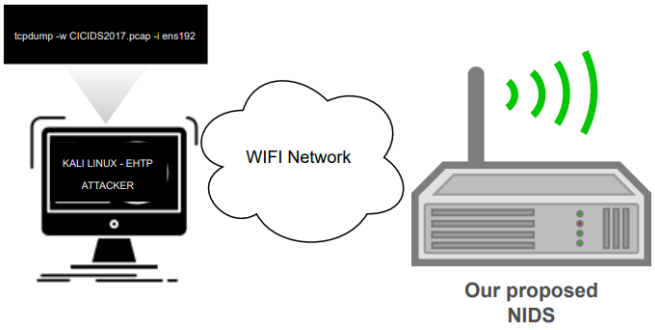


Fig. 18. An overview of the architecture used to reproduce the attack scenarios.

G. Testing the Baseline Model on New, Never-before-seen Data

This section outlines the methodology adopted to evaluate the real-time functioning of the proposed ZT-NIDS system with particular emphasis on its components. Fig. 19 shows the sequential steps involved in testing the system with new network events. Notably, the step "Transform search results using data preparation" reproduces the procedures detailed in the training phase diagram (indicated by a red strikethrough outline in Fig. 16).

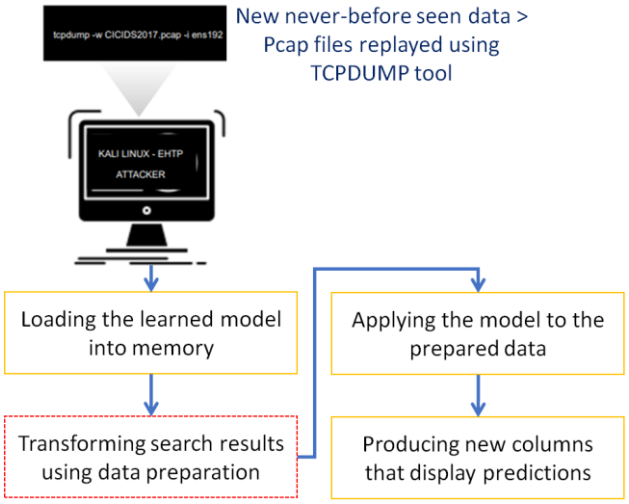


Fig. 19. Worfkow of testing process of ZT-NIDS on never-before-seen network events.

In order to evaluate the modeled baseline in a real network, we used an intact part of the CICIDS2017 dataset in PCAP format, which had not been used during the ZT-NIDS model training phase. The TCPDUMP tool, running under Kali Linux, was used to replay the CICIDS2017 pcap files. These network events were then transmitted to the SIEM system, where they were compared with the baseline and Suricata.

The SIEM then extracts the characteristics of the network events received and compares them to the baseline models, which had previously been stored in memory, and to Suricata's SNIDS. A new column is then generated to store the predicted values generated after application of the base model, as shown in Fig. 20. Further comparisons are made between the values predicted by the model and the actual values, enabling the calculation of performance metrics that serve as evidence of the model's effectiveness.

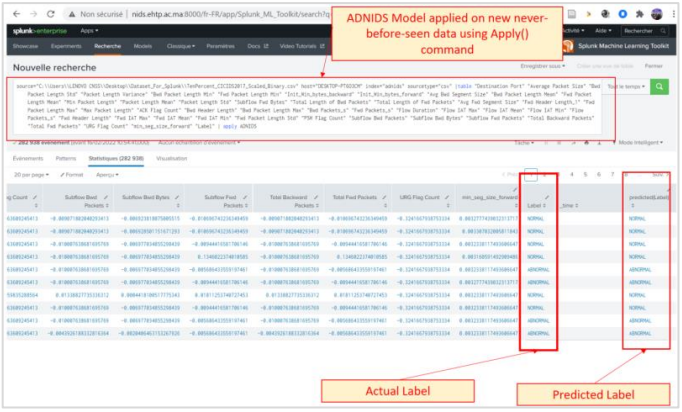


Fig. 20. The baseline model applied on never-before-seen network events.

V. Results and Discussion

A. Results

The following section summarizes all the results obtained after testing different machine learning algorithms on the pre-processed, balanced and binarized CICIDS2017 dataset. In particular, we focused on the recognition of benign traffic, i.e. the creation of the baseline. Consequently, Table VIII presents all the results obtained.

TABLE VIII. Summary of Results

Algorithms	Recall	Precision	Accuracy	F1
Random Forest	0,9999	0.9999	0.9999	0.9999
Decision Tree	0,9999	0.9999	0.9999	0.9999
Naïve Bayes	0,9983	0,9987	0,9985	0,9984
Multi-Layer Perceptron	0,9992	0,9994	0.9993	0,9992

B. Discussion of the Obtained Results

As we have previously noted, there has been limited research addressing the specific challenge of intrusion detection within drone fleet networks. Most existing studies have focused on aspects such as routing protocols, battery autonomy, and other functionalities, often overlooking security concerns. However, given the increasing importance of drone technologies and their modern applications, the security of these networks has become critically significant.

The analysis of drone networks has led us to the conclusion that drones can be regarded as flying computers, and their networks share similarities with modern computer networks. Building upon this insight, we have introduced a model named ZT-NDIS, designed with a multi-agent architecture to implement a detection mechanism in line with the zero-trust principle.

The multi-agent architecture has proven well-suited to the challenge of intrusion detection in drone networks due to its characteristics of distribution, autonomy, and cooperation among the various agents. The detection mechanism comprises two complementary modules: a signature-based detection module and an anomaly detection module.

To materialize the proposed model, we conducted a comprehensive simulation of all components of the ZT-NIDS architecture using a real-world laboratory setup. Initially, we preprocessed, balanced, and reduced the dimensionality of the CICIDS2017 dataset. We then assessed a range of machine learning algorithms to model the baseline of benign network traffic. Cross-validation techniques were employed to ensure classification performance. While all tested algorithms yielded promising results, we ultimately selected the "Decision Tree" model.

Subsequently, we constructed the baseline model using the MLTK framework based on Scikit Learn, implementing the Decision Tree algorithm. To complement the proposed detection mechanism, Suricata was integrated for recognizing attacks with known signatures.

Finally, a real-world testing was conducted by networking the system and replaying CICIDS2017 PCAP files. The results were highly satisfactory, and the system was able to effectively identify both normal and abnormal network.

This proposed system represents a significant contribution and marks the initial step toward enhancing the security of civilian drone networks, which continue to gain popularity in various applications.

VI. Conclusion and Perspectives

We introduced in this work a network intrusion detection system called ZT-NDIS, designed with a multi-agent architecture and a zero-trust detection mechanism adapted to drone networks. The multi-agent architecture is suitable to drone networks given their distributed, autonomous and cooperative characteristics. The zero-trust principle enabled us to detect various types of cyberattack, including zero-day threats, through continuous comparison of network traffic against the modelled baseline of benign traffic.

Thanks to the pre-processing of the CICIDS2017 dataset, we were able to create an effective network baseline model with almost 100% accuracy. The inclusion of Suricata has further enhanced the system's performance, particularly against known attacks.

To evaluate the system's effectiveness, we carried out real laboratory tests. This enabled us to assess its ability to detect emerging cyber threats that could target drone networks.

The proposed system represents an important step towards improving the security of drone networks and gaining a better understanding of their security posture. Future work will focus on developing the various agents, generating real network traffic from drone networks and testing the system in a production environment.

References

[1] Hashimy, S. Q., & Benjamin, M. S. (2023). The Deployment of US Drones in Afghanistan: Deadly Sky and Unmanned Injustice.

[2] Shaikh, E., Mohammad, N., & Muhammad, S. (2021, March). Model checking based unmanned aerial vehicle (UAV) security analysis. In 2020 International Conference on Communications, Signal Processing, and their Applications (ICCSPA) (pp. 1-6). IEEE.

[3] Mekdad, Y., Aris, A., Babun, L., El Fergougui, A., Conti, M., Lazzeretti, R., & Uluagac, A. S. (2023). A survey on security and privacy issues of UAVs. Computer Networks, 224, 109626.

[4] Suresh Kumar, K., Prabakaran, D., Senthil Kumaran, R., & Yamuna, I. (2022). Privacy and security of smart systems. Intelligent Green Technologies for Sustainable Smart Cities, 291-315.

[5] Wojciechowski, P., & Wojtowicz, K. (2023, June). Detection of Critical Infrastructure Elements Damage with Drones. In 2023 IEEE 10th International Workshop on Metrology for AeroSpace (MetroAeroSpace) (pp. 341-345). IEEE.

[6] Keskin, B. B., Griffin, E. C., Prell, J. O., Dilkina, B., Ferber, A., MacDonald, J., ... & Gore, M. L. (2022). Quantitative investigation of wildlife trafficking supply chains: A review. Omega, 102780.

[7] Al-Wathinani, A. M., Alhallaf, M. A., Borowska-Stefańska, M., Wiśniewski, S., Sultan, M. A. S., Samman, O. Y., ... & Goniewicz, K. (2023, May). Elevating Healthcare: Rapid Literature Review on Drone Applications for Streamlining Disaster Management and Prehospital Care in Saudi Arabia. In Healthcare (Vol. 11, No. 11, p. 1575). MDPI.

[8] Botta, A., Rotbei, S., Zinno, S., & Ventre, G. (2023). Cyber Security of Robots: a Comprehensive Survey. Intelligent Systems with Applications, 200237.

[9] Ouiazzane, S., Addou, M., & Barramou, F. (2022). A multiagent and machine learning based denial of service intrusion detection system for drone networks. Geospatial Intelligence: Applications and Future Trends, 51-65.

[10] Mohsan, S. A. H., Khan, M. A., Noor, F., Ullah, I., & Alsharif, M. H. (2022). Towards the unmanned aerial vehicles (UAVs): A comprehensive review. Drones, 6(6), 147.

[11] Silva, M., Reis, A., & Sargento, S. (2023). A Mission Planning Framework for Fleets of Connected UAVs. Journal of Intelligent & Robotic Systems, 108(1), 2

[12] Saffre, F., Hildmann, H., Karvonen, H., & Lind, T. (2022). Self-swarming for multi-robot systems deployed for situational awareness. In New Developments and Environmental Applications of Drones: Proceedings of FinDrones 2020 (pp. 51-72). Springer International Publishing.

[13] Jean-Aimé Maxa. Architecture de communication sécurisée d'une flotte de drones. Réseaux et télécommunications [cs.NI]. Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier), 2017. Français.

[14] Said OUIAZZANE, Fatimazahra BARRAMOU and Malika ADDOU, "Towards a Multi-Agent based Network Intrusion Detection System for a Fleet of Drones" International Journal of Advanced Computer Science and Applications(IJACSA), 11(10), 2020

[15] Eric W Frew and Timothy X Brown. Networking issues for small unmanned aircraft systems. Journal of Intelligent and Robotic Systems, 54(1-3) :21–37, 2009.

[16] Van, Y., Chen, X., Li, R., & Jiang, Y. (2022, August). The Intelligent Pelagic Communication System Architecture of the Fleet based on UAV Swarm Relay. In 2022 9th International Conference on Dependable Systems and Their Applications (DSA) (pp. 960-964). IEEE.

[17] Lakhwani, K., Singh, T., & Aruna, O. (2022). Multi-Layer UAV Ad Hoc Network Architecture, Protocol and Simulation. Artificial Intelligent Techniques for Wireless Communication and Networking, 193-209.

[18] Souli, N., Kolios, P., & Ellinas, G. (2023). Multi-Agent System for Rogue Drone Interception. IEEE Robotics and Automation Letters, 8(4), 2221-2228.

[19] Mahajan, N., Kaushal, S., & Kumar, H. (2023, March). IMS enabled Centralized UAV control system with seamless connectivity over mobile network. In 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA) (pp. 636-641). IEEE.

[20] Mademlis, I., Nousi, P., Lavaux, D., Aubourg, T., Le Barz, C., & Pitas, I. (2023). Secure Communications for Autonomous Multiple-UAV Media Production. In Unmanned Aerial Vehicles Applications: Challenges and Trends (pp. 323-347). Cham: Springer International Publishing.

[21] Omolara, A. E., Alawida, M., & Abiodun, O. I. (2023). Drone cybersecurity issues, solutions, trend insights and future perspectives: a survey. Neural Computing and Applications, 1-39.

[22] Lin, N., Yiyang, Y., Yingjie, Z., & Zihui, L. (2023, May). Network Threat Analysis and Protection Technology of UAV System. In 2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI) (pp. 236-240). IEEE.

[23] Krichen, M. (2022). Défis de sécurité pour les communications par drones: menaces, attaques et contre-mesures possibles.

[24] Bajracharya, R., Shrestha, R., Kim, S., & Jung, H. (2022). 6G NR-U based wireless infrastructure UAV: Standardization, opportunities, challenges and future scopes. IEEE Access, 10, 30536-30555.

[25] Gosain, M. S., Aggarwal, N., & Kumar, R. (2023, April). A Study of 5G and Edge Computing Integration with IoT-A Review. In 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES) (pp. 705-710). IEEE.

[26] Karmakar, G., Petty, M., Ahmed, H., Das, R., & Kamruzzaman, J. (2022, December). Security of Internet of Things Devices: Ethical Hacking a Drone and its Mitigation Strategies. In 2022 IEEE Asia-

Pacific Conference on Computer Science and Data Engineering (CSDE) (pp. 1-5). IEEE.

[27] R., Abdan, M., & Ghasemi, S. (2022). A self-adaptive intrusion detection system for securing UAV-to-UAV communications based on the human immune system in UAV networks. Journal of Grid Computing, 20(3), 22.

[28] Shrestha, R., Omidkar, A., Roudi, S. A., Abbas, R., & Kim, S. (2021). Machine-learning-enabled intrusion detection system for cellular connected UAV networks. Electronics, 10(13), 1549.

[29] Ihekoronye, V. U., Ajakwe, S. O., Kim, D. S., & Lee, J. M. (2022, November). Hierarchical intrusion detection system for secured military drone network: A perspicacious approach. In MILCOM 2022-2022 IEEE Military Communications Conference (MILCOM) (pp. 336-341). IEEE.

[30] Al-Haija, Q., & Al Badawi, A. (2022). High-performance intrusion detection system for networked UAVs via deep learning. Neural Computing and Applications, 34(13), 10885-10900.

[31] Kharchenko, V., & Torianyk, V. (2018, May). Cybersecurity of the internet of drones: Vulnerabilities analysis and imeca based assessment. In 2018 IEEE 9th international conference on dependable systems, services and technologies (DESSERT) (pp. 364-369). IEEE.

[32] Susmaga, R. (2004). Confusion matrix visualization. In Intelligent Information Processing and Web Mining: Proceedings of the International IIS: IIPWM ‘04 Conference held in Zakopane, Poland, May 17–20, 2004 (pp. 107-116). Berlin, Heidelberg: Springer Berlin Heidelberg.

[33] Stiawan, D., Idris, M. Y. B., Bamhdi, A. M., & Budiarto, R. (2020). CICIDS-2017 dataset feature analysis with information gain for anomaly detection. IEEE Access, 8, 132911-132921.

[34] Ouiazzane, S., Addou, M., & Barramou, F. (2022). A Suricata and Machine Learning Based Hybrid Network Intrusion Detection System. In Advances in Information, Communication and Cybersecurity: Proceedings of ICI2C’21 (pp. 474-485). Springer International Publishing.

[35] Ouiazzane, S., Addou, M., & Barramou, F. (2023). Cyberthreat Real-time Detection Based on an Intelligent Hybrid Network Intrusion Detection System. In Big Data Analytics and Intelligent Systems for Cyber Threat Intelligence (pp. 175-194). River Publishers.

[36] Mendoza, M. A., & Amistadi, H. R. (2018). Machine learning for anomaly detection on VM and host performance metrics.

Flood Prediction using Hydrologic and ML-based Modeling: A Systematic Review

A Fares Hamad Aljohani[1], Ahmad. B. Alkhodre[2], Adnan Ahamad Abi Sen[3],
Muhammad Sher Ramazan[4], Bandar Alzahrani[5], Muhammad Shoaib Siddiqui[6]
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, KSA[1, 3, 4, 5]
Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah KSA[2, 6]

***Abstract*—Flooding, caused by the overflow of water bodies beyond their natural boundaries, has severe environmental and socioeconomic consequences. To effectively predict and mitigate flood events, accurate and reliable flood modeling techniques are essential. This study provides a comprehensive review of the latest modeling techniques used in flood prediction, classifying them into two main categories: hydrologic models and machine learning models based on artificial intelligence. By objectively assessing the advantages and disadvantages of each model type, we aim to synthesize a systematic analysis of the various flood modeling approaches in the current literature. Additionally, we explore the potential of hybrid strategies that combine both modeling methods' best characteristics to develop more effective flood control measures. Our findings provide valuable insights for researchers and practitioners in the field of flood modeling, and our recommendations can contribute to the development of more efficient and accurate flood prediction systems.**

***Keywords*—Flood prediction; hydrologic model; machine learning; systematic review**

I. Introduction

Floods are one of the most destructive and widespread natural disasters, inflicting severe environmental and socio-economic impacts worldwide [1]. In recent years, the frequency and intensity of floods have increased, underscoring the urgent need for accurate flood modeling to guide effective disaster response and management. Devastating floods have resulted in massive property damage, loss of human and animal lives, destruction of crops, and the propagation of waterborne diseases [2].

Understanding the intricacies of floods and accurately predicting their occurrence and severity is critical in developing strategies for flood management and mitigating potential damages. In the field of flood modeling, two primary approaches have been widely utilized: hydrologic models and data-driven prediction models. Hydrologic models aim to simulate the complex physical processes and interactions within the hydrological system, relying on high-quality data and hydrological expertise. However, these models face challenges in accurately evaluating uncertainty propagation and predicting real-time flood depths [3]. Conversely, data-driven prediction models, leveraging machine learning techniques, have shown superior accuracy and broad applicability in flood forecasting [3]. Nevertheless, a promising avenue for advancement lies in the hybridization of both approaches, harnessing the strengths of each to overcome the limitations.

The primary objective of this study is to explore and compare different modeling approaches employed in flood modeling, specifically categorizing them into hydrologic models and data-driven machine-learning-based models. The research aims to address the existing knowledge gaps in flood modeling and shed light on the diverse methodologies used in simulating and understanding flood events.

To achieve this objective, the study presents an analysis of the strengths and limitations of both model types. Furthermore, it highlights the potential benefits of integrating these approaches to create more robust and accurate hybrid models. The study analyzes a comprehensive range of sources, including case studies, real-world data, and academic research, to provide a well-rounded evaluation of flood modeling techniques. The article introduces and compares one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D) hydrologic models, with a focus on their capabilities, limitations, and applications in flood hazard simulation and prediction. Additionally, the study delves into the advantages of hybrid models, demonstrating how their integration can contribute to more effective flood management strategies.

Contribution: This article's main contribution is to provide valuable insight into flood modeling, and advancement in the field, and ultimately aid in better understanding and managing flood events. By combining a thorough exploration of modeling approaches with a critical assessment of their performances, this study aims to lay the foundation for more resilient flood management practices in the face of escalating climate challenges. The article is structured as follows:

Section II provides the systematic approach adopted to gather related work and literature review along with bibliometric analysis. Sections III and IV provide introductions to various subdomains in Hydrologic modeling and Machine-learning-based modeling, respectively. Sections V and VI analyze the literature review on Hydrologic modeling and Machine-learning-based modeling, respectively. Section VII provides the analysis and discussion, while Section VIII provides the conclusions.

II. Bibliometric Analysis of Hydrologic Modeling and Machine Learning in Predicting Flood

The study focused on examining the trends and patterns in research publications in the field of flood prediction using machine learning-based modeling and hydrologic modeling. Two widely used databases, Web of Science (WoS) and Scopus, were utilized to gather relevant research papers. Additionally, the Dimension database was included to ensure comprehensive coverage of the literature. To visualize and analyze the data, VOSviewer software [4] was employed, along with traditional methods of data representation.

Using information from the Scopus and WoS databases, the bibliometric analysis produced the two instructive visualizations shown in Fig. 1 and Fig. 2. The size of the circles in Fig. 1 represents the number of citations each document in the hydrologic modeling and flood prediction domain has received. Greater circles indicate more citations, which reflects the importance and influence of the corresponding works on this topic. Similar to Fig. 1, but with a focus on machine learning modeling, Fig. 2 shows the citation distribution for the same domain.

Similarly, Fig. 2 and Fig. 3 show the number of citations in the context of machine learning and Hydrologic Modeling in flood prediction. Bibliometric analysis software VOSviewer was used to create both visualizations. A better knowledge of the current trends and contributions in hydrologic modeling and machine learning applied to flood prediction research is made possible by these numbers, which provide insightful information about the research environment by highlighting the most important and cited articles in both fields.

Table I lists how many articles have been written about using hydrologic modeling and machine learning techniques to anticipate flooding. The article counts or publication records for each type of study are shown in the table, indicating the volume of research done in the field of hydrologic modeling and machine learning for flood prediction across the various databases.

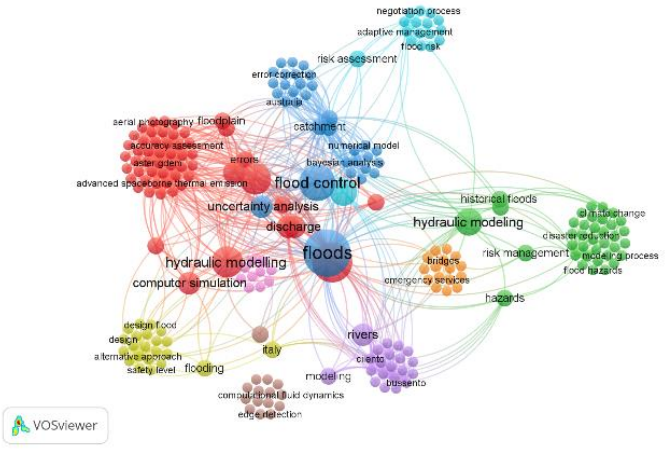


Fig. 1. Map the hydrologic modeling and "Flood prediction" network – Scopus.

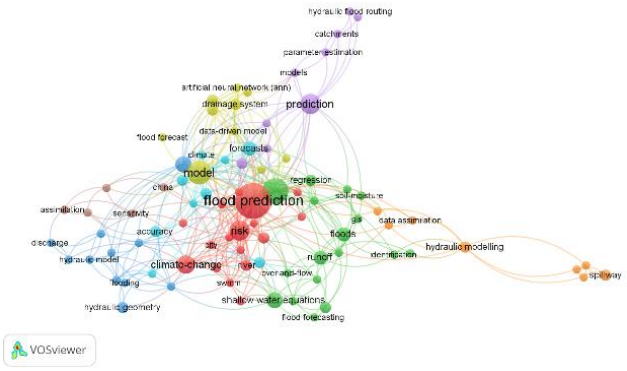


Fig. 2. Map the hydrologic modeling and "Flood prediction" network- WoS.

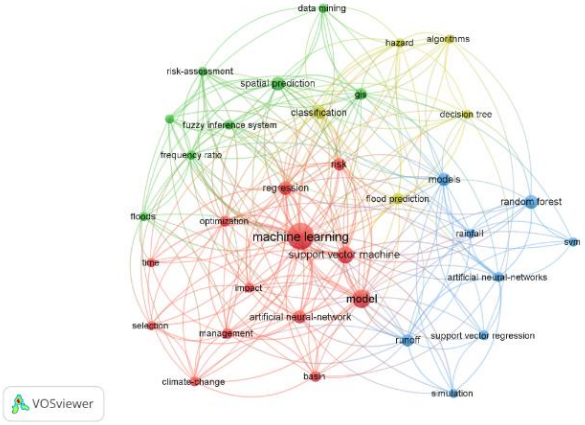


Fig. 3. Map machine learning and "Flood prediction" network WoS.

TABLE I. The Number of Articles Related to the Use of Hydrologic and Machine Learning in Flooding Prediction

Topic \ Database	*WoS*	*Scopus*	*Dimension*
Hydrologic	1493	1249	1255
Machine learning	3,027	2280	3,730

A. Bibliometric Analysis

The bibliometric analysis revealed an interesting trend in the distribution of research papers. It became evident that there is a significantly higher number of research papers dedicated to machine learning compared to those published in the field of hydrologic modeling (see Fig. 4, 5, 6, and 7). This indicates a greater emphasis on the utilization of machine learning techniques in flood forecasting research.

The higher number of research papers on machine learning suggests that it has gained significant attention and interest in the field of flood forecasting. Machine learning techniques offer the potential to improve the accuracy and efficiency of flood prediction models by leveraging large datasets and complex algorithms. On the other hand, the relatively lower number of research papers on hydrologic modeling may indicate a need for further exploration and development in this area. The findings of this study provide valuable insights for researchers and practitioners interested in the field of hydrologic modeling, machine learning, and flood forecasting.

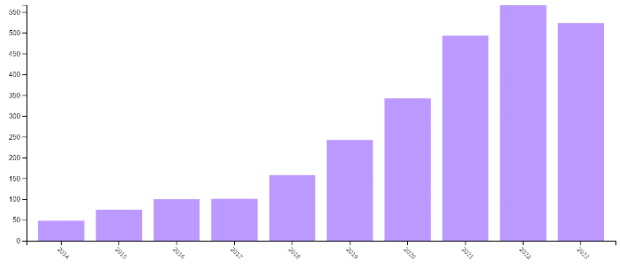


Fig. 4. Publications over time: Machine Learning and "Flood Prediction" WoS.

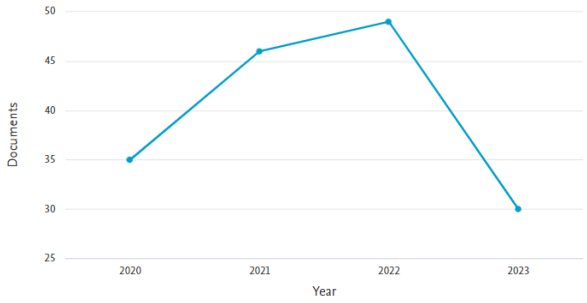


Fig. 5. Publications over time: Machine Learning and "Flood Prediction" Scopus.

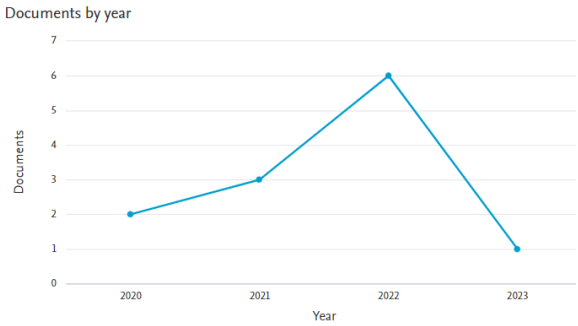


Fig. 6. Publications over time: Hydrologic Modeling and "Flood Prediction" Scopus.

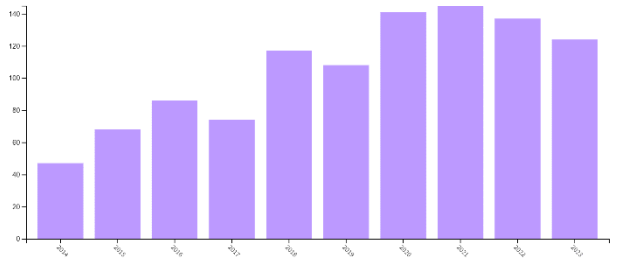


Fig. 7. Publications over time: Hydrologic modeling and "Flood prediction" WoS.

III. Hydrologic Modeling

Hydrologic modeling plays a crucial role in simulating and predicting flood hazards, enabling a deeper understanding of the complex dynamics associated with floods. This modeling approach encompasses various techniques, including one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D) models, each offering unique advantages and limitations. This study delves into these different modeling approaches, exploring their applications, capabilities, and challenges in simulating flood events. By examining the strengths and limitations of each modeling method, the study aims to enhance the understanding of hydrologic modeling and its significance in effectively managing and mitigating the impacts of floods.

Fig. 8 shows the hydrologic model configured for 1D, 2D, and coupled 1D/2D simulations (a, b, and c) [5]. The 1D simulation in panel (a) shows a condensed version of the hydrologic system that works well in settings with linear flow patterns. The 2D simulation, which considers two-dimensional flow characteristics, illustrates the hydrologic processes in more detail in Panel (b). The coupled 1D/2D simulation, which combines both methods in panel (c), enables a more precise and in-depth portrayal of complicated flow interactions in situations when both 1D and 2D models are required. This figure helps academics and practitioners choose the best strategy based on particular modeling requirements and objectives by providing a useful visual reference for understanding the various modeling setups and their applications in hydrologic simulations [5].

A. One-Dimension Hydrologic Modeling

Various authors have discussed the effectiveness of hydrologic modeling methods in simulating and predicting flood hazards. For example, a study conducted by Ambiental Environ-mental Assessment [6] demonstrates that the 1D model effectively captures the interconnected network of a river by linking multiple cross-sections that traverse both the land and the river. Through this model, the water level is simulated and allowed to flow in a single direction along the channel. Additionally, the model accommodates the possibility of reverse water flow, such as in cases where the presence of structures obstructs the passage of water. Pinos et al. [7] highlight that river flood events are among the most frequent and economically burdensome natural disasters. Despite floods being a natural component of the hydrological cycle, they have far-reaching environmental consequences and can cause significant human and financial losses. Consequently, the utilization of modeling techniques becomes essential in simulating and predicting these occurrences.

Pinos et al. [7] conducted a study on the performance of the hydrologic 1D model approach in approximating flood levels for a mountain river. The study utilized HEC-RAS, Mike 11, and Floor Modeler as modeling tools. In the case of HEC-RAS, high-resolution cross-section surveys were conducted at intervals of 25 meters along the river line. The validation of the model was based on historical flood regions with return periods ranging from 2 to 10 years. The findings of research [7] served as reference models for different return periods and were compared to other models. The 1D model is considered an acceptable approximation as long as the water remains within the roadway profile. However, when the flow in the streets exceeds the curbs, there is a potential for the flow direction to shift, making the 2D model more suitable at that point [8]. It is important to note that using the 1D hydrologic modeling

approach has certain limitations. One major limitation is the assumption that the floodplain between the various cross-sections of the river is similar. Additionally, there is a need to determine the specific number and spacing of cross-sections to accurately represent the river channel and neighboring topography, and the guidelines for establishing these parameters are limited [10].

B. Two-dimensional Hydrologic Modeling

2D flood modeling is an approach used to analyze and interpret the two-dimensional flow of water during anticipated flood events. It relies on digital terrain modeling and the bathymetry of water channels to establish the depth of water and depth-averaged velocity on a mesh or grid [11]. One of the advantages of this modeling technique is that it does not require predefined flow routes, allowing for a more flexible representation of the flow dynamics.

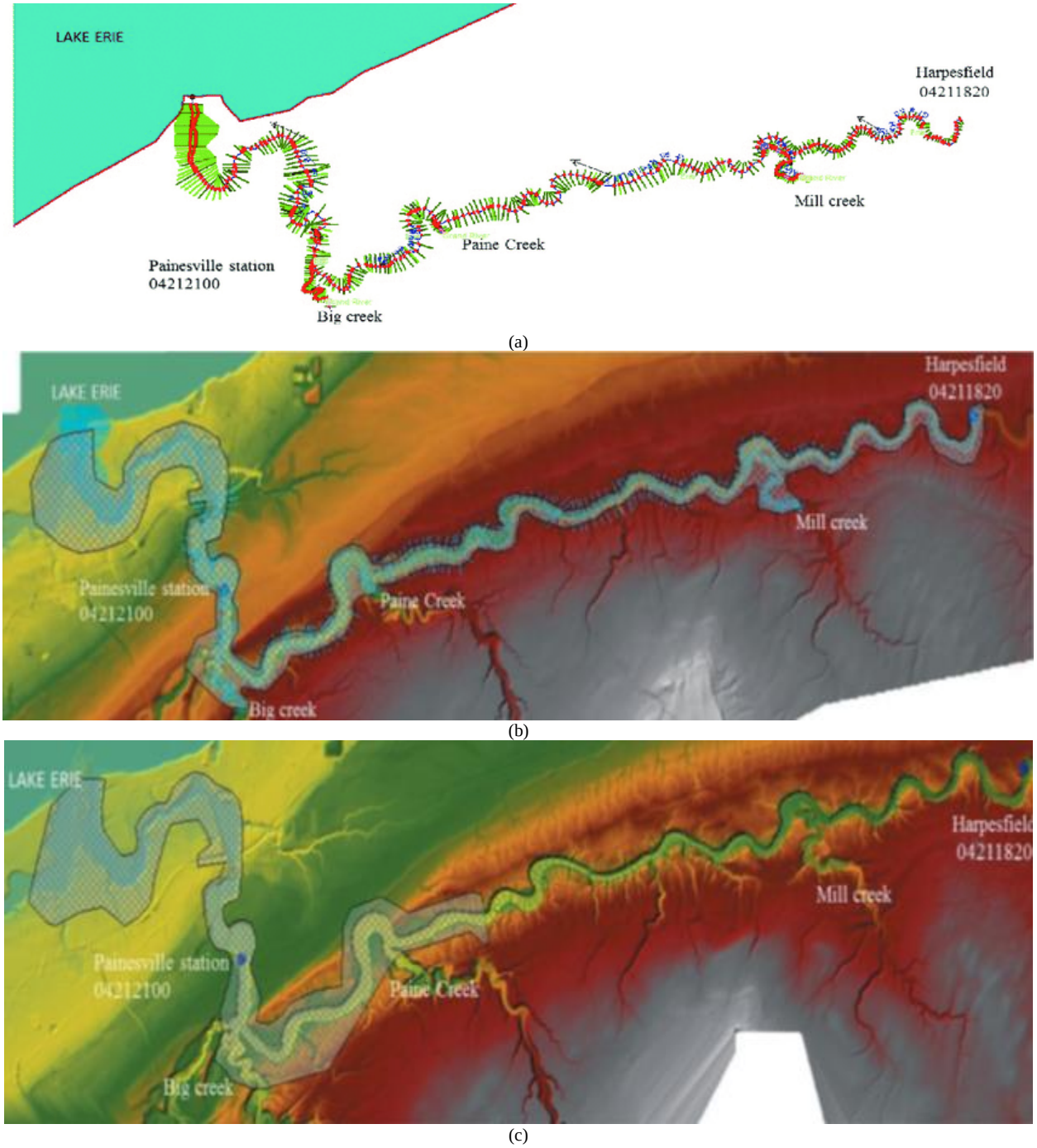


Fig. 8. Hydrologic model set up of (a) 1-D simulation, (b) 2-D simulation, and (c) coupled 1-D/2-D simulation [5].

An example of the application of 2D flood modeling can be seen in the assessment of the Medjerda River conducted by Gharbi et al. [12]. The researchers utilized 2D hydrologic modeling to understand the behavior of the river and accurately estimate the extent of flood zones. Their findings demonstrated that 2D analysis provides a more precise depiction of the flooded region compared to the one-dimensional (1D) unsteady flow study. The visual representation of flood extent through 2D model maps is superior to the traditional water-depth curves used in 1D models [9].

However, it is important to acknowledge the limitations of the 2D modeling approach. Firstly, the complexity of the 2D model makes it more computationally costly compared to 1D models. The computational time required for a complete run of a 2D model is significantly longer, often taking hours as opposed to minutes for a 1D model [8]. This can pose challenges for inexperienced hydrologists who may struggle with the complexity of the model and the efficient transfer of information to relevant departments.

Furthermore, the application of 2D modeling may be hindered by the availability of quality data. High-quality data, obtained through on-site sensors and Internet of Things (IoT) devices, is crucial for achieving accurate results [12]. Additionally, when analyzing flood risk within large cities and complex structures, relying solely on 2D geographical data may not be effective. In such cases, the integration of 3D city model-based GIS solutions becomes necessary to provide decision-makers with comprehensive information [13].

C. Three-dimensional Hydrologic Modeling

3D modeling involves the mathematical representation of surfaces in three dimensions, utilizing specialized software to manipulate vertices, edges, and polygons in a simulated 3D space [14]. When it comes to simulating and predicting flood hazards, 3D models have proven to be computationally expensive but crucial for accurately representing the three-dimensional flow around urban areas and interactions between flood waves and constructed buildings [15].

In the context of flood simulation and its impact on utilities, Adda et al. [16] emphasize that 3D modeling offers robust visual depictions that enable decision-makers to assess the safety of buildings located in flood zones. Their research explores the use of LiDAR (Light Detection and Ranging) data and 3D modeling to analyze flood risk on government utilities and buildings. LiDAR data is highlighted as an inexpensive and comprehensive method for multidimensional 3D mapping [16].

Data collection for 3D modeling often involves ground-controlled points obtained through GPS methods. Adda et al. found that their 3D approach revealed regions that were potentially situated on low-lying terrain prone to flooding. This information is vital for emergency decision-making and prioritization. The use of 3D hydrologic modeling enables rapid reaction, alert, and warning systems, mitigation strategies, and effective planning and management of complex geographic issues.

By utilizing 3D geospatial data, planning challenges can be better addressed, and conditions that may increase the risk of flooding can be identified. This aids in understanding and evaluating the nature of dangers and facilitates the development of clear management strategies for rescue operations. Additionally, one significant advantage of 3D hydrologic modeling is the ability to test infrastructure projects before implementing them on the ground, minimizing potential risks [17].

However, it is important to acknowledge the limitations of the 3D modeling approach. While it offers valuable insights, it is more suitable for localized hydrologic issues and may require substantial details, resulting in longer simulation times [17], [18].

D. Discussions

The evaluation and discussion of 1D, 2D, and 3D modeling methodologies for flood simulation and prediction emphasize the advantages and disadvantages of each methodology as well as some of the field's practical uses.

A common technique for simulating flow in a single direction along a channel is 1D hydrologic modeling. As long as the water stays inside the roadway profile, it offers a simplified portrayal of flood threats and is supposed to be acceptable. The primary benefit of 1D modeling is that it is straightforward and computationally efficient, making it appropriate for use in large-scale applications. A drawback of this method is that it necessitates particular rules for setting cross-section spacing and assumes uniform floodplain features between cross-sections.

On the other hand, 2D hydrologic modeling, which considers the two-dimensional flow of water and incorporates digital terrain modeling and bathymetry, provides a more in-depth simulation of flood events. It enables the depiction of flood inundation patterns and offers a more precise estimation of flood extent. The use of 2D models is particularly valuable when the flow exceeds the curbs and shifts direction, such as in urban areas with complex topography and structures. However, 2D modeling is computationally more expensive and requires high-quality data, including LiDAR data, to achieve accurate results [19].

It may also pose challenges for inexperienced users and information transfer to relevant departments. Additionally, it might make it difficult for new users to transmit information to the appropriate departments.

The study also emphasizes the value of 3D modeling in assessing flood risks, particularly when it comes to examining three-dimensional flow patterns around built-up areas and their interactions. 3D models enable precise simulation of flood impacts on utilities and infrastructure and offer a more thorough understanding of flood wave behavior. They give decision-makers useful information that enables them to evaluate the security of structures in flood-prone locations and prioritize emergency responses. By offering visualization tools, assessing hazards, and testing infrastructure projects before execution, 3D modeling also helps with planning and management. 3D modeling, on the other hand, necessitates specialist tools and comprehensive data and is more difficult and computationally expensive. It may take longer to simulate problems and is best suited for limited hydrologic problems.

IV. Machine Learning Algorithms in Flood Prediction

The destructive nature of floods has necessitated the advancement of flood prediction as a basis for risk reduction, policy suggestions, and minimizing property damage and loss of life. These requirements have led to the development of machine learning algorithms that mimic the mathematical expressions utilized in examining the physical processes of floods over the past two decades. In their study, Mosavi et al. [20] acknowledged that machine learning methods have improved prediction accuracy and provided cost-effective solutions, which has contributed to their increased popularity among hydrologists. Fig. 9 shows the fundamental steps in creating an ML-based model.

A systematic review revealed that several critical factors guide the selection of machine learning methods, including robustness, speed, computation cost, and generalization capability. Lawal et al. [21] discovered that the popularity of machine learning in predicting flood alerts and reducing the impact of floods arises from its low computational requirements, as it relies on observational data. However, a comparative study conducted by Lawal et al. [21], which evaluated logistic regression, support vector classification, and decision tree algorithms, highlighted the importance of considering performance accuracy, recall, and receiver operating characteristics when choosing machine learning algorithms for flood prediction.

Machine learning has made it possible to monitor the changing patterns of river water levels, which helps mitigate the socioeconomic implications caused by floods. According to Mosavi et al. [20], popular machine-learning methods for flood prediction include ANNs, SVM, SVR, ANFIS, WNN, and DTs. However, hybridization through various methods is also common. The study also found that the data decomposition technique is preferred to improve dataset quality and prediction accuracy, while ensemble methods facilitate generalization and reduce prediction uncertainty.

In addition, Mosavi et al. [20] identified that applying add-on optimizer algorithms improves prediction quality by tuning ANNs to optimal neuronal architectures. In a study on the detection of flooding from river water levels, Zehra [22] observed that non-linear (NARX) and support vector machines (SVM) are viable machine learning methods. The study revealed that NARX and SVM utilize hydrological resource variables such as precipitation amount, seasonal flow, peak gust, and river inflow, which are regressed into flood and non-flood classes.

Accurate prediction of floods and other hydrological events is crucial for water resource management techniques, policy development, and evacuation models. Improving prediction systems for short- and long-term flood events is significant in minimizing damage. It is important to note that machine learning (ML) approaches for flood prediction can vary significantly depending on the specific application, dataset, and type of prediction required.

For instance, ML approaches for predicting short-term water levels may differ greatly from those used for predicting long-term stream flows. When building an ML model, all the available data undergoes training, validation, verification, and testing processes [23]. These steps ensure that the model is trained effectively and performs accurately when applied to new data.

Overall, the development and utilization of ML models in flood prediction play a crucial role in improving the accuracy and effectiveness of water resource management, policy ideas, and analyses, as well as evacuation planning.

When accurate data are available, machine learning approaches can be a powerful tool in risk analysis. However, the findings obtained from these approaches may not be as sophisticated or predictable as those from model-driven studies, such as hydrodynamic models [24]. The use of a data-driven approach with machine learning for predictive studies can be relatively straightforward, particularly in the presence of uncertainty related to climate change.

One of the key advantages of using machine learning prediction models is their ability to capture flood nonlinearity based solely on historical data, without requiring an understanding of the underlying physical processes. Data-driven prediction models based on machine learning hold promise as they are easier to construct and require fewer inputs. Over the past two decades, the continuous improvement of machine learning algorithms has demonstrated their usefulness in flood forecasting, often surpassing conventional approaches in terms of performance and accuracy [25]. The distinguishing factor of machine learning technology in flood prediction is its ability to extract crucial information solely from input data without the need for specialized knowledge [26].

It is important to consider certain aspects of machine learning algorithms. Firstly, their performance is only as good as the quality of their training, which involves the system learning the intended task from previous data. Therefore, ensuring robust data enrichment is crucial in machine learning algorithms. Secondly, the competence of a machine learning algorithm varies depending on the specific task, which is commonly known as the "generalization problem." It refers to how effectively a trained system can forecast situations for which it was not specifically trained [20].

Fig. 9. The fundamental steps for creating a machine learning (ML) model.

Wagenaar et al. [27] discovered that the field of flood risk analysis, focusing primarily on rare extreme events, often faces challenges in data collection during such events, resulting in a lack of data for machine learning applications in flood risk and impact modeling, particularly for effective model training. However, advancements have been made in utilizing machine learning for descriptive hazard assessment using data from social media. Machine learning algorithms are effective optimization approaches rather than black box models, offering efficiency, reliability, and quick convergence at low computational costs.

The following are some commonly used ML algorithms in flood prediction:

A. ANNs: Artificial Neural Networks

The most often used algorithms for modeling flood prediction are ANNs. ANNs interpret historical data rather than the physical qualities of a catchment. As a result, ANNs are regarded as trustworthy data-driven tools for building sophisticated and nonlinear black-box models of the links between rainfall and flood. Despite their benefits, ANNs have several disadvantages, including network architecture, data management, and the inability to physically perceive the modeled system. The comparatively low precision while employing ANN is a significant disadvantage [20].

B. MLP: Multilayer Perceptron

The MLP is a class of FFNN that trains its network of interconnected nodes with multiple layers using supervised learning from BP. The MLP is characterized by simplicity, nonlinear activation, and a large number of layers. These qualities led to the model's widespread application in complicated hydrogeological models and flood prediction. MLP models were shown to be more effective and more generalizable in a review of ANN classes used in flood simulation. However, it is typically discovered that the MLP is more challenging to optimize [28].

C. SVM: Support Vector Machine

SVM, a supervised learning machine that operates on the statistical learning theory and the structural risk minimization (SRM) rule, is very well-liked in flood modeling. To reduce the predicted error and overfitting concerns of a learning machine, the SRM principle operates as a tradeoff between the quality and multidimensional character of the approximation function [29]. The SVM's training method creates new, non-probabilistic binary linear classifiers that maximize the geometric margin through inverse problem-solving and minimize the empirical classification error. Hydrologists use SVM extensively for flood prediction [20]. SVM, which is based on the structural risk reduction concept, is a reliable and efficient method for equation fitting, data analysis, hydrological forecasting, and other applications. Furthermore, SVM is used to handle small sample, non-linear, and high-dimensional pattern recognition problems and has unique benefits. SVM may be applied to classification as well as regression issues [30].

SVM applications are widely used in hydrological modeling and flood predictions. SVM's enhanced form as a regression tool supports vector regression (SVR) is a developed and efficient alternative procedure for dealing with regression difficulties during the last two decades by giving alternative loss functions. SVR is based on mapping and solving the original data into a high-dimensional feature space using linear and/or nonlinear regression classification. SVR formulation is based on SRM rather than ERM, which minimizes an upper bound of the generalization error rather than the prediction error on the training set [29].

SVM and other data-driven ML models rely on the quality and amount of training data as well as model optimization parameters. If the data is insufficient and inadequate to cover the differences, their learning falls short and, as a result, they cannot achieve reasonable accuracy. The disadvantages of SVM-type ML models for dealing with the "generalization problem" might be mitigated by a strong and complete understanding of ML techniques, as well as user-specified practical solutions [29].

SVM is essentially a linear machine and can be thought of as a statistical tool that solves issues using an approach akin to Artificial Neural Networks (ANN). Its approximate use of the Structural Risk Minimization (SRM) concept aids in its ability to generalize effectively to new data. While it has all the advantages of ANN, it also addresses some of the fundamental flaws that were observed in the ANN application. [31]. ANNs employ empirical risk minimization, but SVMs use structural risk minimization to handle the overfitting problem by balancing the model's complexity against its success in fitting the training data. [35].

The reason why the SVM algorithm is more popular in flood prediction than other algorithms is SVM may automatically choose the critical vectors in the training process as support vectors and delete the nonsupport vectors from the model. As a result, the model performs effectively in noisy environments. Furthermore, with certain crucial real training vectors encoded in the models as support vectors, SVM can trace back historical occurrences to enhance future forecasts with lessons learned from the past. Because the input vectors of SVM are fairly versatile, it is quite simple to integrate other relevant elements into the model (such as temperature, evaporation, date, etc.). Because SVM parameter optimization is a convex issue, there is only one optimal point, unlike ANN which has more than one optimal [33].

D. DT: Decision Tree

Because DTs are rapid algorithms, ensemble models to simulate and predict floods have become increasingly popular. The classification and regression tree (CART) is a common DT type used in machine learning. The decision tree is very useful for determining the level of risk of flooding [34].

E. GA: Genetic Algorithm

A genetic algorithm had been created by Holland. The survival of the fittest is the foundation of the idea. It uses chromosomes, which have several genes on each one. Every gene represents a choice variable (or model parameter), and every chromosome represents a potential best-case scenario [35].

F. ACO: Ant Colony Optimization

Dorigo developed the ACO after becoming curious about how ants choose the quickest route between their colony and a food source. It was discovered that although ants cannot see, they can communicate with one another via a chemical called a pheromone. Each spreads a scent along its course. Ants are therefore likely to select the route with the highest concentration of pheromone. According to this, ants will finally take the shortest way if there are both lengthy and short routes leading from the nest to the food source [35].

V. Analysis of Hydrologic Modeling Techniques

When it comes to hydrologic modeling, deciding between a steady-state and a non-steady-state flow is significantly easier. Although 2D modeling can yield better results in some cases, there are also scenarios where 1D modeling can produce outcomes equally as good as or better than 2D models, with less work and computing resources. Many situations are complex, and it is possible to include both the positive and bad aspects of each approach depending on the context.

A few instances of scenarios in which it is believed that 2D modeling is preferable to 1D modeling are as follows:

Water may flow in several directions if a levee is broken or overtopped in a model region located behind a levee system. Before it reaches the lowest point and begins to pool and maybe overtop or breach the levee on its lower end, water can flow overland in the protected region in several different ways thanks to the slopes present there. When the protected area is relatively small, and the entire area eventually fills to the level of a pool, a 1D model may accurately predict the ultimate water surface and the extent to which the region will be inundated.

Tides, river flows, and other water sources entering an estuary or bay can cause the water to flow in various directions—a place or occurrence where the water's flow path is not entirely clear. It is challenging to forecast flood occurrences due to the episodic character of flow evolutions on alluvial fans. This is because the channels' whole direction may shift while the event takes place, making it impossible to generate accurate predictions.

It is best to avoid making sharp bends when there is a good chance that considerable super elevation may occur. Because flood plains are expansive and level, the water that leaves the overbank zone may go in various directions. Measurements of precise velocities are required to correctly analyze the hydrology of flow around an item.

Because of the complexity of urban terrain, flows on the urban surface are often substantially different from flows in channels. In recent years, several examples have been explored and applied to coupled 1D/2D techniques, in which the urban surface is represented using two-dimensional (2D) flow approaches and combined with a 1D pipe network model. Roads, buildings, barriers, and other elements of metropolitan surfaces abound [30]. These structures, particularly buildings, will alter the direction and velocity of the flood water, resulting in a variety of complicated flow pathways. The information on the buildings may be distorted or lost if the grid resolution is too coarse. Models with finer grid resolutions may offer more precision and a more accurate depiction of physical processes.

Constructing an entirely 3D model is more complex, but once it has been constructed, changes to the design may be made methodically and straightforwardly. Applying design changes in a 2D model is more challenging than in 3D.

There are three types of instruments for predicting hydrological variables: conceptual, physically based, and "black-box" models. The underlying physics of the first two categories, which may be represented by either simplified relations or partial differential equations in one or two dimensions, must be understood. Furthermore, using these models to forecast rainfall/runoff processes and/or river routing also calls for a significant amount of topographic, land-use, and other information that might not be accessible. Additionally, the lengthy calculation requirements associated with this method, particularly when two-dimensional models are required, sometimes limit real-time forecasting derived from physically based models [36].

VI. Analysis of Machine Learning Techniques for Flood Prediction

This study on hydrologic modeling and machine learning in flood hazard assessment gives a thorough examination of machine learning techniques for flood prediction in this section. This analysis's goal is to assess the usefulness and applicability of several machine-learning strategies for anticipating flood dangers.

A. Evaluation Standards for Prediction

Establishing evaluation standards that cover accuracy, dependability, robustness, consistency, generalization, and timeliness is essential for creating accurate flood prediction models. These standards act as the basic rules for evaluating the efficacy of flood prediction models and guaranteeing their dependability in practical implementations.

B. Metrics for Performance Evaluation

Root-mean-square error (RMSE), mean error (ME), mean squared error (MSE), Nash coefficients (E), and correlation coefficient (CC or R^2) are some of the performance evaluation metrics for flood prediction models that are frequently utilized. These measures allow for a quantitative evaluation of the model's prediction skills and make it easier to compare various strategies.

C. Analysis of Various ML Algorithms in Flood Prediction

The study identifies the advantages of ANNs, such as enabling working with huge datasets, and the benefits and drawbacks of particular ANNs, such as Backpropagation Neural Networks (BPNN), functional networks, and the NARX network. The study also investigates how the inclusion of autoregressive models can improve the precision of flood forecasts.

The performance of the MLP and various DT models, such as the ADT model, the Rotation Forest (RF), and the M5 model tree (MT), is specifically examined. The study draws attention to their strength and effectiveness, particularly in cases with lengthy lead times.

SVM's excellent generalization capacity, promising hourly flood prediction results, and uncertainty evaluation for potentially dangerous flood quantiles are identified.

D. Hybrid Designs

The study investigates the effectiveness of hybrid models like the Adaptive Neuro-Fuzzy Inference System (ANFIS) and Wavelet Neural Network (WNN) for longer-term flood predictions that last longer than two hours. ANFIS's excellent capacity to predict flash floods in real-time, as well as its high accuracy and dependability is also identified. The study also looks at the advantages of sophisticated ANFIS hybrid models calibrated by Support Vector Regression (SVR) for nonlinear and real-time flood prediction, highlighting their enhanced prediction accuracy and cost-effectiveness.

E. Ensemble Methods

Finally, the study looks into cutting-edge hybrid models and ensemble techniques that combine statistical, soft computing, and machine learning techniques to improve flood prediction models. It examines Ensemble Prediction System (EPS) techniques, such as ANN, MLP, SVM, and RF ensembles, which show promise in enhancing prediction precision, and robustness, and lowering model uncertainty.

Table II lists the machine-learning techniques that have been applied to flood modeling. Although data-driven technologies, artificial neural networks (ANNs) have limits when it comes to understanding systems. The Multilayer Perceptron (MLP) is straightforward but difficult to optimize. Small sample sizes are effectively handled by a Support Vector Machine (SVM), but it requires high-quality data. Although Decision Tree (DT) is rapid and appropriate for flood modeling, more study is required. The Genetic Algorithm (GA) seeks the best options, but it depends on accurate encoding. For hydrogeological modeling and flood prediction, these methods have advantages and factors to consider.

An accurate forecast should be judged by its accuracy, dependability; robustness; consistency; generalization; and timeliness [37]. Durable and simple models are the best way to ensure that projects are completed on schedule. Using various root-mean-square errors (RMSE) can also evaluate the accuracy of forecasting models, mean error (ME), mean squared error (MES), and R^2 correlation coefficients, as well as the mean and squared errors for each model tested (CC). *RMSE* and R^2 values close to one indicate that flood forecasting models are generally reliable (Calculated using Eq. (1) and Eq. (2) respectively).

The flood forecasting models' reliability can be determined by examining their *RMSE* and R^2 values, where values close to one indicate higher reliability (calculated using Eq. (1) and Eq. (2) respectively). By referencing [20] and reviewing approximately 45 references, the study extracted and presented the results in Fig. 10 and Fig. 11. Evaluations of this study considered various factors, including the dataset, processing cost, and specific application. The study also assessed the method's generalizability, speed, installation cost, ease of use, and maintenance expenses. Standard deviation (RMSE) was measured using a single unit for accurate representation, and thorough confirmation ensured the absence of errors. R^2 and *RMSE* were utilized to assess the performance of single and hybrid ML approaches for short-term flood forecasting, as depicted in Fig. 10 and Fig. 11, based on Mousavi's research [20].

$$R^2 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{(\sum_{i=1}^{N}(x_i-\bar{x})^2)}*(\sum_{i=1}^{N}(y_i-\bar{y})^2)} \quad (1)$$

$$RMSE = \sqrt{\frac{(\sum_{i=1}^{N}(x_{obs,i}-x_{model,i})^2)}{n}} \quad (2)$$

where, the values are defined as follows:

x_i, y_i	The values that were observed and forecasted, along with the residue correspond to the i[th] data point.
$\bar{x}$ & $\bar{y}$	The arithmetic means of those values
X_{obs}	Observed value
X_{model}	The forecasted values for the specific year i

An ANN is a short-term forecasting technology that is widely viewed as promising. Improved methods for greater effectiveness Although ANNs performed severely in some early research, especially in the generalization component, they showed improved results when dealing with massive datasets. In this case, BPNNs and functional networks should be avoided. The models can handle noisy datasets accurately, efficiently, and quickly. In contrast, the NARX network outperformed the BPNN network. Even so, incorporating autoregressive models could improve accuracy.

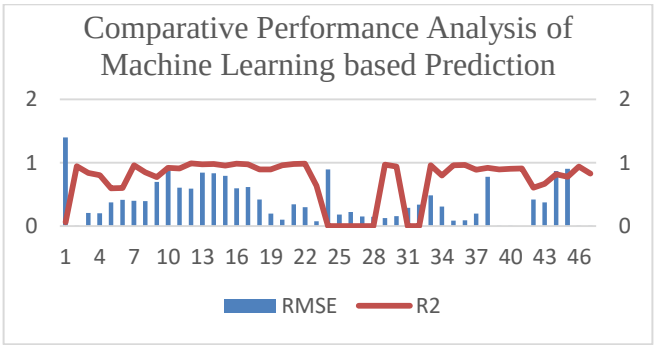


Fig. 10. Comparison of machine learning technologies in flood prediction based on R^2 and *RMSE* [20].

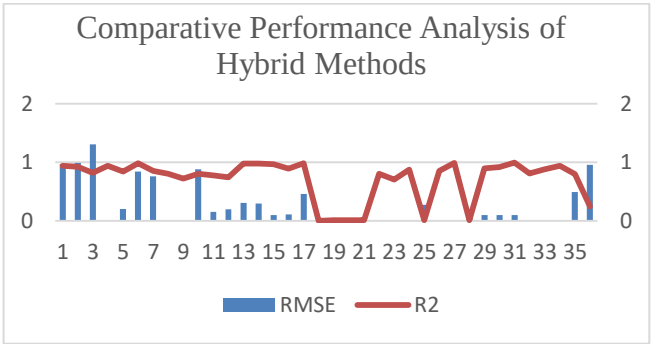


Fig. 11. R^2 and *RMSE* comparison of ML technologies in flood prediction [20].

TABLE II. A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR FLOOD MODELING

Algorithm	*Description*	*Advantages*	*Disadvantages*	*References*
ANNs (Artificial Neural Networks)	Interpret historical data, build sophisticated and nonlinear models	Trustworthy data-driven tools	Network architecture limitations, inability to perceive the modeled system	[30]
MLP (Multilayer Perceptron)	FFNN with multiple layers, widely used in hydrogeological models	Simplicity, nonlinear activation, effectiveness	Challenging to optimize	[39]
SVM (Support Vector Machine)	Supervised learning machine based on structural risk minimization	Reliable and efficient, handles small sample and nonlinear problems	Dependence on training data quality, challenges in generalization	[29], [30], [31][32], [43] [8]
DT (Decision Tree)	The rapid algorithm, increasingly popular in ensemble models	Speed, suitability for flood modeling	Further research is needed for flood prediction	[30]
GA (Genetic Algorithm)	The survival of the fittest concept uses chromosomes and genes	Ability to search for optimal solutions	Dependent on suitable chromosome encoding	[47]

TABLE III. A SYSTEMATIC REVIEW OF KEY RESEARCH PAPERS FOCUSED ON FLOOD MODELING AND PREDICTION

Location	*Methodology*	*Results*	*Key Factors*	*Implications*	*Reference*
Kalvan watershed, Iran	Tested five ML algorithms (ERT, RRF, PRF, RF, BRT) with 15 climatic and geo-environmental variables	ERT yielded the highest AUC (0.82)	Topographical and hydrological parameters	Aid in flood mitigation planning	[43]
Various urban settings	Reviewed prevailing flood modeling approaches	Overview of methods for pluvial flood modeling	-	Guide urban flood managers in selecting appropriate methods	[44]
Quannan area, China	Compared NBTree, ADTree, and RF methods using 13 flood explanatory factors	RF demonstrated high accuracy for flood susceptibility assessment	Multiple environmental factors	Support flood prediction in the study area	[45]
Khiyav-Chai watershed, Iran	Employed MDA, CART, SVM, and ensemble modeling with various factors	MDA had the highest predictive accuracy (89%)	Slope, drainage density, distance from river	Identify flood-prone areas for prevention	[46]
Various urban areas	Summarized calculation methods for urban flood numerical simulation	Identified trends for improving model accuracy and computational efficiency	1D-2D coupling, finite volume method, unstructured meshing	Guide hydrologists in model selection	[47]
Damansara River catchment, Malaysia	Combined FR approach with SVM using 13 flood conditioning parameters	Effective for flood risk management along an expressway	Environmental parameters	Replicable in other areas for flood risk assessment	[48]
Various locations in Thailand	Used MIKE-11 hydrologic model and ML techniques to improve runoff forecasting	Enhanced flood prediction accuracy	Hydrological data	Real-time flood prediction for better management	[49]
Jakarta, Indonesia	Utilized environmental factors and satellite imagery	Supported flood susceptibility prediction for flood risk management	Environmental conditions	Effective flood risk assessment	[50]
Indian monsoon forecasting	Employed neural networks for ISMR prediction	Demonstrated superior accuracy over existing models	Indian monsoon data	Improved ISMR forecasting	[51]
Various locations	Utilized various ML models for flood prediction	Compared models to improve prediction accuracy	Environmental and hydrological data	Enhanced flood prediction for management	[52]

Overall, the single prediction models examined could produce reasonably accurate short-term forecasts. However, hybrid models such as ANFIS and WNN scored better for forecasts lasting more than two hours. Non-linear and actual flood predictions were made more accessible and accurate with SVR-tuned advanced ANFIS hybrid models.

The capacity of (ML) models to capture the intricate (potentially unknown) nonlinear interactions between predictor (input) and predict and (output) variables sets them apart from other hydrologic modeling techniques, which solely base their predictions on previously observed data. Another benefit of these flexible models is their relatively high computational efficiency, which has increased their appeal over the past 20 years because of the continuing advancements in computing power [20].

Table III provides a comprehensive overview of various studies that have employed these approaches, encompassing a diverse range of locations and methodologies. The studies demonstrate the effectiveness of ML algorithms in improving flood prediction accuracy, emphasizing the importance of

environmental and hydrological factors in flood susceptibility assessment. Furthermore, the table highlights the development of various modeling techniques for urban flood modeling and the potential of integrated approaches combining ML and traditional methods for enhanced flood risk management. These findings underscore the transformative role of ML and other advanced techniques in addressing flood-related challenges, paving the way for more effective flood preparedness and mitigation strategies.

VII. Analytical Discussion

This systematic review offers valuable insights into the state of the art in flood modeling and prediction, providing a foundation for informed decision-making in flood risk management and urban planning. It underscores the potential of machine learning techniques in enhancing the ability to predict and mitigate the impacts of flash floods and pluvial flooding in urban areas. The study also highlights the strengths and weaknesses of both hydrologic models and data-driven machine learning models, paving the way for potential hybrid modeling approaches that can provide more accurate and efficient flood management strategies.

Additionally, the review identifies trends in research focus and geographic areas, which can guide future research directions and flood management efforts. Overall, this comprehensive analysis contributes to the ongoing efforts to develop effective flood control measures and adapt to the increasing challenges posed by climate change and urbanization.

Regarding computed water surface elevations and flow/stage hydrographs, 1D modeling can be just as accurate as 2D and 3D modeling while requiring less computational time and effort. The following are some examples of where this might be possible:

Rivers and floodplains where the predominant flow and force directions and paths follow the overall river flow. For most river systems, this is believed to be true, despite debates about the influence of lateral and vertical velocity on predicted water surface heights as well as the flood inundation boundary.

Gravity-driven streams with sloping beds tend to have very little overbank area. According to the river's dams, levees, pumping stations, and bridges, the river's predicted level and flow are affected by these gated projects.

The hydrologic flow characteristics present in many of the river systems are something that no 2D model has been able to represent adequately. It is a case where 1D models are much ahead of 2D models in terms of technological sophistication. These characteristics can be implemented in 2D models, but a popular 2D model with such a comprehensive collection of features is yet to be developed.

Medium to big rivers are considered when modeling a significant portion of the river system (100 or more miles) (i.e., 2-week to 6-month forecasts). Even with multi-processor computing and GPU (Graphics Processing Units) computation, 2D models have substantial geographical and simulation time restrictions regarding real-time forecasting. This is going to change over time. There is no evidence to justify using a 2D model in these situations. Many of the benefits of a 2D model will be thwarted by inaccurate topographical data due to a lack of information in the overbank and channel bathymetry.

As a result, the correctness of a 1D or 2D modeling approach for a given application is frequently in question. It is not as simple as choosing whether to solve the Saint Venant equations in one or two dimensions. There are other variables to consider. It was concluded that there are knowledge and tool gaps when determining whether to employ 1D, 2D, or 3D. It is necessary to use 1D and 2D models in the modeling efforts, and Hydrologic modeling software must be improved in this area.

Le et al. [38] [39] proposed many ANNs for seasonal flood forecasting and compared the outcomes. Data from 1970–1985 was employed as a training tool, while the 1986–1987 dataset was used to verify the results. The ANNs were able to identify whether the dataset was incomplete accurately. According to [20], employing ANNs to speed up data analysis could lower analytical expenses. ANNs have also been used to create precipitation forecast models, as seen in the [40]. An ANN model was used for a historical dataset spanning the years 1900–2001 to evaluate prediction accuracy. For this dataset, more than 100 floodstream localities were examined. The ANN, on the other hand, had issues with generalization. Despite this, water management found the ANN to be helpful in this instance.

Prediction models for heavy rain and flooding were developed by [37] using a variety of BPNNs. This dataset covered 1871–2010, and it did so every month. It was discovered that BPNN models using virtual networks were ideal for nonlinear flood forecasting since they were both fast and resilient. The following source may be long-term flood projections: BPNN and LLR-based models were used by Shamim et al. [41] to explain nonlinear floods better. An estimated two decades' worth of rainfall, evaporation, and water level statistics stretches back to 1988 in this dataset. According to their findings, LLR outperformed the BFGSNN neural network model in terms of efficiency and durability. In contrast to the other approaches, BPNN performed well.

According to [20], the most reliable ANN for long-term flood prediction is the BPNN model. For the long-term forecasting of flood discharge, ANNs performed better than BPNNs and MLPs in reference. There were promising outcomes while employing MLP. There was, however, the problem of generality. Mulualem and Liou [40] used an SVM model to forecast streamflow and reservoir inflow long-term. For comparison, they used neural networks and ARMA. Monthly river-flow flows from 1974–1998 were used to train the models, and data from 1999–2003 was used to evaluate them. According to a comparison of model performance, SVM outperformed the ANN when predicting long-term discharges.

ANNs are the most extensively utilized ML tool for dealing with complex flood features and incomplete data sets because of their accuracy, high fault tolerance, and parallel solid processing. ANN, on the other hand, has a problem with generalization. In [30], ANFIS, MLP, and SVM outperformed ANNs. As suitable data pre-processing, wavelet transformations may increase the performance of most machine

learning (ML) methods, which have been suggested. WANs, as opposed to traditional ANNs, have few benefits.

According to Hosseini et al. [42], short-term and long-term rainfall-runoff models' accuracy, precision, and performance were all improved by deconstructing ML algorithms (such as WNN). However, while WNNs have proven a success, long-term forecasts are limited. Hybrid WNN/autoregressive models WMRA and WARM were developed to improve the precision of one-year-ahead forecasts.

Through deconstruction, models performed far better in some circumstances, resulting in more accurate results. For example, wavelet–neuro-fuzzy models outperformed standalone ANFIS and ANNs in terms of accuracy and speed [42]. However, as the lead time lengthened, so did the degree of uncertainty in the predictions. Future research should take into account the accuracy of the model. An essential part of developing hybrid techniques was using data decomposition methodologies such as autoregressive, wavelet transformations (DWT), wavelet–autoregressive, IIS, and EMD.

Extrapolative prediction systems are another advancement in prediction accuracy and generalizability (EPS). Recent ensemble techniques have significantly changed speed, accuracy, and generality. Many non-traditional approaches to machine learning were used in developing ANN and WNN model training algorithms, such as BB sampling and genetic programming, in addition to typical ML techniques like the basic average and Bayesian inference. Conversely, ensembles outperformed models that did not include human decision-making as an input component. New decomposition–ensemble prediction models appropriate for monthly forecasts were the most significant hybrid models. Their accuracy and generalization improved significantly compared to SVM, ANFIS, and ANNs.

Predicting floods using machine learning models is still a developing field. An overview of machine learning models used in flood forecasting is presented in this study, along with the development of a classification strategy to examine the literature. More than 6,000 items were analyzed and investigated in the survey. Several original and significant studies compared the precision of at least two machine learning models. Models were classified into two groups depending on the lead time, with hybrid and single-method subgroups further subdividing.

Considering performance comparisons from previous literature helped in accessing and analyzing how these approaches perform. All approaches were examined using R^2 and RMSE, as well as a generalization, robustness, and computing costs/speeds. Despite the previous optimistic results, there was a lot of research and testing to enhance and develop the most popular machine learning algorithms like ANNs and SVM, SVR and ANFIS, WNN, and DTs. Four essential topics emerged from the research on improving prediction models' accuracy and general models.

The initial stage was to use both conventional and soft computing in conjunction with at least two different types of machine learning algorithms. Secondly, data segmentation methods were used to increase the dataset's quality, resulting in much higher accuracy in the predictions made from the data. Generalizability and predictive power were significantly improved and decreased by employing several approaches. Add-on optimizer algorithms, for example, can be used to increase the quality of neural network models.

Flood prediction is projected to improve significantly in the near and long term due to the development of these four leading technologies. Developing these new machine learning approaches relies heavily on applying soft computing principles in algorithm design. As a result, future hybrid machine-learning methods will rely heavily on soft computing techniques, as detailed in the study.

VIII. Conclusion

In conclusion, the study has underscored the critical impact of floods on the environment, emphasizing the need for effective prediction models to mitigate their adverse effects, such as loss of life, crop destruction, and increased waterborne ailments. The investigation has focused on the application of one-dimensional, two-dimensional, and three-dimensional hydrologic modeling for flood hazard forecasting, providing valuable insights for the development of preventive interventions. Notably, the study has revealed the growing superiority of machine learning (ML) techniques over traditional hydrologic models in predicting flood occurrences.

The findings indicate that ML, armed with sophisticated algorithms and extensive datasets, excels in its ability to provide accurate flood predictions. Notably, the study suggests that ML models can offer effective solutions by correctly estimating complex hydrological parameters, as exemplified in the accurate determination of water flow through structures like the hydrologic manifold P-10. This superiority of ML over hydrologic modeling systems is further supported by the bibliometric investigation, which revealed a burgeoning interest in utilizing machine learning for flood prediction.

Two key recommendations emerge from this study to further advance flood prediction accuracy. First, there is a discernible shift toward the use of machine learning techniques, and it is recommended to collect real-world data on flood events from Jeddah municipality. This data would serve as a valuable resource for training machine learning algorithms, enhancing their accuracy and applicability in the specific context of the study. Second, acknowledging the need for a comprehensive approach, the study recommends developing a hybrid model that combines time-series data for machine learning-based algorithms with the expertise of 1D/2D/3D hydrologic modeling. This hybrid approach aims to leverage the strengths of both methodologies for improved flood prediction, acknowledging the importance of evolving hydrologic modeling science while embracing the advancements offered by machine learning.

References

[1] Akshya, J., and P. L. K. Priyadarsini. "A hybrid machine learning approach for classifying aerial images of flood-hit areas." In 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), pp. 1-5. IEEE, 2019.

[2] Memon, F. S. (2015). Catastrophic effects of floods on environment and health: Evidence from Pakistan. Memon, FS and Sharjeel, MY, 72-84. http://dx.doi.org/10.22555/pjets.v5i2.903

[3] Hosseiny, H., Nazari, F., Smith, V., & Nataraj, C. (2020). A framework for modeling flood depth using a hybrid of hydrologics and machine learning. Scientific Reports, 10(1), 1-14.

[4] Van Eck, Nees, and Ludo Waltman. "Software survey: VOSviewer, a computer program for bibliometric mapping." scientometrics 84, no. 2 (2010): 523-538.

[5] Ghimire, Ekaraj, Suresh Sharma, and Niraj Lamichhane. "Evaluation of one-dimensional and two-dimensional HEC-RAS models to predict flood travel time and inundation area for the flood warning system." ISH Journal of Hydraulic Engineering 28, no. 1 (2022): 110-126.

[6] Ambiental Environmental Assessment. (2020). Detailed Hydrologic (Flood) Modelling. https://www.ambiental.co.uk/services/detailed-hydrologic-flood-modelling/

[7] Pinos, J., Timbe, L., & Timbe, E. (2019). Evaluation of 1D hydrologic models for the simulation of mountain fluvial floods: A case study of the Santa Bárbara River in Ecuador. Water Practice and Technology, 14(2), 341-354. https://doi.org/10.2166/wpt.2019.018.

[8] Leandro, J., Chen, A. S., Djordjevic, S., & Savic, D. A. (2009). Comparison of 1D/1D and 1D/2D coupled (sewer/surface) hydrologic models for urban flood simulation. Journal of hydrologic engineering, 135(6), 495-504.

[9] Leandro, J., Djordjević, S., Chen, A. S., & Savić, D. A. (2009). Flood inundation maps using an advanced visualization algorithm for 1D/1D models

[10] Ali, A. B. M. (2018). Flood inundation modeling and hazard mapping under uncertainty in the Sungai Johor basin, Malaysia. CRC Press.

[11] Liu, Q., Qin, Y., Zhang, Y., & Li, Z. (2015). A coupled 1D–the 2D hydrodynamic model for flood simulation in flood detention basin. Natural Hazards, 75(2), 1303-1325. DOI: 10.1007/s11069-014-1373-3.

[12] Gharbi, M., Soualmia, A., Dartus, D., & Masbernat, L. (2016). Comparison of 1D and 2D hydrologic models for floods simulation on the Medjerda River in Tunisia. J. Mater. Environ. Sci, 7(8), 3017-3026. https://www.jmaterenvironsci.com/Document/vol7/vol7_N8/315-JMES-1772-Gharbi.pdf

[13] Al Kalbani, K., & Rahman, A. A. (2021). 3D city model for monitoring flash flood risks in Salalah, Oman. International Journal of Engineering and Geosciences, 7(1), 17-23.

[14] Falter, D., Dung, N. V., Vorogushyn, S., Schröter, K., Hundecha, Y., Kreibich, H., ... & Merz, B. (2016). Continuous, large - scale simulation model for flood risk assessments: Proof - of - concept. Journal of Flood Risk Management, 9(1), 3-21. https://doi.org/10.1111/jfr3.12105

[15] Stamataki, I. (2020). Experimental and numerical investigation of flash floods and their interaction with urban settlements (Doctoral dissertation, University of Bath).

[16] Adda, P., Mioc, D., Anton, F., McGillivray, E., Morton, A., Fraser, D., & Eb, C. (2010). 3D flood-risk models of government infrastructure. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. ISPRS Arch, 38, 6-11. https://www.isprs.org/proceedings/XXXVIII/4-W13/ID_39.pdf.

[17] Sufiyan, I., & Zakariya, R. B. (2018). Monitoring Simulation for Flood Risk Prediction Using 3D and Swat in Terengganu Watershed. J Pollut Eff Cont, 6(216), 2

[18] Hosseini, F. S., Choubin, B., Mosavi, A., Nabipour, N., Shamshirband, S., Darabi, H., & Haghighi, A. T. (2020). Flash-flood hazard assessment using ensembles and Bayesian-based machine learning models: application of the simulated annealing feature selection method. Science of the total environment, 711. https://doi.org/10.1016/j.scitotenv.2019.135161

[19] Seibert, P., Raßloff, A., Kalina, K. A., Gussone, J., Bugelnig, K., Diehl, M., & Kästner, M. (2023). Two-stage 2D-to-3D reconstruction of realistic microstructures: Implementation and numerical validation by effective properties. Computer Methods in Applied Mechanics and Engineering, 412, 116098.

[20] Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: A literature review. Water, 10(11), 1536.

[21] Lawal, Z., Yassin, H., & Zakari, R. (2021). Flood prediction using machine learning models: A case study of Kebbi State Nigeria. 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). https://doi.org/10.1109/csde53843.2021.9718497

[22] Zehra, N. (2020). Prediction analysis of floods using machine learning algorithms (NARX & SVM). International Journal of Sciences: Basic and Applied Research (IJSBAR), 49(2), 24-34. Retrieved from https://www.gssrr.org/index.php/JournalOfBasicAndApplied/article/view/10719

[23] Kumar, V., Sharma, K. V., Caloiero, T., Mehta, D. J., & Singh, K. (2023). Comprehensive Overview of Flood Modeling Approaches: A Review of Recent Advances. Hydrology, 10(7), 141.

[24] Park, S. J., & Lee, D. K. (2020). Prediction of coastal flooding risk under climate change impacts in South Korea using machine learning algorithms. Environmental Research Letters, 15(9), 094052.

[25] Abbot, J., & Marohasy, J. (2014). Input selection and optimization for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. Atmospheric Research, 138, 166-178.

[26] Bayat, M., & Tavakkoli, O. (2022). Application of machine learning in flood forecasting. Future Technology, 1(1).

[27] Wagenaar, D., Curran, A., Balbi, M., Bhardwaj, A., Soden, R., Hartato, E., ... & Lallemant, D. (2020). Invited perspectives: How machine learning will change flood risk and impact assessment. Natural Hazards and Earth System Sciences, 20(4), 1149-1161.

[28] Senthil Kumar, A. R., Sudheer, K. P., Jain, S. K., & Agarwal, P. K. (2005). Rainfall - runoff modeling using artificial neural networks: comparison of network types. Hydrological Processes: An International Journal, 19(6), 1277-1291.

[29] Hussain, F., Wu, R. S., & Wang, J. X. (2021). Comparative study of very short-term flood forecasting using physics-based numerical model and data-driven prediction model. Natural Hazards, 107(1), 249-284.

[30] Yan, J., Jin, J., Chen, F., Yu, G., Yin, H., & Wang, W. (2018). Urban flash flood forecast using support vector machine and numerical simulation. Journal of Hydroinformatics, 20(1), 221-231.

[31] Sivapragasam, C., Liong, S. Y., & Pasha, M. F. K. (2001). Rainfall and runoff forecasting with SSA–SVM approach. Journal of Hydroinformatics, 3(3), 141-152.

[32] Simoes, N. E. D. C. (2012). Urban pluvial flood forecasting (Doctoral dissertation, Imperial College London).

[33] Bray, M., & Han, D. (2004). Identification of support vector machines for runoff modelling. Journal of Hydroinformatics, 6(4), 265-280.

[34] Marín-García, D., Rubio-Gómez-Torga, J., Duarte-Pinheiro, M., & Moyano, J. (2023). Simplified automatic prediction of the level of damage to similar buildings affected by river floods in a specific area. Sustainable Cities and Society, 88, 104251.

[35] Tayfur, G., Singh, V. P., Moramarco, T., & Barbetta, S. (2018). Flood hydrograph prediction using machine learning methods. Water, 10(8), 968.

[36] Dazzi, S., Vacondio, R., & Mignosa, P. (2021). Flood stage forecasting using machine-learning methods: a case study on the Parma River (Italy). Water, 13(12), 1612.

[37] Singh, P., Borah, B. (2013). Indian summer monsoon rainfall prediction using artificial neural network. Stoch Environ Res Risk Assess 27, 1585–1599. https://doi.org/10.1007/s00477-013-0695-0.

[38] Le, T., & Ngoc, T. (2020). Floods and household welfare: Evidence from Southeast Asia. Economics of Disasters and Climate Change, 4(1), 145-170. https://link.springer.com/article/10.1007/s41885-019-00055-x

[39] Le, X. H., Ho, H. V., Lee, G., & Jung, S. (2019). Application of long short-term memory (LSTM) neural network for flood forecasting. Water, 11(7), 1387. https://doi.org/10.3390/w11071387

[40] Mulualem, G. M., & Liou, Y. A. (2020). Application of artificial neural networks in forecasting a standardized precipitation evapotranspiration index for the Upper Blue Nile.

[41] Shamim, M.A., Hassan, M., Ahmad, S. et al. (2016). A comparison of Artificial Neural Networks (ANN) and Local Linear Regression (LLR) techniques for predicting monthly reservoir levels. KSCE J Civ Eng 20, 971–977. https://doi.org/10.1007/s12205-015-0298-z

[42] Haribabu, S., Gupta, G. S., Kumar, P. N., & Rajendran, P. S. (2021, July). Prediction of Flood by Rainf All Using MLP Classifier of Neural Network Model. In 2021 6th International Conference on Communication and Electronics Systems (ICCES) (pp. 1360-1365). IEEE..

[43] Band, S. S., Janizadeh, S., Chandra Pal, S., Saha, A., Chakrabortty, R., Melesse, A. M., & Mosavi, A. (2020). Flash flood susceptibility modeling using new approaches of hybrid and ensemble tree-based machine learning algorithms. Remote Sensing, 12(21), 3568.

[44] Bulti, D. T., & Abebe, B. G. (2020). A review of flood modeling methods for urban pluvial flood application. Modeling Earth Systems and Environment, 6(3), 1293-1302. https://doi.org/10.1007/s40808-020-00803-z

[45] Chen, W., Li, Y., Xue, W., Shahabi, H., Li, S., Hong, H., & Ahmad, B. B. (2020). Modeling flood susceptibility using data-driven approaches of naïve bayes tree, alternating decision tree, and random forest methods. Science of The Total Environment, 701, 134979.

[46] Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., & Mosavi, A. (2019). An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. Science of the Total Environment, 651, 2087-2096.

[47] Luo, P., Luo, M., Li, F., Qi, X., Huo, A., Wang, Z., ... & Nover, D. (2022). Urban flood numerical simulation: Research, methods and future perspectives. Environmental Modelling & Software, 105478.

[48] Mojaddadi, H., Pradhan, B., Nampak, H., Ahmad, N., & Ghazali, A. H. B. (2017). Ensemble machine-learning-based geospatial approach for flood risk assessment using multi-sensor remote-sensing data and GIS. Geomatics, Natural Hazards and Risk, 8(2), 1080-1102.

[49] Noymanee, J., & Theeramunkong, T. (2019). Flood forecasting with machine learning technique on hydrological modeling. Procedia Computer Science, 156, 377-386.

[50] Priscillia, S., Schillaci, C., & Lipani, A. (2021). Flood susceptibility assessment using artificial neural networks in Indonesia. Artificial Intelligence in Geosciences, 2, 215-222.

[51] Singh, P., Borah, B. (2013). Indian summer monsoon rainfall prediction using artificial neural network. Stoch Environ Res Risk Assess 27, 1585–1599. https://doi.org/10.1007/s00477-013-0695-0

[52] Wu, J., Liu, H., Wei, G., Song, T., Zhang, C., & Zhou, H. (2019). Flash flood forecasting using support vector regression model in a small mountainous catchment. Water, 11(7), 1327.

An Improved Depth Estimation using Stereo Matching and Disparity Refinement Based on Deep Learning

Deepa[1], Jyothi K[2], Abhishek A Udupa[3]
N.M.A.M Institute of Technology, Affiliated to NITTE (Deemed to be University), Nitte, Karkala,
Visvesveraya Technological University, Belagavi, Karnataka, India[1]
J.N.N College of Engineering, Shimoga, Visvesveraya Technological University, Belagavi, Karnataka, India[2]

***Abstract*—Stereo matching techniques are a vital subject in computer vision. It focuses on finding accurate disparity maps that find its use in several applications namely reconstruction of a 3D scene, navigation of robot, augmented reality. It is a method of obtaining corresponding matching point in stereo images to get disparity map. With additional details, this disparity map could be converted into a depth of a scene. Obtaining an efficient disparity map in the texture less, occluded, and discontinuous areas is a difficult job. A matching cost using an improvised Census transform and an optimization framework is proposed to produce an initial disparity map. The classic Census transform focus on the value of pixel at the center. If this pixel is prone to noisy condition, then the census encoding may differ which leads to mismatches. To overcome this issue an improved census transform based on weighted sum values of the neighborhood pixels is proposed which suppresses the noise during stereo matching. Additionally, a deep learning based disparity refinement technique using the generative adversarial network to handle texture less, occluded, and discontinuous areas is proposed. The suggested method offers cutting-edge performance in terms of both qualitative and quantitative outcomes.**

***Keywords*—Census transform; deep learning; depth; generative adversarial network; occlusion; stereo matching**

I. INTRODUCTION

Stereo matching has gathered attraction recently because of its applications in fields like visual entertainment, 3D reconstruction, autonomous driving, object detection [1], outdoor mapping, navigation and 3DTV [2], [3]. It is a research area that tries to imitate vision systems in humans by using two or several 2D views of the same scene to get three-dimensional depth details of the scene. It intends to find the corresponding relationship between matching pixels. A stereo matching algorithm uses stereo images that are rectified as an input [4], [5]. The horizontal displacement between the matching pixels is called disparity. With additional details, a disparity map could be transformed to a depth of scene. Disparity map accuracy is very crucial as small inaccuracies may affect the result. Obtaining an efficient and precise disparity map is a tedious task because of the existence of noise, occlusions, low textures, ill-posed regions, and the lighting conditions. Hence, it is significant to create a good disparity map.

Stereo matching techniques are classified as conventional algorithms and deep learning methods. Conventional algorithms are grouped into local and global algorithms. In local approaches, disparity is computed by comparing small areas [6] [7]. The disparity calculation relies on intensity in a defined support area. In real time the stereo images collected may be prone to noise, lighting distortions which reduces the efficiency of these algorithms. To overcome these drawbacks, a census transform in stereo matching is proposed in [8] which can decrease the effect of amplitude distortion. It aims at mapping the pixels to a binary string and then calculates the similarity between the pixels by means of Hamming distance. But, this method relies mainly on the central pixel, leading to false matching in a noisy environment. To reduce this shortcomings, a three-state census is proposed in [9] which is tolerant to any noise and enhances the robustness of stereo matching. An algorithm is implemented in [10] to perform census transform that reduces the noise interference and amplitude distortions in the images. A star-census transform (SCT) is introduced [11] that initiates the neighborhood pixel sampling in a symmetrical order that excludes the central pixel in the matching window. An improvised AD-Census stereo matching using gradient fusion (ADSG) is introduced in [12]. The absolute difference is used along with census transform for cost calculation, the result is then combined with gradient cost. These methods focus only on the information locally and hence have a low complexity and execute in shorter time. But the results generated by these local methods in the areas of occlusion, texture less and discontinuities is not satisfying.

The semi global algorithm was proposed in [13]. The accuracy and computational efficiency of semi global algorithms lies in between that of local and global algorithms. A global method considers disparity computation as a global energy minimization method for all disparity values. The energy function has two terms namely data term which penalizes pixels with inconsistent values and smoothness term with enforces smoothing constraint by considering the neighboring pixels. Some of the commonly used global algorithms are graph cuts algorithm [14] and belief propagation technique [15]. A disparity estimation based on tree structure named Pyramid-tree is introduced in [16]. It performs cross regional smoothing that can handle low texture regions. Global methods can generate a good quality disparity map, but they are also quite expensive and time-consuming.

Deep stereo methods are popular these days. Zbontar et al. [17] used a network to get patch-wise details to compute matching cost. The proposed network is trained to find the similarity that exists between a pair of images. It is then processed using classic post processing. The GC-Net [18] is a network with a high performance. It applied 3D convolution kernel to the correspondence space and proposed disparity refinement. This provided improvement over the previous approach. A pyramid stereo matching network is proposed in [19]. It improved the feature extraction by means of multi scale feature extraction network [20]. A network namely cascaded residual learning [21] was introduced which uses a DispNet. This is made up of two sub parts called DispFullNet and a DispResNet. The first network computes the raw disparity map. The second network tries to optimize the raw disparity map by computing the multiscale residual information. Williem et al. [22] introduced a method known as self-guided cost aggregation that uses a convolution network for local stereo matching. The network is made up of emotional weight network and descent filtering network. In LEA Stereo [23] a search is performed to streamline matching pipeline. Shivam Duggal et al. [24] developed a trainable network. Many recent papers introduced refinement components steps to improvise the disparity map quality. The MSMD-Net [25] introduced multi scale technique in which the stereo images are processed using multi resolution pyramid network. The RAFT- stereo [26] consists of a network for stereo estimation along with refinement stage. The deep learning-based methods can produce depth map from a given stereo image pairs, but these stereo methods still find it difficult to find correct correspondences in texture less and the occluded regions.

Though several techniques have been proposed to improvise the matching accuracy, the low accuracy in the occluded and texture less regions has not been handled very well. A depth estimation technique using improvised census transform and disparity refinement using deep learning to enhance the results in occluded and texture less regions is proposed . The weighted sum of the center pixel and its four neighbors is used to calculate the center pixel value in the improvised census transform in order to reduce noise in initial disparity map. The occluded and texture less regions of initial disparity map are refined using Generative adversarial network (GAN) deep learning framework. The extensive experiments performed on Middlebury datasets shows the efficacy of our method. Our method improvises the efficiency of disparity map by a considerable amount. The suggested method is explained in Section II. The outcomes of the suggested method are shown in Section III. In Section IV, the paper's conclusions are discussed.

II. METHODOLOGY

The proposed method applies improved census transform is applied for the stereo images and a matching cost is obtained using Hamming distance. Then, a cost aggregation is carried using semi global method to compute an initial disparity map. Finally, a disparity refinement network using GAN is proposed to increase the efficiency of disparity map from which depth is estimated. An overview of the whole methodology is Fig. 1.

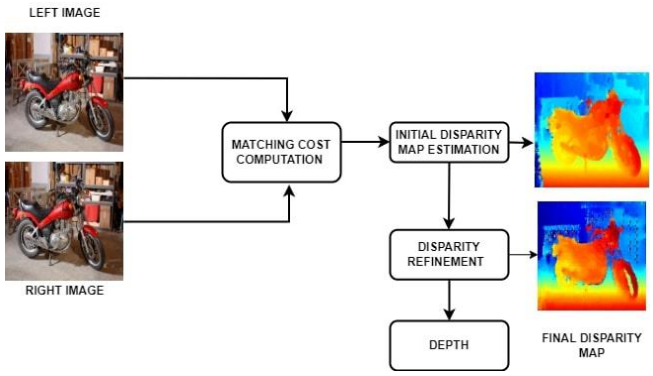


Fig. 1. Block diagram of the methodology.

A. Improves Census Transform

Methods for stereo matching based on intensity difference contain a lot of errors, especially for the outdoor images. To overcome these drawbacks Census transform (CT) method is used for computing matching cost. It is a local method that relies on relative ordering of pixels rather than intensity within a fixed window. Hence it can efficiently handle radiometric variations like lighting changes and illumination differences and discontinuities. The traditional CT is shown in Fig. 2. Census transform consider the center pixel value, compares with all the remaining pixels and assigns the 1 if the center pixel value is less than the compared pixel, otherwise 0 is assigned. It is then represented as a binary bit string.

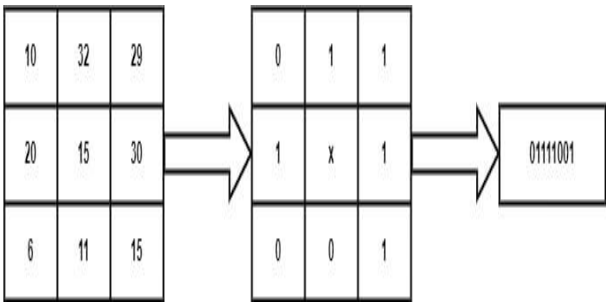


Fig. 2. Traditional census transform.

Census transform is represented by the following equation

$$T_{cen} = \otimes_{i \in w_p} \xi[I(p), I(m)] \quad (1)$$

$$\xi[I(p), I(m)] = \begin{cases} 1, & I(p) \le I(m) \\ 0, & otherwise \end{cases} \quad (2)$$

Here, $\otimes$ is a bitwise operation, w_p represent window with p as centre pixel, m is any point in w_p and $I(p), I(m)$ are pixel values of points p and m respectively.

The traditional CT can reduce the impact of distortions in amplitude, but it depends heavily on the middle pixel. It is prone to noise, as it measures the relative difference of the neighboring pixels based on middle pixel. When the center pixel is affected by noise, encoding from the census transform might vary drastically which may lead to mismatched pixels. Due to noise if the center pixel value changes from 15 to 35, the traditional CT transformation for a 3 X 3 patch of image is depicted in Fig. 3. Since the traditional CT depends on the

pixel at the center, the noisy center pixel value 35 is considered to compare with the remaining pixels in the image patch. The census code obtained is 00000000. Here there is a difference in 5 bits as compared to the initial code 01111001.

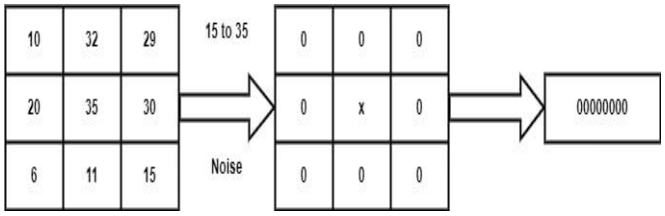


Fig. 3. Traditional census transform in noisy condition.

Aiming to overcome this drawback, an improvised census method is proposed in this paper where the weighted summation of pixel at the center and the four neighboring pixels is used to update the center pixel.

Let $I(m,n)$ be the center pixel. The weight distribution of pixel of the center and the four neighboring pixels are:

$$\mathrm{wt(m,n)} = 0.4 \quad (3)$$

$$\mathrm{wt(m+1,\ n)} = 0.15 \quad (4)$$

$$\mathrm{wt(x-1,\ y)} = 0.15 \quad (5)$$

$$\mathrm{wt(m,n+1)} = 0.15 \quad (6)$$

$$\mathrm{wt(m,\ n-1)} = 0.15 \quad (7)$$

The weights are assigned in such a manner that the weight of each pixel lies in between 0 and 1 and the total weighted sum of the pixel at the center and four neighboring pixel is 1.

The weighted sum of pixel centered at (x, y) is computed using the following equation.

$$I_{wt}(m,n) = I(m,n)wt(m,n) + I(m+1,n)wt(m+1,n) + I(m-1,n)wt(m-1,n) + I(m,n+1)wt(m,n+1) + I(m,n-1)wt(m,n-1) \quad (8)$$

The following equation is used to update the value of the center pixel.

$$I_{mid}(m,n) = \begin{cases} \{ I(m,n), & |I_{wt}(m,n) - I(m,n) \le T| \\ I_{wt}(m,n), & |I_{wt}(m,n) - I(m,n) > T| \end{cases} \quad (9)$$

If the variation between the weighted sum of center pixel and the original center pixel is more than the threshold $T = 6$, then the pixel value in the center is updated by the weighted sum otherwise the original is used.

The improvised technique for census transform proposed in the paper is depicted in Fig. 4.

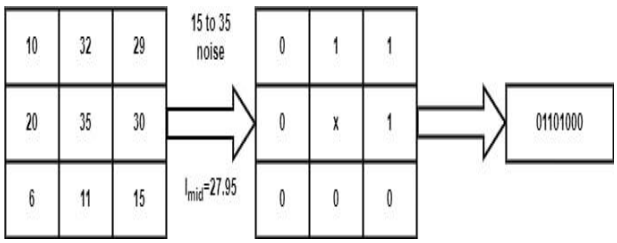


Fig. 4. Census transform of the proposed method in noisy condition.

Due to noise if center pixel value change from 15 to 35, the improvised method proposed in the paper updates the value of center pixel to 27.95 using the weighted sum as shown in Fig. 4 (i.e 35*0.4+11*0.15+32*0.15+30* 0.15+20*0.15). This value is compared with the remaining pixels in the patch to get the census code 01101000. The pixels differ only by 2 bit as compared to original code 01111001. This demonstrates how the approach is noise-resistant and improves matching performance.

B. Matching Cost Computation

To ascertain whether the values between two pixels indicate the matching point of a scene, a matching computation of cost is carried out. After the census transform the correspondence of pixels can be determined using Hamming distance [8]. The Hamming distance between matching points is found to estimate the correspondence between matching points. Let $S_{cenL}(p)$ be a binary bit array of pixel p in the left stereo image and $S_{cenR}(q)$ be the binary bit array of pixel q in right stereo image for disparity d. The following calculation uses the Hamming distance to compute the matching census cost between p and q.

$$C(p,d) = Hamming[S_{cenL}(p), S_{cenR}(q)] \quad (10)$$

The cost computation for the center pixel of $3\,X\,3$ image patch is depicted in Fig. 5.

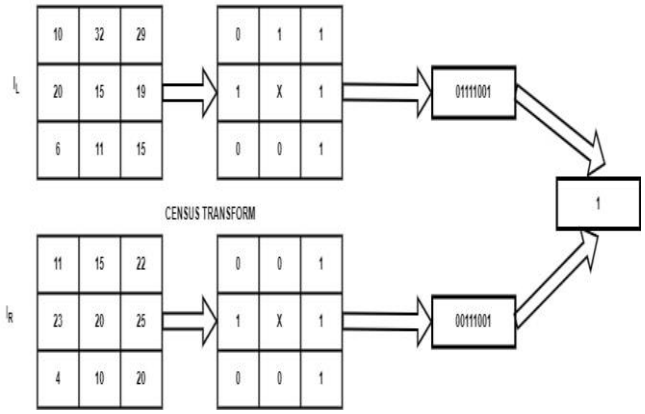


Fig. 5. Matching cost computation.

C. Initial Disparity Estimation

Due to noise, the pixel wise cost may produce ambiguous results. Hence additional constraint is included to get a smooth disparity by penalizing the changes in the neighboring pixels [27]. The smoothness constraint and pixel wise cost is represented by the energy function $E(D)$.

$$E(D) = \sum_P C(p, D_P) + \sum_{x \in N_p} P_1\, T\left[\left|D_p - D_x\right| = 1\right] + \sum_{x \in N_p} P_2\, T\left[\left|D_p - D_x\right| > 1\right] \quad (11)$$

$C(p, D_P)$ is the cost summation of all the pixel for disparity D. P_1 is the penalty applied to pixels x in N_P with low disparity difference. P_2 is the penalty for pixels in N_p with high disparity difference.

The stereo matching problem aims to minimize the energy function$E(D)$. Finding the minimum energy function $E(D)$ is computationally expensive. The energy function is

approximated by aggregating the matching cost from all directions $'r'$. The total number of directions r is 8. The cost at direction r is represented by,

$$S_r\ (p,d) = C(p,d) + min\big[S_r(p-r,d), S_r\ (p-r,d-1) + P_1, S_r(p-r,d+1), S_r(p-r,i) + P_2\big] \quad (12)$$

$C(p,d)$ is cost for pixel p and $S_r(p-r,d)$ is the pixel cost at direction r with disparity d, $S_r\ (p-r,d-1)$ is cost at direction $'r'$ and disparity $'d-1'$. $S_r(p-r,d+1)$ denotes cost for disparity $'d+1'$ and direction $'r'$. S_r is minimum pixel cost at direction r.

The following equation is then used to determine the initial disparity.

$$D(p) = argmin \sum S_r\ (p,d) \quad (13)$$

D. Disparity Refinement

The initial disparity calculated may contain wrongly matched disparities at the object boundaries, occluded areas and the texture less regions. Finding the correct disparities in these areas is a difficult task. Hence, an appropriate disparity refinement method is needed. A disparity refinement is performed based on deep learning-based technique using the generative adversarial network (GAN) to handle texture less, occluded, and discontinuous areas. The method proposed uses GAN network introduced by Good fellow [28] for disparity refinement. GAN includes two networks called generator and a discriminative network that are implemented based on neural networks. The generator takes initial disparity map as its input and focuses on generating a refined disparity map. The discriminator is fed with ground truth disparity along with disparity map produced by the generator. The discriminator aims to differentiate the ground truth disparity and generated refined disparity map. The feedback from the discriminator is given to the generator to fine tune the generated image. This procedure is repeated until the resulting disparity resembles the ground truth disparity. The disparity refinement network proposed in the paper is depicted in Fig. 6.

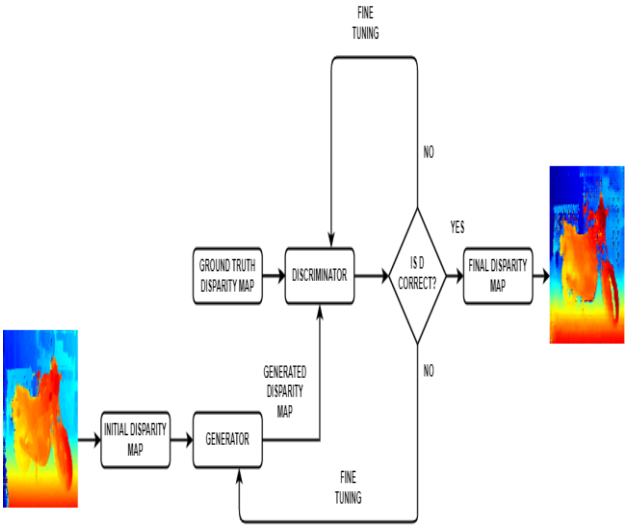


Fig. 6. Architecture of disparity refinement network.

A Pix2Pix GAN [29] is used to refine the disparity map. Pix2Pix GAN is an adversarial network. Pix2Pix GAN is known for the capacity of producing high quality images. The initial disparity is given as input to the generator. The generator generates the disparity map which is then fed to the discriminator. The various generator networks available are UNET 128, ResNet 6 and Resnet 9. The proposed disparity refinement network uses UNET128 as it can learn with few training images. The architecture of UNET 128 generator used in the proposed approach is depicted in Fig. 7.

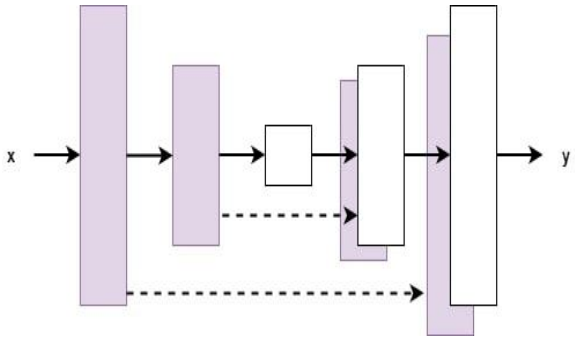


Fig. 7. Architecture of generator.

It uses a network consisting of several convolutional layer, batch normalization, dropout, and activation layers. It is trained using adversarial loss and then revised by means of L_1 loss. This loss drives the generator to generate image close to ground truth disparity. The generator is then updated using a sum of L_1 loss and a loss called as adversarial loss. A comparative study of UNet 128 with other networks such as ResNet 6 and ResNet 9 is given in Table I. ResNet 6 and Resnet 9 are the deep residual networks which include 6 residual blocks and nine residual blocks respectively. The information details are passed via a shortcut connection. Convolutional, batch-normalization, and corrected Liner Unit (ReLU) layers make up a traditional residual block. For evaluation, measurements like squared relative difference (SRD) and absolute relative distance (ARD) are used. Lower values of the above metrics indicate better performance. The efficiency of the generator architecture is shown in Table I. Results obtained for UNet 128 is better than the ResNet 6 and ResNet 9.

$$ARD = \frac{1}{N}\sum \frac{d_t\,(p,q) - d_g(p,q)}{d_t(p,q)} \quad (14)$$

$$SRD = \frac{1}{N}\sum \frac{\left|d_t(p,q) - d_g(p,q)\right|^2}{d_t(p,q)} \quad (15)$$

Here d_t is generated disparity map, d_g is ground truth disparity map. N is the total pixels.

TABLE I. COMPARATIVE STUDY OF GENERATOR ARCHITECTURE

	ResNet 6	ResNet 9	UNet 128	
ARD	0.058	0.051	0.037	Lower is better
SRD	0.409	0.413	0.403	

The discriminator is based on Patch GAN model. This PatchGAN model provides extremely high frequency information. The GAN's primary objective is described as,

$$L_{GAN}(G,D) = E_{p,q}\left[logD(p,q)\right] + E_{p,r}\left[log\left(1 - D\big(p,G(p,r)\big)\right)\right] \quad (16)$$

Here, p de represent the a ground truth disparity, q denote the generated disparity and r denotes the initial disparity map

The generator G attempts to decrease the objective as response to the discriminator D which attempts to increase it. The result is as follows:

$$G^* = argmin_g max_d L_{GAN}(G,D) \quad (17)$$

The aim of G is to decrease the objective and the generator updates itself using Loss L_1. It is computed as,

$$Loss_{L_1}(G) = E_{p,q,r}\left[\|\big(q - G(p,r)\big)\|_1\right] \quad (18)$$

The objective is updated as

$$G^* = argmin_g max_d L_{GAN}(G,D) + \lambda Loss_{L_1}(G) \quad (19)$$

The performance of training model is depicted in Fig. 8. Here the training loss decreases gradually as the number of epochs increases. Lower the loss, the more effective the model is.

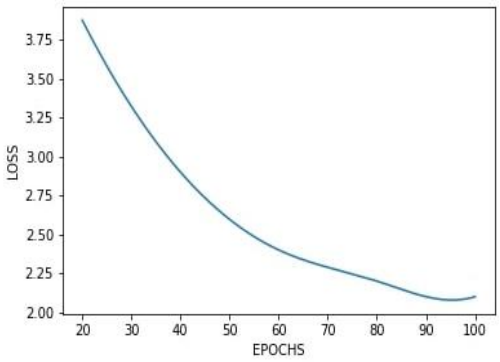


Fig. 8. Training loss versus epochs.

E. Depth

Once the disparity map is generated for the stereo images, the depth Z is estimated using the following formula:

$$Z = \frac{f\ X\ B}{D} \quad (20)$$

Fig. 9. Point cloud generated for motorcycle image.

Here f is focal length, B is stereo camera baseline. These values are obtained from the stereo calibration. Once the depth is estimated the exact coordinates of each pixel in the scene can be computed. These coordinates are made used to construct point clouds. The point cloud generated for Motorcycle image is shown in Fig. 9. The coordinates are stored in polygon format file. The output ply file is plotted using ply file plotters. The point cloud shown above was constructed using open3d.

III. Results and Discussion

The experiments were performed using Middlebury dataset [30], [31] images to analyze the performance. The refinement network is trained using the Pytorch framework on a personal machine. The computer hardware environment used is a Dual Intel-Xeon E5-2609V4 8C having 1.7 GHz 20M 6.4 GT/s and 128GB Memory. A Dual NVDIA Tesla server P100 GPU having 3584Cores and maximum of 18.7 TeraFLOPS is used. The datasets are downscaled to 256 pixels width and 256 pixels height for computational purposes. The Adam optimizer is used to optimize the discriminator. The learning rate is 0.0002. The GAN models does not converge, hence a balance has to be established between the generator and discriminator. The number of epochs is 100.

A. Middlebury Dataset

The Middlebury dataset includes rectified stereo images from indoor and outdoor surroundings utilizing a stereo vision concept. These images are complex, and it has images of different characteristics such as, different resolutions and low texture areas. Hence, the dataset consists of complex images for framework evaluation. Our stereo matching technique is robust to occluded and non-textured regions. The details of testing images like Cones, Teddy and Venus from Middlebury 2001 and 2003 are given in Table II. The details of higher resolution images from Middlebury 2014 are given in Table III.

TABLE II. Image of Middlebury 2001 and 2003

Images	Disparity level	Image resolutions
Cones	60	450 × 375
Teddy	60	450 × 375
Venus	20	434 × 383

TABLE III. Image of Middlebury 2014

Images	Disparity level	Image resolutions
Adirondak	73	718 × 496
ArtL	64	347 × 277
Jadeplant	160	659 × 497
Motorcycle	70	741 × 497
Pipes	75	735 × 485
Playroom	83	699 × 476
Playtable	73	680 × 463
PlaytableP	73	681 × 462
Recycle	65	720 × 486
Shelves	60	738×497

B. Noise Resistance Test

The traditional CT heavily depends on the center pixel. If this pixel is prone to noisy condition, then the census encoding may differ which leads to mismatches. To analyze the

efficiency of the proposed improved census transform in a noisy condition, salt and pepper noise of 2% and 5% noise is applied to Cones, Teddy and Venus images. The qualitative results for Teddy image when 2% salt and pepper noise is applied is represented in Fig. 10 and Table IV shows the percentage of bad matching pixels (PBMP) of the initial disparity map for the proposed improved census transform and traditional CT. The outcome conclude that results of the method proposed in the noisy condition is remarkably good than the traditional CT.

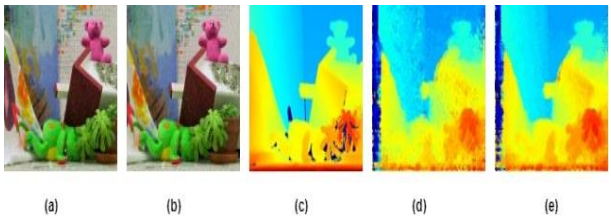


Fig. 10. Visual results for initial disparity map on noisy Teddy image (a) Left reference image (b) Right image (c) Ground Truth Disparity (d) Initial Disparity Map using traditional CT (e) Initial Disparity Map using proposed method.

TABLE IV. PBMP OF INITIAL DISPARITY MAP

	Salt & Pepper Noise (2%)		Salt & Pepper Noise (5%)	
	Traditional CT	**Proposed**	**Traditional CT**	**Proposed**
Cones	8.984	7.556	12.990	8.980
Teddy	13.492	9.715	19.492	12.103
Venus	3.229	1.370	6.062	2.740

C. Qualitative Results

The initial and improved disparity maps estimated by the suggested method are shown in Fig. 11. In the Fig. 11, the Fig. 11(a) is the left reference image. Fig. 11(b) is the right image. The ground truth disparity is given Fig. 11(c). The fourth column Fig. 11(d) represent initial disparity map. The refined disparity map is represented in Fig. 11(e). The red rectangular regions marked in Fig. 11(d) represent the occluded areas which are filled in the refined disparity map obtained by the suggested approach. The yellow circular region marked in the initial disparity of the Venus image shows the texture less region which is filled in the refined disparity map. It is discovered that the suggested method effectively creates high-quality disparity maps in noisy, textureless, and occluded regions.

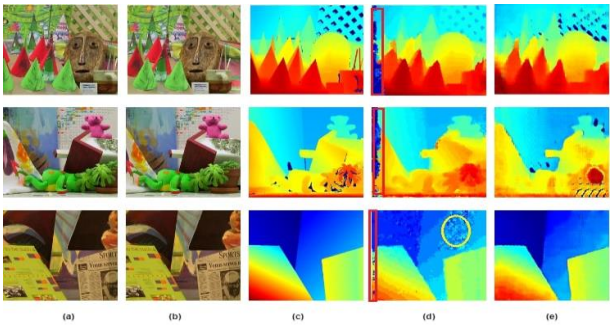


Fig. 11. Visual results on Cone, Teddy, and Venus images (a) Left image (b) Right image (c) Ground Truth Disparity (d) Initial Disparity Map (e) Refined disparity map.

The disparity maps generated for images such as Jade Plant, Adirondack, Motorcycle and Recycle are presented in first, second, third and fourth rows respectively in Fig. 12. The first row of the Fig. 12 shows a Jade Plant image from Middlebury dataset. This image is very challenging to match due to brightness difference. But, the proposed method has correctly discovered the disparities. The second and third rows of Fig. 12 shows Adirondack and Motorcycle images. The texture less surfaces in the initial disparity map of Adirondack image is highlighted by the yellow circular region. These regions are well recreated by the proposed approach. The fourth row of Fig. 12 shows Recycle image. The occluded areas in the initial disparity map is highlighted by the red rectangular region. The possibility of getting wrong matches in these regions are very high. These occluded areas are filled accurately in the estimated disparity map. We find that the proposed method produces efficient results in occluded and texture-less regions.

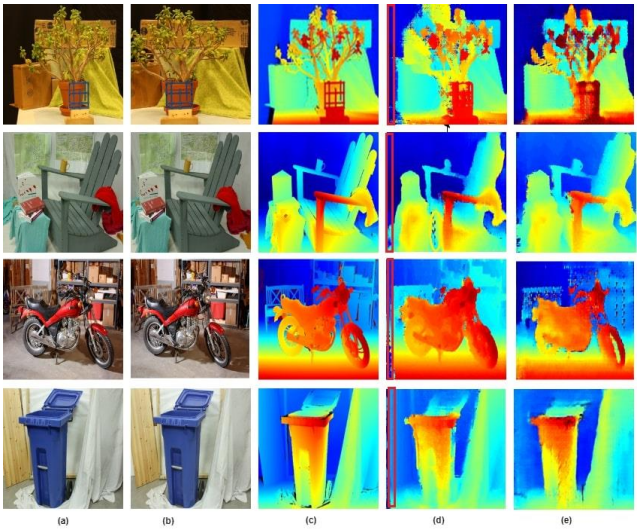


Fig. 12. Visual results on Jade Plant ,Adirondack, , Motorcycle and Recycle images (a) Left reference image (b) Right image (c) Ground Truth image (d) Initial Disparity Map (e) Refined disparity map.

D. Evaluation Metrics

The quantitative analysis is performed using the evaluation metrics namely root mean square error (RMSE) and PBMP. The efficiency increases when PBMP and RMSE values decrease. N be the number of pixels. d_t and d_g be the disparity map estimated and ground truth disparity maps respectively.

RMSE is calculated as:

$$RMSE = \left[\frac{1}{N}\sum\left|d_t(x,y) - d_g(x,y)\right|^2\right]^{\frac{1}{2}} \quad (21)$$

PBMP is calculated as follows:

$$PBMP = \left[\frac{1}{N}\sum\left|d_t(x,y) - d_g(x,y)\right| > T\right] * 100 \quad (22)$$

E. Comparison with Existing Methods

The proposed method is compared with ADSG [12] and Deep Pruner [24]. The results for the methods compared are obtained from Middlebury evaluation leader board. An improved AD-Census method using gradient fusion is used in

ADSG. The absolute difference and census transform is used for cost calculation, which is then combined with gradient cost. This method focus only on local information and hence do not give satisfying results in the areas of occlusion, texture-less and discontinuities. Deep Pruner uses a trainable network. It do not produce satisfactory results in the occluded regions. Tables V and VI demonstrate that the comparison of RMSE and PBMP results.

TABLE V. COMPARISON OF RMSE RESULTS

Images	ADSG	Deep Pruner	Proposed
Adirondack	19.5	6.18	5.14
ArtL	24.6	9.50	6.22
Jade plant	25.8	28.2	5.44
Motorcycle	79.6	10.3	6.52
Pipes	32.1	13.9	6.74
Playroom	35.2	8.91	7.85
Playtable	50.0	4.89	3.52
PlaytableP	19.9	4.74	3.54
Recycle	17.6	3.81	5.55
Shelves	21.9	4.28	5.66
Avg	32.62	9.471	5.618

TABLE VI. COMPARISON OF PBMP FOR THRESHOLD=1

Images	ADSG	Deep Pruner	Proposed
Adirondack	38.9	39.7	23.97
ArtL	35.5	41.8	42.36
Jade plant	49.8	62.8	44.50
Motorcycle	43.2	45.3	45.92
Pipes	41.5	53.8	34.78
Playroom	57.8	57.7	26.08
Playtable	64.4	48.2	46.28
PlaytableP	42.2	41.7	47.15
Recycle	37.5	36.8	27.36
Shelves	65.0	54.2	45.20
Avg	47.58	48.2	38.36

Our technique yields the lowest average RMSE and PBMP, as shown in Table V and Table VI. This signifies the accuracy and competitiveness of our method as compared to ADSG [12] and Deep Pruner [24]. The proposed method produces average RMSE 5.68 and PBMP 38.36%. The improved census transform in matching cost is robust to noise. Additionally the disparity refinement based on deep learning-based technique using the generative adversarial network (GAN) handle texture less, occluded, and discontinuous areas and produce a good quality disparity map.

IV. CONCLUSION

A stereo matching method that is based on improvised census transform along with an optimization framework is proposed to determine the initial disparity map. Further disparity refinement is carried out using GAN to obtain the depth of a scene. The traditional census transform heavily depends on center pixel. If this pixel is prone to noise, then census encoding generated will differ which may lead to false matching. To handle this issue an improved census cost that relies on the weighted sum values is proposed. In the disparity refinement stage a deep learning based network using GAN is proposed which can handle outliers and enhance the correctness of matching. The efficiency of the suggested strategy is assessed using images from Middlebury benchmark. The comparison with the current system showed that the proposed method works better than other methods.

REFERENCES

[1] H. M. Wang, H. Y. Lin, and C. C. Chang, "Object detection and depth estimation approach based on deep convolutional neural networks," *Sensors*, vol. 21, no. 14, 2021, doi: 10.3390/s21144755.

[2] M. Menze, C. Heipke, and A. Geiger, "Object Scene Flow," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 60–76, 2018, doi: 10.1016/j.isprsjprs.2017.09.013.

[3] S. Hong, M. Li, M. Liao, and P. Van Beek, "Real-time mobile robot navigation based on stereo vision and low-cost GPS," *IS T Int. Symp. Electron. Imaging Sci. Technol.*, pp. 10–15, 2017, doi: 10.2352/ISSN.2470-1173.2017.9.IRIACV-259.

[4] R. A. Hamzah and H. Ibrahim, "Literature survey on stereo vision disparity map algorithms," *J. Sensors*, vol. 2016, 2016, doi: 10.1155/2016/8742920.

[5] K. Y. Kok and P. Rajendran, "A review on stereo vision algorithms: Challenges and solutions," *ECTI Trans. Comput. Inf. Technol.*, vol. 13, no. 2, pp. 134–151, 2019, doi: 10.37936/ecti-cit.2019132.194324.

[6] C. S. Huang, Y. H. Huang, D. Y. Chan, and J. F. Yang, "Shape-reserved stereo matching with segment-based cost aggregation and dual-path refinement," *Eurasip J. Image Video Process.*, vol. 2020, no. 1, pp. 1–9, 2020, doi: 10.1186/s13640-020-00525-3.

[7] Deepa and K. Jyothi, "A Robust Disparity Map Estimation for Handling Outliers in Stereo Images," *2021 5th Int. Conf. Electr. Electron. Commun. Comput. Technol. Optim. Tech. ICEECCOT 2021 - Proc.*, no. December, pp. 38–43, 2021, doi: 10.1109/ICEECCOT52851.2021.9708034.

[8] R. Zabih and J. Woodfill, "Non parametric local transforms.pdf," 1994.

[9] Y. Men, N. Ma, G. Zhang, X. Li, C. Men, and P. Sun, "A stereo matching algorithm based on Census transform and improved dynamic programming," *Harbin Gongye Daxue Xuebao/Journal Harbin Inst. Technol.*, vol. 47, no. 3, pp. 60–65, 2015, doi: 10.11918/j.issn.0367-6234.2015.03.010.

[10] N. Y. C. Chang, T. H. Tsai, B. H. Hsu, Y. C. Chen, and T. S. Chang, "Algorithm and architecture of disparity estimation with mini-census adaptive support weight," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 792–805, 2010, doi: 10.1109/TCSVT.2010.2045814.

[11] J. Lee, D. Jun, C. Eem, and H. Hong, "Improved census transform for noise robust stereo matching," *Opt. Eng.*, vol. 55, no. 6, p. 063107, 2016, doi: 10.1117/1.oe.55.6.063107.

[12] H. Liu *et al.*, "Stereo matching algorithm based on two-phase adaptive optimization of AD-census and gradient fusion," *2021 IEEE Int. Conf. Real-Time Comput. Robot. RCAR 2021*, pp. 726–731, 2021, doi: 10.1109/RCAR52367.2021.9517511.

[13] H. Hirschmüller, "Accurate and efficient stereo processing by semi-global matching and mutua information," *Proc. - 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, CVPR 2005*, vol. II, no. 2, pp. 807–814, 2005, doi: 10.1109/CVPR.2005.56.

[14] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2, pp. 508–515, 2001, doi: 10.1109/iccv.2001.937668.

[15] J. Sun, H. Y. Shum, and N. N. Zheng, "Stereo matching using belief propagation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2351, no. 7, pp. 510–524, 2002, doi: 10.1007/3-540-47967-8_34.

[16] C. Xu, C. Wu, D. Qu, F. Xu, H. Sun, and J. Song, "Accurate and Efficient Stereo Matching by Log-Angle and Pyramid-Tree," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 4007–4019, 2021, doi: 10.1109/TCSVT.2020.3044891.

[17] J. Žbontar and Y. Lecun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol.

17, pp. 1–32, 2016.

[18] A. Kendall *et al.*, "End-to-End Learning of Geometry and Context for Deep Stereo Regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 66–75. doi: 10.1109/ICCV.2017.17.

[19] R. Qin, X. Huang, W. Liu, and C. Xiao, "Pairwise stereo image disparity and semantics estimation with the combination of U-net and pyramid stereo matching network," *Int. Geosci. Remote Sens. Symp.*, vol. 2019-July, pp. 4971–4974, 2019, doi: 10.1109/IGARSS.2019.8900262.

[20] X. Li, T. Lai, S. Wang, Q. Chen, C. Yang, and R. Chen, "Weighted feature pyramid networks for object detection," *Proc. - 2019 IEEE Intl Conf Parallel Distrib. Process. with Appl. Big Data Cloud Comput. Sustain. Comput. Commun. Soc. Comput. Networking, ISPA/BDCloud/SustainCom/SocialCom 2019*, pp. 1500–1504, 2019, doi: 10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00217.

[21] J. Pang, W. Sun, J. S. J. Ren, C. Yang, and Q. Yan, "Cascade Residual Learning: A Two-Stage Convolutional Neural Network for Stereo Matching," *Proc. - 2017 IEEE Int. Conf. Comput. Vis. Work. ICCVW 2017*, vol. 2018-Janua, pp. 878–886, 2017, doi: 10.1109/ICCVW.2017.108.

[22] Williem and I. K. Park, "Deep self-guided cost aggregation for stereo matching," *Pattern Recognit. Lett.*, vol. 112, pp. 168–175, 2018, doi: 10.1016/j.patrec.2018.07.010.

[23] X. Cheng *et al.*, "Hierarchical neural architecture search for deep stereo matching," *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, no. NeurIPS, pp. 1–12, 2020.

[24] S. Duggal, S. Wang, W. C. Ma, R. Hu, and R. Urtasun, "Deeppruner: Learning efficient stereo matching via differentiable patchmatch," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-Octob, pp. 4383–4392, 2019, doi: 10.1109/ICCV.2019.00448.

[25] Z. Shen, Y. Dai, and Z. Rao, "MSMD-Net: Deep Stereo Matching with Multi-scale and Multi-dimension Cost Volume," 2020, [Online]. Available: http://arxiv.org/abs/2006.12797

[26] L. Lipson, Z. Teed, and J. Deng, "RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching," *Proc. - 2021 Int. Conf. 3D Vision, 3DV 2021*, pp. 218–227, 2021, doi: 10.1109/3DV53792.2021.00032.

[27] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, 2008, doi: 10.1109/TPAMI.2007.1166.

[28] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020, doi: 10.1145/3422622.

[29] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, pp. 5967–5976. doi: 10.1109/CVPR.2017.632.

[30] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, no. June, pp. 195–202, 2003, doi: 10.1109/cvpr.2003.1211354.

[31] X. Jiang, J. Hornegger, and R. Koch, "Pattern recognition: 36th German Conference, GCPR 2014 Münster, Germany, September 2–5, 2014 proceedings," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8753, no. 2, pp. 31–42, 2014, doi: 10.1007/978-3-319-11752-2.

Federated-Learning Topic Modeling Based Text Classification Regarding Hate Speech During COVID-19 Pandemic

Muhammad Kamran[1], Ammar Saeed[2], Ahmed Almaghthawi[3]
Department of Cybersecurity, College of Computer Science and Engineering, University of Jeddah, 21959, Kingdom of Saudi Arabia[1]
Department of Computer Science, COMSATS University Islamabad, Wah Campus, WahCantt, Pakistan[2]
Department of Computer Science, College of Science and Art at Mahayil, King Khalid University, Abha 62529, Saudi Arabia[3]

Abstract—One of the most challenging tasks in knowledge discovery is extracting the semantics of the content regarding emotional context from the natural language text. The COVID-19 pandemic gave rise to many serious concerns and has led to several controversies including spreading of false news and hate speech. This paper particularly focuses on Islamophobia during the COVID-19. The widespread usage of social media platforms during the pandemic for spreading of false information about Muslims and their common religious practices has further fueled the existing problem of Islamophobia. In this respect, it becomes very important to distinguish between the genuine information and the Islamophobia related false information. Accordingly, the proposed technique in this paper extracts features from the textual content using approaches like Word2Vec and Global Vectors. Next, the text classification is performed using various machine learning and deep learning techniques. The performance comparison of various algorithms has also been reported. After experimental evaluation, it was found that the performance metric like F1-score indicate that Support Vector Machine performs better than other alternatives. Similarly, Convolutional Neural Network also achieved promising results.

Keywords—Knowledge extraction; text mining; pandemics and society; hate speech; Islamophobia

I. INTRODUCTION

One of the most challenging tasks in knowledge discovery is extracting the semantics of the content regarding emotional context from the natural language text. The COVID-19 pandemic gave rise to many serious concerns and has led to several controversies including spreading of false news and hate speech. Hate speech becomes more emotionally hurting if it targets someone's belief. In this paper, we particularly focus on Islamophobia which is a type of racism that is being practiced by anti-Muslim communities, individuals, groups, and organizations against Islam and Muslims [1]. It is one of the most visible forms of racism in the modern-day and several relevant incidents are reported on daily basis. but it is still not being given due attention and consideration as a global issue. The internet and social media are one of the primary means of disseminating fake news and false information around the world [2], [3], [4], [5]. Consequetenly, Muslim community is facing several challenges in their daily as well as professional life where they are in minority in different parts of the world [6]. Moreover, global Islamophobia has increased significantly because of COVID-19. On social media platforms, false information, hate speech, and conspiracy theories regarding Muslims have been circulated, further stigmatizing them. Also, the stigmatization of Muslims and others of Asian heritage has resulted from the pandemic's genesis in Wuhan, China. Discriminatory laws, such as the travel bans imposed by some nations on nations with most Muslims, have made the issue worse by feeding already-existing anti-Islamic attitudes. As a result, the epidemic has acted as a trigger for the escalation of Islamophobia, maintaining prejudice and unfavorable views towards Muslims. The role of social media usage during the pandemic has evidently played a major role in spreading Islamophobia [7], [8].

Conspiracy theories and false allegations about Muslims being to blame for the virus's spread have been propagated over social media during COVID-19. The spread of this misinformation on social media sites has fueled an upsurge in anti-Muslim sentiment. Muslims have been held responsible for the virus's spread on multiple occasions. For instance, after a religious gathering was conducted in New Delhi in March 2020, there were rumors of Muslims being held responsible for spreading the disease throughout India. Islamophobia increased as a result, with some calling the illness the "Muslim virus" or the "Tablighi virus" in India [9], [10].

Although some of the social media platforms implement the procedures for preventing the spread of hate speech; however, automation of such measures is still an ongoing research area. In this respect, the Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) can assist developing such automated methods [11]. These ML and DL techniques are most widely used in the sentiment analysis of the textual content from social media platforms and have yielded excellent outcomes thus far. Moreover, they are several other applications of ML and DL techniques in various domains like mentioned in [12], [13] and [14]. Our goal is to use them for tackling the spread of hate speech such that results in Islamophobia.

In this paper, we focus on the identification and classification of Islamophobic content originated during the COVID-19 pandemic. First of all, we perform the data collection step followed by preprocessing of the data. For this, we extracted one dataset from the Google fact-checking

platform while another dataset was collected from the tweets of social media platform X (formerly known as Twitter). We also performed the analysis of data using approaches like) Bag of Words (BoW), Term Frequency – Inverse Document Frequency (TF-IDF), etc. For classification, we used: (i) ML techniques like, Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression (LR), and Random Forest (RF); (ii) NLP transformer-based algorithms like Generative Pre-Trained Transformer (GPT) and Bidirectional Encoders Representation from Transformers (BERT); and (iii) DL models Long Short-Term Memory Network (LSTM) and Convolutional Neural Network (CNN). We also conducted a comparative study of these methods usingvarious performance measures. To summarize, the major contributions of the proposed work are:

- The extraction of various features from the textual data containing COVID-19 and Islamophobia related content.
- Data preprocessing for making it suitable for use for knowledge extraction using various ML and DL techniques.
- Classification of the data using various likes like ML, DL, BERT, and GPT.
- Evaluation of the performance of various classification techniques using performance metrics like Accuracy, Precision, Recall, F1-score, and AUC (area under the curve).

The rest of the paper bas been organized as follows. Section II provides the relevant details about the related work. The working of proposed approach has been described in Section III followed by its experimental evaluation in Section IV.Finally, Section V provides conclusion of the proposed work.

II. Related Work

Various studies have been conducted on Twitter datasets for detecting and removing hate speech related to COVID-19 and hate speech. Chandra et al. [15] used the coronaBias dataset containing 410,990 tweets related to COVID-19 and explored three different approaches for feature derivation and sentiment learning, including LDA, NMF, and Top2Vec. Mehmood et al. [16] worked on a dataset of 1290 tweets for hate speech detection and utilized 1D CNN with RNN for feature extraction and classification. Khan et al. [17] collected 8438 English and 8790 Hindi tweets from Twitter, and used Word2Vec, GloVe, BERT, and n-gram methods for the classification of tweets polarity classes. Alraddadi et al. [18] performed Arabic text classification using a dataset compiled using the Octoparse scrapping tool, and utilized ML algorithms such as KNN,SVM, LR, MNB, and NB for data classification. Vidgen and Yasseri [19] proposed a technique for classifying Islamophobic hate speech using KNN, SVM, LR, MNB, and CNN. To detect hate speech related to COVID-19 and Islamophobia, these studies utilized various approaches for feature extraction, classification, and sentiment learning. The datasets used in these studies were annotated and passed through various pre-processing steps such as case folding, tokenization, stop words removal, cleaning, and normalization. The ML algorithms used for classification included KNN, SVM, LR, MNB, and NBfor Arabic text classification, and DT, RF, LR, NB, SVM, and CNN for X datasets. The best results were obtained by using various combinations of ML algorithms with feature extraction methods such as GloVe, Word2Vec, n-gram and BERT methods. These studies provided fruitful results for detecting and removing hate speech from social media platforms, and their findings can be further utilized for developing efficient tools for hate speech detection.

In their work, Massey et al. [20] analyzed data from social platforms for the detection of Islamophobic content using machine learning and trend analysis approaches. The dataset, used in this work, was scraped using predetermined Islamic keywords and includes political opinions from the left, right, and center. ML techniques such NB, SVM, Boosting, MAXENT, CART, and RF were used by the researchers using 10-fold cross-validation to 400 hand-labeled comments. The accuracy of the Bagging and RF classifiers was practically identical at 0.66%, according to the data collected using multiple performance indicators, and stemming did not enhance the outcomes in this instance. In a further study, Gata and Bayhaqy [21] examined tweets concerning Islamophobia in the wake of the 2019 Christchurch assault in New Zealand. A dataset of 3115 collected tweets from March 15, 2019, the day of the incident, was used in the study. The dataset underwent various steps of preparation, including scraping, stop words elimination, and tokenization. The two ML models, NB and SVM, were combined with the random oversampling technique for result derivation and comparison. The best accuracy of 91.390% was provided by SVM with SMOTE, which was superior to other combinations. Ayan et al.'s [22] sentiment analysis of Twitter data was done to look for anti-Islamic content. From August to September 2018, the researchers gathered 162,000 tweets that had been manually positive and negative rated by professional annotators. To prepare the data for pre-processing by ML algorithms like Ridge Regression (RR) and NB, weblinks, converted letters, word-level TF-IDF, and redundancy removal were removed from the input. The Bayesian classifier took more time and had a lower accuracy of 98.1% than the RR classifier. In a study by F. González-Pizarro and S. Zannettou [23], nasty attitudes on political data from Papasavva were analyzed using contrastive learning. 134.5 million Political postings from June 2016 to November 2019 were included in the collection, coupled with a dataset of 5,859,439 photos from Zannettou. The data was pre-processed, and severe toxicity levels were calculated to identify and classify Islamophobic content. Another study [24] by Saha et al. looked at hate speech in Hindi and the rise in hate crimes in India. They made use of the 2019 HASOC dataset, which was made available to the public and included translations in English, German, and Hindi. The Gradient Boosting model, along with mBERT and LASER embeddings, was used to achieve language neutrality. Due to the unbalanced data, the model they constructed performed better on Hindi data than on English and German data.

In [25], 5,846 Lebanese and Syrian political tweets, categorized as normal, abusive, or hostile were used by Mulki et al. [25] to construct the L-HSAB dataset. An integration of

SVM and NB classifiers was used with n-gram BoW and TF-IDF vectorization methods. ML classifiers with n-gram vectorization frequently outperform neural networks for text classification. These strategies, however, are domain-specific and might not work well if the context of the information is removed or if negative remarks are given good connotations. Gitari et al. [26] provided a three-step methodology for classifying hate speech. A rule-based approach is utilized to determine the text's subject in the first stage followed bythe creation of a lexicon for hate speech. Finally, a text is deemed to be hate speech if it contains any of the three characteristics like hate verbs, negative polarity, and theme-based grammatical patterns. Despite being simple to understand, lexicon-based approaches are not totally reliable. A multi-class classifier was used by Davidson et al. [27] to distinguish between political correctness, offensive language, and hate speech. They created a precise model with L2 regularization using LR, and the results were encouraging. By merging different techniques, hybrid approaches have also been employed to detect hate speech. For the classification of hate speech, Wester et al. [28] have presented a hybrid technique that blends learning and lexical-based methods. Using a lexicon-based technique, complicated syntactic and semantic aspects are extracted using this method, and a learning algorithm is then used. In comparison to the distinct lexical and learning approaches, the hybrid model has performed better. Although, the work in [28] address hate speech but a relevant problem that needs is Interest in the detection of Islamophobic textual content has increased because of the rise is: Islamophobic occurrences during COVID-19. However, because to the dearth of publicly accessible datasets and the sparse application of numerous textual features and transformer-based core NLP approaches, there has been little research in this field. There is a huge research gap because of the majority of studies concentrating on either traditional textual features or word embeddings with ML and DL models. For this, in this work, we attempt to overcome these research issues and construct an efficient model for accurate Islamophobic content identification in the proposed work.

III. Proposed Work

This section will delve into the detailed discussion of the proposed framework along with justification of adopted methods.

A. Proposed Fremwork

Here, the presented framework will be demonstrated and discussed in detail. Fig. 1 provides a compact overview of proposed model.

For classifying Islamophobia related social media content,first, datasets were collected from X and Google fact-checking API. This step is followed by data preprocessing using various techniques such as stop words removal, data balancing, lemmatization, and tokenization. Additionally, for feature extraction through word embeddings and n-grams,different methods like Word2Vec, GloVe, TF-IDF, and BoW are used. For classification, transformer-based techniques BERT and GPT, and topic modelingare utilized. After that, the classification is performed using some selected conventional ML, DL, and transformed-based techniques. Performance evaluation measures Accuracy, Precision, Recall, F-Measure, and AUC are recorded evaluating the performance of these classifiers.

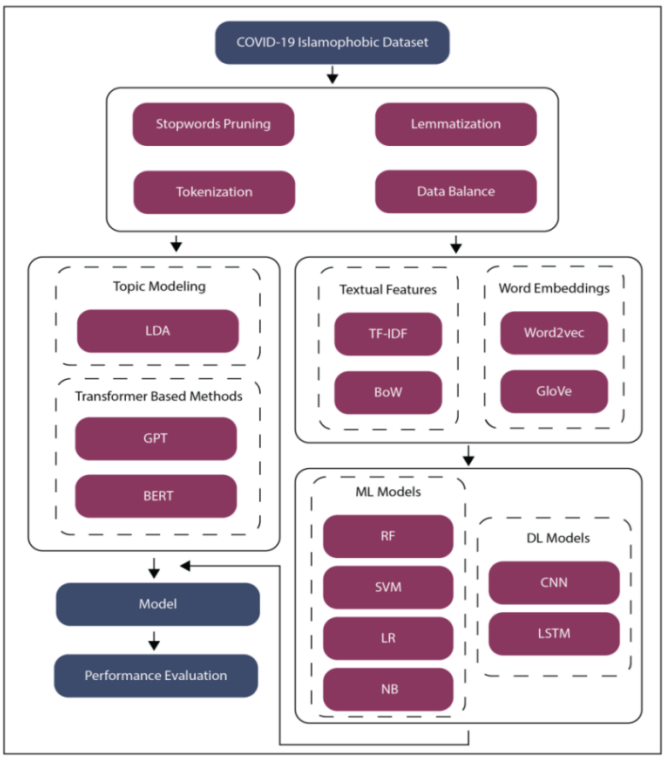


Fig. 1. Proposed model framework.

B. Feature Extraction

This phase involves extracting features that are particularly useful in experiments for achieving desirable outcomes. We present the detail of this phase in the following.

1) TF-IDF: TF-IDFcalculates the frequency of words in each document by taking into the account the inverse frequency of such words appearing in multiple documents consistently [29]. The weight of each document in the corpus can be computed using Eq. (1):

$$wgt_{d,c} = freq^{t}_{d,c} x\ log\ log(\frac{N}{freq_d} \quad (1)$$

Where, $wgt_{d,c}$ is denoted as the total weightage of both data points, $freq^{t}_{d,c}$ computes the frequency of occurrences of the data point d in c. N represents total number of documents in the corpus, $log\ log(\frac{N}{freq_d})$ computes the log of all the documents present in the corpus with the frequency of data point d.

2) BoW: To extract useful features from textual data for classification purposes, the Bag of Words (BoW) [30] method is employed. This approach considers a document or phrase as a set of its constituent words and checks for the presence of familiar words irrespective of their order. BoW generates word bags using Eq. (2), as follows:

$$doc_c = \sum_{d=1}^{N} \quad weight^{c}_{d} x weight_d \quad (2)$$

Where doc_c denotes the documents that contain the concerned data point d. $weight_d^c$ are the scalar weights of the frequent word d for the data point c in the document. While $weight_d$ indicates the weight of frequent word d.

3) Word2vec: The Word2Vec approach uses a three-layer deep neural network to analyze the context of a document and connect related phrases. Unlike BoW, Word2Vec offers two models - Continuous Bag of Words (CBoW) and Skip-Gram [31]. To ensure proper word embedding, it is recommended that Word2Vec is trained on a large and high-quality dataset. The computation of Word2Vec through the Skip-Gram method for an M-dimensional data corpus contain a word wd_o at location q can be seen in Eq. (3).

$$\frac{1}{T}\sum_{q=1}^{M} \sum_{-s\le a\le s, a\neq 0} \log \text{logprob}(\text{wd}_{q+1}|\text{wd}_o) \quad (3)$$

Where $log\ logprob(wd_{q+1}|wd_o)$ denotes the logarithm of wd_o with respect to placements and co-occurrences within the document.

4) GIoVE: To perform unsupervised learning, GloVe generates word embedding by constructing a count-based matrix based on word co-occurrence and analyzing each term individually [32]. It uses a less-weight approach to produce factors and creates a lower-dimensional matrix. The entire working logic of GloVe can be seen in Eq. (4).

$$\text{h} = \sum_{c,d=1}^{m} \text{g}(\text{f}_{c,d})(\text{d}_b^t \text{d}_a - \log \text{logf}_{c,d})^2 \quad (4)$$

5) Transformer-based models: Language models in NLP are built on transformers, which consist of an Encoder and Decoder. In this work, we used two transformer-based models GPT and BERT. GPT is an autoregressive decoders model, and it has two versions: GPT-2 and GPT-3 while BERT is a variant of bidirectional encoders-based models, and there are several types of BERT models. BERT uses Mask Language Modeling to overcomes the unidirectional constraint by utilizing and an attention mechanism and utilizes an encoder, decoder, and various layers. On the other hand, GPT is an autoregressive decoder model that works to benefit from unlabeled text datasets for using them on limited supervised datasets. GPT has two variants: GPT2 and GPT3, with the latter having 175 billion parameters and the capability to perform several NLP tasks such as text classification, question answering, text generation, and named entity recognition. Overall, these transformer-based models have revolutionized NLP tasks and continue to provide state-of-the-art performance.

C. Topic Modeling

To identify topics from a collection of documents, topic modeling is a useful technique. LDA is an effective approach for text classification in which the text of a document is classified based on its relation to a particular topic. The fundamental principle of LDA's functioning is demonstrated by Eq. (5).

$$\text{p}(\text{s}, \text{k}) = \text{q}(\text{k}|\text{d}) * \text{q}(\text{s}|\text{k}) \quad (5)$$

Where $q(k, d)$ is the probability of the topic per document and $q(s, k)$ is the probability of words per topic equaling the $p(s, k)$ denoted as the probability of word with the topic.

IV. Experiments and Results

To evaluate the performance of the proposed scheme, the experiments were conducted in systematic manner. In the first set of experiments, the n-gram method was used to extract features and classify the data using four ML algorithms. In the second set of experiments, word embedding features were classified using deep LSTM and CNN models. Next, LDA was applied to the data for topic modeling, and the classification step was performed. Finally, core NLP transformer-based methods, namely BERT and GPT, were evaluated. The dataset wqs balanced before conducting experiments.

A. Datasets

To investigate the global prevalence and impact of Islamophobia, we collected two distinct datasets. The first was obtained from the Google Fact Check platform and consisted of news articles that were fact-checked by websites such as PolitiFact and Snopes. We extracted articles relevant to Islam, including those with terms like Islam, Muslims, Quran, Jihad, and women, resulting in a total of 1555 articles. The second dataset was sourced from Twitter and included posts from users worldwide. We used predetermined hashtags, including #fuckIslam, #Jihadi, #Coronajihad, #Tablighijamat, and #TablighiJamaatVirus, as well as lexicons from Hatebase, to collect tweets from January 2020 to August 2020. The dataset is diverse as it retrieves data using an unbiased mechanism. The English-language dataset consists of 9612 tweets and was pre-annotated by three English-proficient annotators. During the annotation process, the annotators were not provided with any information about users' identities. The annotators were tasked with categorizing each tweet into one of three categories: Islamophobic, related to Islam but not Islamophobic, or neither about Islam nor Islamophobic. The annotations were assigned with great care, and in cases of disagreement, a majority vote was utilized. Of the 2930 tweets marked as Islamophobic, 4336 were related to Islam but not Islamophobic, and 2346 were neither Islamophobic nor related to Islam.

B. Dataset Preprocessing and Balancing

In the proposed work, various pre-processing techniques were applied, including converting all letters to lowercase, removing stop words and hyperlinks, and half-sentences. It also involves lemmatization, and tokenization. For data balancing, it is made sure in this phase that the balanced data is used for experiments and result analysis.

After performing pre-processing and balancing the dataset, the vocabulary size was determined for the English data. The vocabulary size for unigrams was found to be 17861 with an average tweet length of 14 words. After pre-processing the data to contain 8 words per tweet, the vocabulary size decreased to 16580 unigrams. Table I presents some of the most frequent words extracted from the dataset as part of feature extraction.

TABLE I. WORDS FRQUENCY IN DATASET AFTER PRE-PROCESSING

Sr. No	Words
1	Muslim
2	Islam
3	Islamic
4	Quran
5	Pakistan
6	Allah
7	Radical
8	Jehadi
9	Mohammed
10	Hindu

It is important to mention here that the utilized dataset had an imbalanced distribution, which was addressed using under-sampling. Tokenization and lemmatization were then applied to the pre-processed dataset, and during tokenization, both unigrams and bigrams were used, and the LDA model was employed to identify the best topics, which were visualized using an Intertypic Distance Map. The top 20 phrases from the first topic are visualized in Fig. 2 which account for 12.6% of the tokens. The bigram themes in the bar chart are distinguished from one another using an underscore. The use of these techniques can help identify the most pertinent phrases and topics in large datasets, making it easier to analyze and understand the data.

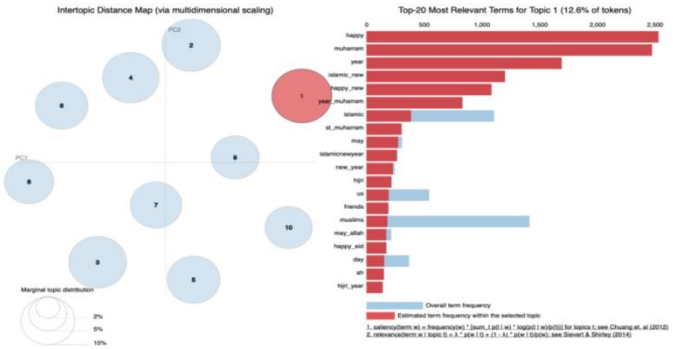


Fig. 2. Visualizing the most salient terms of first topic using topic modeling.

C. Machine Learning Algorithm with Textual Features

In initial experiment, we evaluated the n-grams based features using SVM. Tables II and III present the performance of various ML classifiers with BoW and TF-IDF, respectively, while maintaining the performance standards mentioned earlier. In the proposed work, SVM with n-gram based textual feature extraction techniques were applied to the categorical Islamophobia data in Python language. A train/test ration of 90/10 was used. The SVM model combined with the BoW method achieved a slightly higher accuracy of 91.7% compared to other counter parts.

In the next experiment, the RF classifier is used to detect Islamophobic content based on the same n-gram features as before. RF-BoW achieves significantly higher accuracy than RF-TF-IDF. The following experiment uses LR classifier for categorical data classification. LR-BoW outperforms LR-TF-IDF. In the last experiment, GNB is used for classification with the same features as before. TF-IDF was observed to achieve better results than BoW.

TABLE II. RESULTS OF VARIOUS ML MODELS WHILE USING TF-IDF

TF – IDF			
Algorithm	**Accuracy (%)**	**F1 Score (%)**	**AUC (%)**
RF	86.7	87.0	97.3
SVM	90.5	91.0	97.8
LR	90.3	90.0	97.8
NB	86.9	87.0	90.2

TABLE III. RESULTS OF ML MODELS WITH BOW

BoW			
Algorithm	**Accuracy (%)**	**F1 Score (%)**	**AUC (%)**
RF	87.6	88.0	97.0
SVM	91.7	92.0	98.0
LR	91.6	92.0	98.5
NB	77.4	77.0	82.6

D. Word Embeddings with Deep Learning Algorithms

We examined the performance of four ML models that used derived n-gram features and then explored the effectiveness of DL models with word embeddings as input. We experimented with a customized CNN, which is a type of deep neural network designed for rapid classification of vectorial data, using features extracted from the GloVe and Word2Vec word embedding models. We trained and tested the CNN model using the same data split as the ML algorithms, first with Word2Vec features using 32 epochs and a batch size of 10 and then with GloVe features using 100 epochs and a batch size of 32. For validation, the batch size remained the same while the number of epochs was set to 5. The results of both embedding models with CNN showed that CNN performs marginally better with GloVe than Word2Vec, exhibiting better accuracy and evaluation rates. The next experiment involves LSTM, which uses a batch size of 10, 20 epochs, and essential layers, including embedding, dense, and SoftMax layers. The accuracy of the LSTM model improves over time for both GloVe and Word2Vec features with a decrease in the loss ratio as the number of epochs increases. It was observed that the results of Word2Vec Features with LSTM were better than results with GloVe with Accuracy =88.6%, Precision, recall, and F1-score=89%, and AUC=97.2%.

The results indicate that Word2Vec embeddings provide better representation of the text data for the LSTM model. These findings are consistent with previous research that suggests that the choice of word embeddings can significantly impact the performance of deep learning models in natural language processing tasks. Therefore, selecting the appropriate word embeddings is crucial for the effectiveness of the model.

In another set of experiments, we test various machine learning algorithm with topic modeling which includes the experiments conducted using the LDA algorithm and ML

models. The derived topics are then scaled using a standard scalar before being classified. The selection of these topics is based on their grammar weightage, which helps identify the most relevant and significant terms for each topic.

Next experiment involved extracting unigrams and bigrams from the pre-processed dataset, which were then used as input for LDA. The LDA algorithm generated extracted topics after fine-tuning of the model. Again, four ML classifiers were then evaluated on the selected topics using the same split of 90/10 for training and testing sets respectively. The results of this experiment have been presented in Fig. 3.

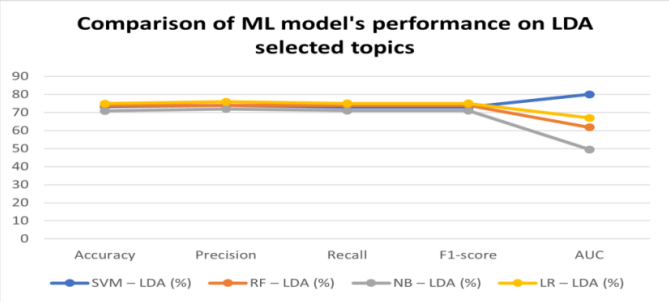


Fig. 3. Comparison of ML model's performance with LDA selected topics.

E. Using Transformer-based Techniques for Classification

As mentioned earlier, we also investigated the use of two popular transformer-based models, BERT and GPT, for the NLP task. BERT model is fed with the pre-processed dataset as input, which is then encoded into an embedding representation. After performing several transformations on the embeddings, the representations are decoded back into vocabulary-based representations. From the results, it was observed that GPT outperforms BERT, achieving an accuracy of 91.6% compared to BERT accuracy of 89%%. Similarly, the precision, recall, and f1-score values also show a similar trend, where GPT outperforms BERT. These results indicate that GPT is more effective in extracting contextual features and capturing the nuances of the text data for classification tasks.

The performance of the transformer-based NLP models, BERT and GPT, was also evaluated. The pre-processed dataset is fed into BERT and classification results are recorded. Fig. 4 shows the results of this experiment. The BERT model outperforms other models with a significantly higher accuracy of 89.31%. The results of this experiment demonstrate the effectiveness of BERT in text classification tasks and highlight its potential as a powerful NLP tool.

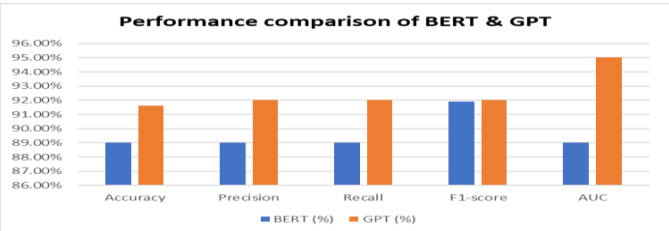


Fig. 4. Comparison of BERT and GPT based on various performance measures.

In the final part of the experiment, GPT-2 was used for text classification. The results of this experiment indicated that the GPT-2 model achieved an accuracy of 91.6%, which is significantly higher than the accuracies by other models.

F. Comparison of Results among Applied Techniques

From the results reported in the previous section, we got the motivation for comparing the results of various techniques. The results of the first experiment showed that BoW outperformed other techniques, particularly when TF-IDF is used with GNB models. These results suggest that BoW-based features are better suited for use during classification. During the classification, SVM showed the best performance achieving an accuracy of 90.7%. We believe this is because SVM is able to get good parameter settings without parameter tuning. Next, two DL models, LSTM and CNN, were compared using GloVe and Word2Vec. The results indicated that the custom CNN model outperformed LSTM in when evaluated against various performance metrics. This experiment highlights the importance of selecting the appropriate DL model and word embedding for achieving optimal performance in natural language processing tasks. It is worth noting that the performance of the ML and DL models can vary depending on the specific task and dataset. Therefore, it is crucial to conduct comprehensive experiments and compare the results before selecting the optimal model for a particular task. Fig. 5 shows the comparison of performance of CNN while using Word2vec and GloVe.

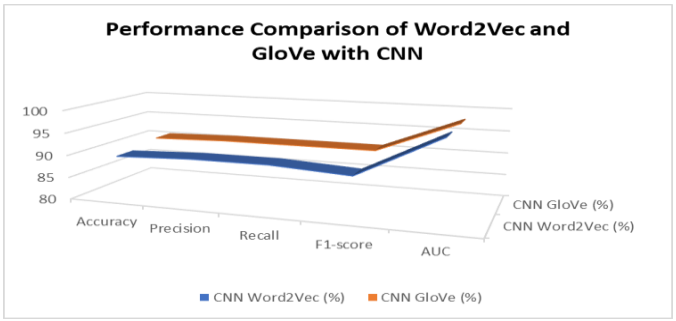


Fig. 5. Comparison of performance of Word2Vec and GloVe while using CNN.

In the second experiment, the LSTM model was used to classify Islamophobic content using the same word embedding models, Word2Vec and GloVe. The results showed that the LSTM-Word2Vec model had decent performance when compared to GloVe. In contrast to the CNN model, where the combination of CNN and Word2Vec performed better, the LSTM model with Word2Vec had better results. This comparison is also presented in Fig. 6. The comparison of different models and feature extraction techniques is important to determine the best approach for a given task. In this study, it was found that BoW-based features performed better than TF-IDF-based features when used with ML models, while CNN-GloVe outperformed LSTM-Word2Vec in DL models for classifying Islamophobic content. These findings can be useful in future studies and real-world applications for detecting and addressing hate speech online.

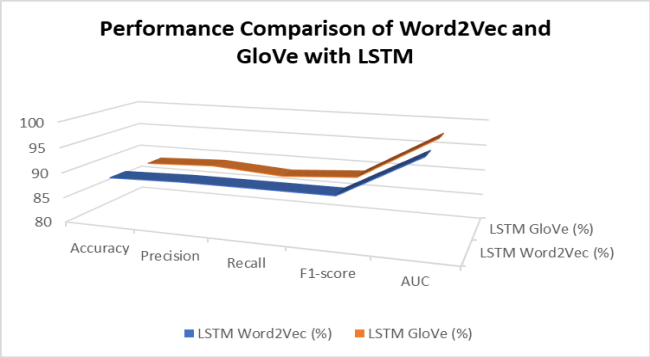


Fig. 6. Comparison of performance of Word2Vec and GloVe while using LSTM.

The second DL model, CNN, achieved the highest accuracy of 90.6% in the second experiment. LDA topic modeling combined with ML classifiers showed that LR and RF performed best among the classifiers. LR achieved the highest accuracy of 74.9%. The last experiment includes transformer-based models BERT and GPT, with BERT achieving an accuracy of 89.31%, which is in between the maximum accuracies of ML and DL models. Fig. 7 shows the comparison of BERT results with ML's best performing algorithm: SVM, DL models, and LDA's best performing model LR.

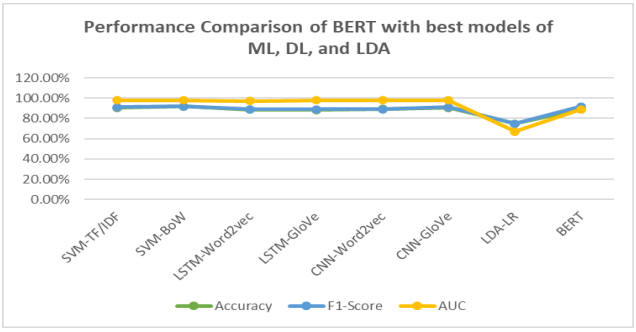


Fig. 7. Performance comparison of BERT with other classifiers.

In the next phase of the experiment, GPT-2 was utilized for the classification of Islamophobic content. The results indicate that GPT-2 outperformed all the previous methods used in this study. The F1 score achieved by GPT-2 was 92%, which is significantly higher than the other models. The comparison of GPT-2 results with other best-performing models can be visualized in Fig. 8. These findings highlight the effectiveness of GPT-2 in the classification of Islamophobic content and suggest that it could be used in similar tasks.

G. Comparison of Proposed Technique with Existing Islamophic Classification Techniques

In this section, we compare the results of the proposed study with those of prior art. It should be noted that previous studies did not use both textual features and word embeddings to test the performance of classification models. Nevertheless, we have compared their results with those achieved in the proposed study. Table IV presents a comparison of the F1 score results obtained by previous studies and the proposed study for Islamophobic content detection. It is evident from this table that the proposed study achieved better results than the previous studies, indicating that utilizing both textual features and word embeddings is an effective approach for improving the performance of classification models in this domain.

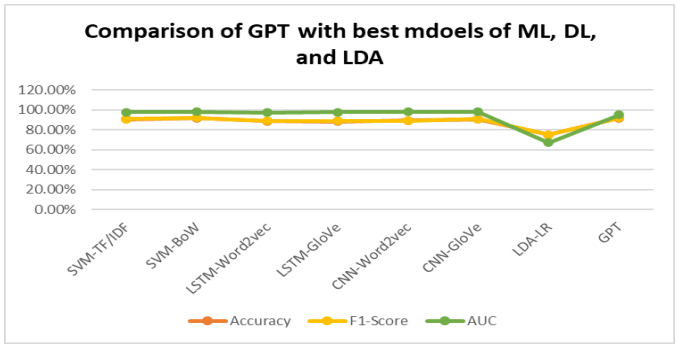


Fig. 8. Comparison of GPT with other classifiers.

TABLE IV. PERFORMANCE COMPARISON OF PROPOSED SCHEME WITH PRIOR ART

Ref	Algorithm	Results
Chandra et al. [15]	BERT	F1 score = 88.0
Mehmood et al. [16]	CNN	F1 score = 90.1
Alraddadi et al. [18]	NB	F1 score = 89.0
Vidgen et al. [19]	SVM	F1 score = 77.3
Massey et al. [20]	RF	F1 score = 66.0
Proposed Model	CNN, RF-LDA BERT, RF, SVM, GPT	CNN – F1 score = 91.0 RF – F1 score = 88.0 BERT – F1 score = 91.9 SVM – F1 score = 92.0 GPT – F1 score = 92.0

Table IV demonstrates that the proposed study outperformed the earlier investigations, even though the earlier studies did not make use of various Transformer technique variations or DL techniques with various word embeddings.

The main limitations of the proposed work, as noted by the findings of the experimental evaluation, are that more data is required to get better result that needs to be improved and can be a potential research area. Similarly, knowledge extraction from other hate speech related content related to any pandemic and its potential impact on the society can also to be investigated by extension of the proposed work.

V. CONCLUSION

Globally, there has been a substantial increase in content that is anti-Islamic because of the COVID-19 pandemic. Rapidly proliferating erroneous information and narratives have influenced negative attitudes and actions. Automated methods based on data science and AI, however, have emerged as useful resources for identifying and classifying racist content, enabling the detection and avoidance of damaging narratives. The proposed classifier performance evaluation in this study extracted significant features from the data using processes like Word2Vec, GloVe, etc. In addition, important themes were identified using topic modeling using LDA. Several ML and DL methods, such as LSTM and CNN with word embeddings and transformer-based models like BERT and GPT, were tested in this study. With an F1 score of 92%,

GPT was found to be the model that performed the best. Future studies might concentrate on employing various GPT iterations, such as GPT-3, and investigating additional DL models, such as RNN and GANs. Additionally, expanding the dataset can enhance the precision of the findings. Society may lessen Islamophobia and foster greater acceptance and tolerance by using these tools. To find bad content, stop it from spreading, and encourage a more open and tolerant society, it is essential to keep developing and improving automated tools.

ACKNOWLEDGMENT

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under Grant No. (UJ-22-DR-82). The authors, therefore, acknowledge with thanks the University of Jeddah technical and financial support.

REFERENCES

[1] L. Cervi, S. Tejedor, and M. Gracia, ‘What Kind of Islamophobia? Representation of Muslims and Islam in Italian and Spanish Media’, Religions, vol. 12, no. 6, p. 427, 2021.

[2] R. A. Alraddadi and M. I. E.-K. Ghembaza, “Anti-Islamic Arabic Text Categorization using Text Mining and Sentiment Analysis Techniques,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 8, 2021.

[3] Moreno-Vallejo, Patricio Xavier, Gisel Katerine Bastidas-Guacho, Patricio Rene Moreno-Costales, and Jefferson Jose Chariguaman-Cuji. "Fake News Classification Web Service for Spanish News by using Artificial Neural Networks," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 3, pp. 301–306, 2023.

[4] BAYRAKLI, Enes, and Farid HAFEZ. "European Islamophobia Report (EIR) 2022." Medya ve Din AraştırmalarıDergisi 6, no. 1, pp. 221-225, 2023.

[5] J. Qian, ‘Historical Ethnic Conflicts and the Rise of Islamophobia in Modern China’, Ethnopolitics, pp. 1–26, 2021.

[6] Sukabdi, Zora Arfina, Muhammad Adlin Sila, Chandra Yudistira Purnama, FathulLubabinNuqul, Seta AriawuriWicaksana, Ali Abdullah Wibisono, and Yanwar Arief. "Islamophobia Among Muslims in Indonesia." Cogent Social Sciences, vol. 9, no. 1pp. pp. 1-29, 2023.

[7] Riaz, Marwa, Khadija Shahbaz, and Maryam Ali. "Islamophobia in the US and Europe: An Analytical Study." Annals of Human and Social Sciences, vol. 4, no. 2, pp. 615-625, 2023.

[8] T. Mirrlees and T. Ibaid, ‘The Virtual Killing of Muslims: Digital War Games, Islamophobia, and the Global War on Terror’, Islam. Stud. J., vol. 6, no. 1, pp. 33–51, 2021.

[9] Ahuja, K.K. and Banerjee, D.. The “labeled” side of COVID-19 in India: Psychosocial perspectives on Islamophobia during the pandemic. Frontiers in Psychiatry, vol. 11, p.604949, 2021.

[10] Rajan, B. and Venkatraman, S., Insta-hate: An Exploration of Islamophobia and Right-wing Nationalism on Instagram Amidst the COVID-19 Pandemic in India. Journal of Arab & Muslim Media Research, vol. 14, no.1, pp.71-91, 2021.

[11] Alghamdi, Jawaher, Yuqing Lin, and Suhuai Luo. "A Comparative Study of Machine Learning and Deep Learning Techniques for Fake News Detection." Information, vol. 13, no. 12, pp. 1-28, 2022.

[12] Ouassil, Mohamed-Amine, Bouchaib Cherradi, Soufiane Hamida, Mouaad Errami, Oussama EL Gannour, and Abdelhadi Raihani. "A Fake News Detection System based on Combination of Word Embedded Techniques and Hybrid Deep Learning Model." International Journal of Advanced Computer Science and Applications (IJACSA), vol. 13, no. 10, pp. 525–534, 2022.

[13] Kamran, Muhammad, and Ahmed Abdul-Aziz Almaghthawi. "Case-Based Reasoning Diagnostic System for Antenatal Research Database." International Journal of Online & Biomedical Engineering, vol. 18, no. 7, pp. 176-187, 2022.

[14] Alam, Furqan, Ahmed Almaghthawi, Iyad Katib, AiiadAlbeshri, and Rashid Mehmood. "IResponse: An AI and IoT-enabled Framework for Autonomous COVID-19 Pandemic Management." Sustainability, vol. 13, no. 7, pp. 3797, 2021.

[15] Chandra, Mohit, Manvith Reddy, Shradha Sehgal, Saurabh Gupta, Arun Balaji Buduru, and Ponnurangam Kumaraguru. "" A Virus Has No Religion": Analyzing Islamophobia on Twitter During the COVID-19 Outbreak." In Proceedings of the 32nd ACM conference on hypertext and social media, pp. 67-77, 2021.

[16] Mehmood, Qasim, Anum Kaleem, and Imran Siddiqi. "Islamophobic Hate Speech Detection from Electronic Media Using Deep Learning." In Mediterranean conference on pattern recognition and artificial intelligence, pp. 187-200, 2021.

[17] Khan, Heena, and Joshua L. Phillips. "Language agnostic model: Detecting Islamophobic content on social media." In Proceedings of the 2021 ACM Southeast conference, pp. 229-233. 2021.

[18] R. A. Alraddadi and M. I. E.-K. Ghembaza, ‘Anti-Islamic Arabic Text Categorization using Text Mining and Sentiment Analysis Techniques’, Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 8, 2021.

[19] B. Vidgen and T. Yasseri, ‘Detecting weak and strong Islamophobic hate speech on social media’, J. Inf. Technol. Polit., vol. 17, no. 1, pp. 66–78, 2020.

[20] T. Massey, C. Amrit, and G. C. van Capelleveen, ‘Analysing the trend of Islamophobia in Blog Communities using Machine Learning and Trend Analysis’, presented at the 28th European Conference on Information Systems, ECIS 2020: Liberty, Equality, and Fraternity in a Digitizing World, pp.1–14, 2020.

[21] W. Gata and A. Bayhaqy, ‘Analysis sentiment about islamophobia when Christchurch attack on social media’, TELKOMNIKA Telecommun. Comput. Electron. Control, vol. 18, no. 4, pp. 1819–1827, 2020.

[22] B. Ayan, B. Kuyumcu, and B. Ciylan, ‘Detection of Islamophobic Tweets on Twitter Using Sentiment Analysis’, Gazi Univ. J. Sci. Part C, vol. 7, no. 2, pp. 495–502, 2019.

[23] F. González-Pizarro and S. Zannettou, ‘Understanding and Detecting Hateful Content using Contrastive Learning’, ArXivPrepr. ArXiv220108387, 2022.

[24] P. Saha, B. Mathew, P. Goyal, and A. Mukherjee, ‘Hatemonitors: Language agnostic abuse detection in social media’, ArXivPrepr. ArXiv190912642, 2019.

[25] Mulki, Hala, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. "L-hsab: A levantine twitter dataset for hate speech and abusive language." In Proceedings of the third workshop on abusive language online, pp. 111-118, 2019.

[26] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, ‘A lexicon-based approach for hate speech detection’, Int. J. Multimed. Ubiquitous Eng., vol. 10, no. 4, pp. 215–230, 2015.

[27] Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. "Automated hate speech detection and the problem of offensive language." In Proceedings of the international AAAI conference on web and social media, vol. 11, no. 1, pp. 512-515, 2017.

[28] Wester, Aksel, Lilja Øvrelid, Erik Velldal, and Hugo Lewi Hammer. "Threat detection in online discussions." In Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 66-71. 2016.

[29] Kumar, Vipin, and Basant Subba. "A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus." In 2020 national conference on communications (NCC), pp. 1-6. IEEE, 2020.

[30] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, ‘The influence of preprocessing on text classification using a bag-of-words representation’, PloS One, vol. 15, no. 5, p. e0232525, 2020.

[31] K. W. Church, ‘Word2Vec’, Nat. Lang. Eng., vol. 23, no. 1, pp. 155–162, 2017.

[32] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543.

A Neural Network-based Approach for Apple Leaf Disease Detection in Smart Agriculture Application

Shengjie Gan*, Defeng Zhou, Yuan Cui, Jing Lv
Hunan Chemical Vocational Technology College
Zhuzhou 412004, Hunan, China

***Abstract*—Plant diseases significantly harm agriculture, which has an impact on nations' economies and levels of food security. Early plant disease detection is essential in smart agriculture. For the diagnosis of plant diseases, a number of methods, including imaging, have been used recently. Some of the existing methods for plant disease detection using imaging have limitations as firstly, high computational cost, some methods require complex image processing algorithms or manual design of features that can increase the time and resources needed for the detection. Secondly, low accuracy, most of the methods rely on simple classifiers or handcrafted features that may not capture the subtle differences between different diseases or healthy leaves. Thirdly, dependency on expert knowledge, some methods need human intervention or prior knowledge of the diseases and pests to perform the detection. These limitations are not suitable for the problem at hand because they can affect the efficiency of the detection system. In this study, three apple tree leaf diseases—apple black spot, Alternaria, and Minoz blight—are detected using a neural network (NN) and a digital image processing technique. The sample images are prepared, processed, and used to extract attributes using a digital image processing approach, and the NN is used to classify the diseases. An evaluation of the proposed system's performance in identifying illnesses in apple trees shows satisfactory accuracy and strong overall performance. Additionally, when compared to other techniques already in use, this strategy is more effective at diagnosing.**

***Keywords*—Smart agriculture; plant disease; apple leaf disease; image processing; neural network**

I. Introduction

According to the statistics of FAO, America has managed to earn 1.1 billion dollars by producing 4.08 million tons of apples on average, in addition to meeting the domestic demand to be the largest exporter of this product worldwide [1-3]. Principally, the price and balanced market of agricultural products depend on the quality. Plant diseases can significantly reduce product quality and quantity, leading to a decline in the economies of nations that depend on exporting agricultural products [4]. If identified in the earliest stages of development, these disorders may, in some cases, be prevented and controlled. In this regard, some countries are looking for ways to diagnose plant diseases early [5].

Among the most common solutions in the past has been the visual diagnosis of the disease by experienced experts, which, in large farms, typically costs a lot and calls for ongoing plant pathologist monitoring [6]. However, many diseases don't have obvious symptoms in the first stages of the disease, and it is difficult to identify them with the naked eye. For this reason, it is necessary to investigate and provide a fast, automatic, low-cost, and precise tool for identifying plant diseases. In recent decades, the growth of technology has increased progress in various branches of science, industry, and agriculture; therefore, various researchers have sought to use new technologies for the early and timely diagnosis of various plant diseases [7]. The technologies used are mainly in two parts, the type of sample processing system and classification models. Among the different sample processing systems, we can mention the types of imaging systems, the smelling machine, and the tasting machine. The different types of statistical models, data mining, and artificial intelligence can also be mentioned from the different classification models [8]. Usually, sample processing systems are selected according to the type and symptoms of plant diseases [2].

The imaging system has attracted much attention recently due to its advantages, such as non-destructiveness, reduction of human resources costs, and high accuracy. Different researchers of this system diagnose palm leaf nutritional disease, classify minnow, black spot, and Alternaria (apple tree leaf) diseases, and diagnose tea leaf disease [9]. The diseases are internal powdery mildew, bacterial angular spot, ring spot, spot, gray rot, anthracnose, and powdery mildew [10]. Since no signs of them can be seen in the beginning and spread of the disease, more advanced systems than usual imaging should be used to extract the characteristics of the disease. Therefore, the researchers considered the hyperspectral imaging system because this system, in addition to the spatial features related to the disease, also extracts the spectral features, which are very useful in the early diagnosis of the disease. On the other hand, this system is costly, time-consuming, and requires trained people, which may only be possible for some farmers to use. The odor machine is another type of sample processing system that has advantages such as non-destructiveness, high reliability, and easy and fast use, and it is not effective in cases where there are no signs of discoloration of the disease in the samples [11].

Researchers have also developed more accurate classification models by studying the recent research done by different researchers to diagnose plant diseases, which can be advanced image processing systems [12]. In this regard, the classification model can be used in a study to diagnose and classify diseases (mosaic, leaf rust, and round spot) of apple tree leaves as well as diseases (brown spot, grey spot, and round spot) of corn.

This study proposes a method to identify and categorize three apple tree leaf diseases using digital image processing and a neural network (apple black spot, Alternaria, and Minoz blight). The proposed method aims to overcome these limitations by using a neural network and a digital image processing technique that can extract features and classify diseases with high accuracy and low complexity.

The main contributions of this study are as follows:

1) The study, which focuses on the disease of apple tree leaves, provides a digital image processing technique specifically made for identifying plant diseases. The preparation, processing, and extraction of pertinent properties from sample images made possible by this technique enable precise and effective disease identification.

2) This study investigated a neural network (NN) to classify diseases and trained it using information gathered from digital image processing technology. This strategy shows how NNs are effective at correctly classifying and diagnosing diseases of apple trees, advancing automated and trustworthy plant disease diagnosis systems.

The rest of this paper structure as follows, Section II reviews the related works. Section III discuss about proposed method. Section IV presents results and discussion. Finally, this paper concludes in Section V.

II. Related Works

In study [13], a method presented for detecting apple leaf disease. This approach is according to technologies for pattern recognition and image processing. The input RGB (Blue, Green, and Red) picture was first given a color transformation structure, and the RGB model was then transformed into the Hue, Saturation, and Intensity (HSI), YUV, and grey models. Following removing behind-the-scenes, the sickness spot image was segmented using a region-growing algorithm (RGA) and a predefined threshold value. Each spot picture had thirty-eight color, texture, and form classification characteristics extracted. Combining (GA) with a genetic algorithm allows for correlation-based feature selection of the most critical factors, enhancing the apple leaf disease detection accuracy while minimizing the feature space's complexity (CFS).

An automated disease leaf recognition approach is proposed for identifying and grading leaf diseases using machine vision and digital image processing [14, 15]. The proposed system is broken up into two phases. The plant is recognized in the first stage according to the characteristics of its leaves. This phase involves pre-processing leaf image data and feature extraction, followed by training and classification using an artificial neural network to identify leaf features [16]. In the second phase, the disease that affects the leaf is categorized, which entails segmenting the defective area using k-means, feature extraction from the wrong area, and disease classification using an artificial neural network (ANN). The degree of disease present in the leaf is then considered when assigning a disease grade.

A prediction model for plant leaf disease detection and classification utilizing computer vision and machine learning approaches was proposed in [17]. Pre-processing, segmentation, and extraction of properties, including shape, color, texture, vein, and so on, are done on the raw picture of a leaf. Several machine-learning classifiers are used to categorize the leaf image. The experimental outcomes are assessed and contrasted with those of K-Nearest Neighbor, Random Forest, Support Vector Machine, and Artificial Neural Network.

For precisely identifying plant leaf disease, the Boosted support vector machine-based Arithmetic optimization algorithm (BSVM-AOA) has been presented [18]. In this instance, the greyscale co-occurrence matrix is employed for feature extraction, and the vector value active contour model is used for picture segmentation. Furthermore, performance indicators, such as f-rating, recall, accuracy, and specificity, are used to gauge how well the proposed technique performs. The proposed method is compared to the various existing processes in a comparative analysis. Their findings revealed that the BSVM-AOA technique had a 98.6%.

III. Proposed Method

In this study, to diagnose apple tree leaf diseases, a method consisting of an image processing method and a neural network model has been developed. Fig. 1 depicts the proposed approach.

A. Pre-processing

1) Dataset collection: This research investigated three apple tree leaf diseases: Apple black spot, Alternaria disease, and minnow pest. Some sample images are collections inspired by [19]. Six hundred forty different samples of apple tree leaves, including data augmentation described in the next section, the dataset contains 320 leaves infected with Alternaria disease, 184 infected with apple black spot disease, and 136 infected with Minoz pest.

2) Data augmentation: Preparing images of leaf samples to prepare pictures of leaves and produce sample images using data augmentation. In this data augmentation, filters such as noising, blurring, contrast, and brightness are added to the sample images to extend the dataset [19]. Fig. 2 displays examples of the dataset's photos.

3) Image conversion: This step is image conversion RGB channel to L*a*b* (Lab*) channel. The Lab* is a color space that is based on the opponent color model of human vision, where red and green form an opponent pair and blue and yellow form an opponent pair. It expresses color as three values: L* for perceptual lightness and a* and b* for the four unique colors of human vision: red, green, blue and yellow.

RGB to Python programming language and the Open CV library are employed for image processing and system implementation [20]. Image processing steps are composed of removing the background, removing leaf tails, removing unwanted regions such as noises, and the RGB channel to the L*a*b* channel conversion of the picture due to the proximity of this channel to the human visual system and removing the L component to eliminate the effect of brightness.

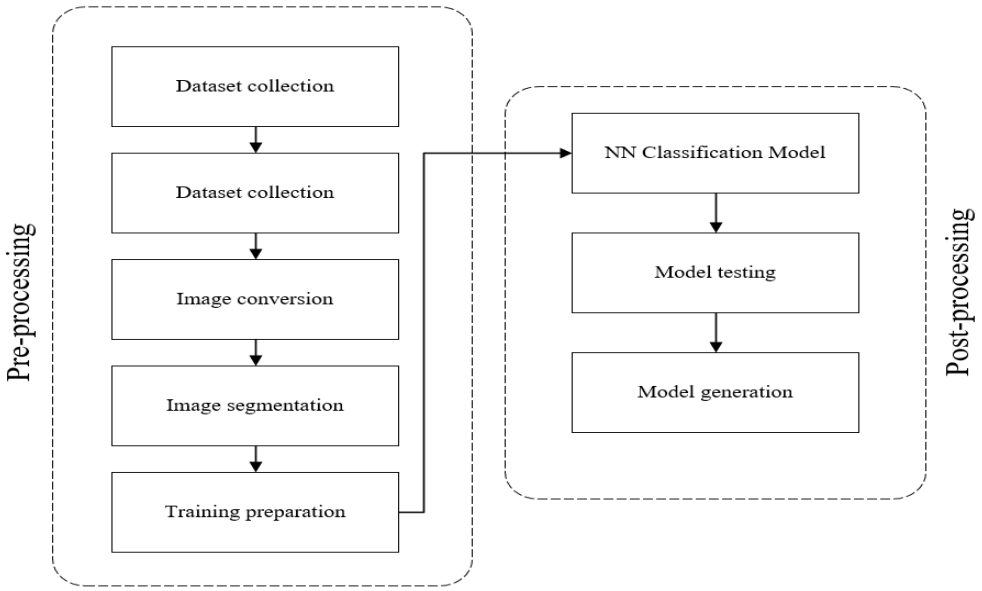


Fig. 1. The proposed method.

Fig. 2. Sample images in the dataset.

4) Images segmentation: Segmentation was done based on the area by extracting the contaminated area using the k-means technique. Image clustering commands were applied to only *a and *b components using the k-means method, and the images were divided into healthy and infected leaf areas. After this stage, because the disease spot had a smaller surface than the leaf surface, this area was selected. The remaining image processing and feature extraction operations were used on the picture of the diseased spot.

5) Feature extraction: In this stage, a collection of characteristics is extracted to describe the contaminated regions. It is according to color, shape, and textural attributes used to describe the regions. Since the elements of color and texture are different for each apple leaf disease, only these features are used for classification. The wavelet and co-occurrence matrix from the grey level is used for texture features. Hence, a maximum element including 14 color features including intensity, angle, four statistical features belonging to the *a as well as *b components of the *L*a*b and R space from the RGB space, and 24 wavelet features, two statistical elements for four coefficients in three Wavelet level, 32 features consisting of the co-occurrence matrix, eight statistical characteristics in four directions of 0, 45, 90 and 145 degrees were extracted for each type of disease.

6) Training preparation: In this section, the structure of different sets of neural networks is generated. This network structure divides the dataset into training, verification, likewise testing sets. In this study, training data accounted for 70% of the total data, 10% was used for validation data, and 20% was considered for the testing set.

B. Post-Processing

1) Proposed NN classification model: Choosing a suitable architecture for the neural network significantly impacts its classification and diagnosis performance. In this study, a multi-layer perceptron (MLP) neural network has been used,

where the first layer has 33 input nodes, and the last layer contains one output node.

It is also essential to choose the correct number of hidden layers and neurons for each layer. According to extensive experimental and results comparison, the structure of the MLP involves 70 nodes for the input layer and one node for the output layer. Furthermore, one and two hidden layers were used, and the number of 3 to 5 neurons was also considered for each of these layers. For the neurons of the network's hidden layers, the hyperbolic and sigmoid tangent transfer function was selected. For the last layer, it was decided to use the linear transfer function.

2) Model testing: After proposing the NN model, it is required to test the classification using testing data. In this research, to assess the accuracy of the proposed model, two assessment metrics are used Correlation Coefficient (CC) and Mean Square Error (MSE).

$$CC = \frac{\sum_{i=1}^{N}(X_i-\underline{X})(Y_i-\underline{Y})}{\sqrt{(X_i-\underline{X})(Y_i-\underline{Y})}} \quad (1)$$

$$MSE = \sqrt{\frac{1}{N}\sum_{N=1}^{N}(X_n - Y_n)^2} \quad (2)$$

In the above Equations, N is the number of data, Y_i is the predicted value, X_i is the measured value, and $\underline{X}$, $\underline{Y}$ respectively represent the average measured and predicted values. In the performance evaluation of the proposed NN model, the input data enters the network through the first layer. After passing through the different layers of the network, the network's output is obtained. Now, having the network output and the measured output value, the error value (RMSE) is calculated.

3) Model generation: After ensuring the effectiveness of the proposed system in classification, it must be utilized as a tool for diagnosing apple tree leaf disease.

IV. Results and Discussion

This section provides the experimental result, performance evaluation and discussion of the proposed method in this study. In the next sub-section, the discussion is also presented to discuss in detail the results and clarify the efficiency of the method.

A. Experimental Results and Performance Evaluation

The first step to collecting the proposed model results is to prepare the data presented in full in the previous section. The next step is choosing the neural network architecture. To achieve the best neural network model, various network architectures were implemented and evaluated. As mentioned above, the MLP architecture was used as a high-efficiency architecture for neural network training. Fig. 3 shows experimental results using the generated model. In Fig. 3, leaf disease and non-disease are shown in red and green color boxes.

Table I evaluates the ability of the neural network trained by the NN model algorithm to diagnose apple tree diseases. Moreover, experimental results for the proposed method and Support Vector Regression (SVR) model were collected for apple tree disease detection, and their results were evaluated and compared. Table II indicates the result of the SVR model.

Six hundred forty data samples are available in the dataset to test the models. Among the available data, 320 of these leaves were infected with Alternaria disease, 184 were infected with apple black spot disease, and 136 were infected with Minoz blight. Table III shows how each model could correctly diagnose the number of disease samples. For example, the neural network model has been able to accurately analyze 316 samples of 320 samples related to Alternaria disease and wrongly diagnose four samples as black spot disease.

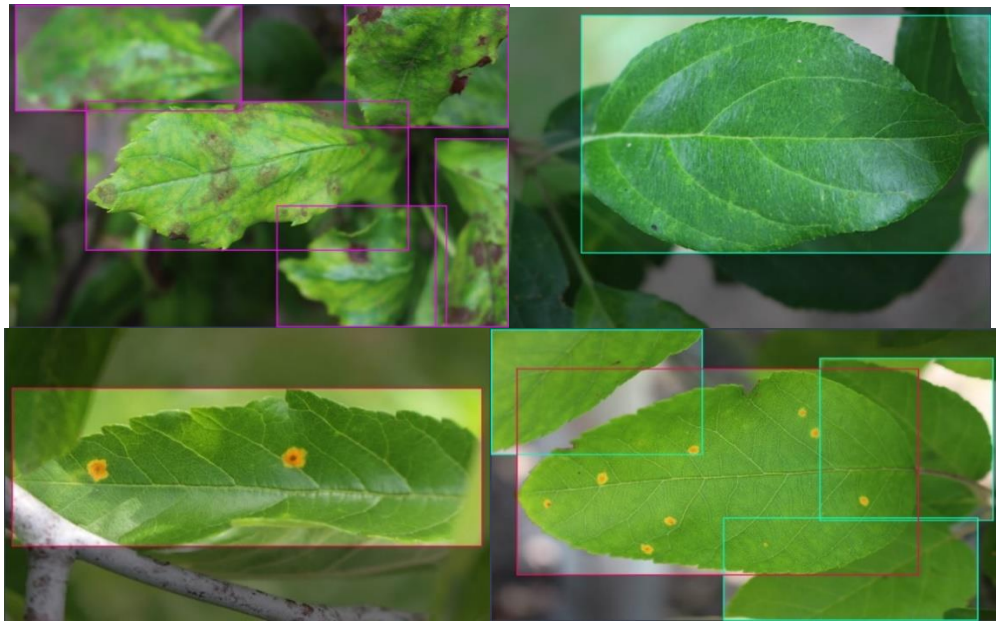


Fig. 3. Experimental results (includes leaf disease and non-disease).

TABLE I. PERFORMANCE EVALUATION FOR THE PROPOSED MODEL

Class	The proposed NN model						
	0	1	2	3	4	Total	Performance
1	0	316	4	0	0	320	0.98
2	0	4	179	1	0	184	0.97
3	0	1	5	130	0	136	0.95
Total	0	321	188	131	0	640	-

TABLE II. PERFORMANCE EVALUATION FOR THE SVR MODEL

Class	SVR model						
	0	1	2	3	4	Total	Performance
1	2	313	3	2	0	320	0.97
2	1	2	174	7	0	184	0.95
3	3	0	5	128	0	136	0.94
Total	6	315	182	137	0	640	-

TABLE III. RESULTS OF EVALUATION METRICS

Model	Training data		Testing **data**	
	RMSE	CC	RMSE	CC
ANN	0.091	0.978	0.098	0.976
SVR	0.12	0.978	0.30	0.963

Table III demonstrates that the proposed model's CC index for the test data is equivalent to 0.976, which is higher than the CC values obtained from other models. In addition, the RMSE index of the proposed model for the test data is equal to 0.098, which is lower than the values of other models. As a result, it can be said that the proposed model performs superior to other models in diagnosing the type of apple tree leaf disease.

B. Discussion

This section presents a discussion and in more detail of the proposed methodology in this research. As discussed earlier, this research emphasizes evaluating and comparing the proposed approach for identifying apple tree illnesses, which is based on a neural network (NN) model algorithm. The proposed method's outcomes are contrasted with the Support Vector Regression (SVR) model.

The evaluation of the neural network trained using the NN model technique for identifying diseases in apple trees is shown in Table I. We gathered and compared the experimental findings for the SVR model and the proposed method. The dataset utilized to test the models included 640 data samples, as shown in Table II. 320 of these samples contained Alternaria disease, 184 contained apple black spot disease, and 136 contained Minoz blight infections.

The effectiveness of each model in diagnosing the various illness samples is detailed in Table III. For instance, the neural network model correctly identified Alternaria disease in 316 out of 320 samples but mistakenly identified black spot disease in four samples. The proposed model outperformed other models in terms of the CC index, which is shown in Table III as having a value of 0.976 for the test data. The proposed model's RMSE index for the test data is also 0.098, which is lower than the results produced by previous models.

Finally, based on these findings, we observed that the model performs better than previous models in identifying the specific apple tree leaf disease type. The proposed method outperforms previous models in detecting and classifying apple tree illnesses, according to the high CC index and low RMSE index.

V. CONCLUSION

The paper presents a novel method for detecting three apple tree leaf diseases: apple black spot, Alternaria, and Minoz blight. The method consists of two main components: digital image processing and neural network (NN) techniques. The digital image processing component is responsible for preparing, processing, and extracting features from the leaf images. The neural network component is responsible for classifying the diseases based on the features. The paper describes the details of each component and evaluates the performance of the method using a dataset of 300 leaf images. The paper also compares the method with other existing methods and shows that the proposed method achieves higher accuracy and performance in identifying the apple tree leaf diseases. The paper concludes by suggesting some future directions for improving the method, such as extending it to a deep learning-based model and exploring various convolutional neural networks (CNN) for better feature extraction and classification.

ACKNOWLEDGMENT

This research was supported by the Natural Science Foundation of Hunan Province, Systematical Development and Application of Intelligent Greenhouse based on the Internet of Things (2022JJ50090).

REFERENCES

[1] A. A. Bharate and M. S. Shirdhonkar, "A review on plant disease detection using image processing," 2017: IEEE, pp. 103-109.

[2] V. Binnar and S. Sharma, "Plant Leaf Diseases Detection Using Deep Learning Algorithms," 2023: Springer, pp. 217-228.

[3] Z. Iqbal, M. A. Khan, M. Sharif, J. H. Shah, M. H. ur Rehman, and K. Javed, "An automated detection and classification of citrus plant diseases using image processing techniques: A review," Computers and electronics in agriculture, vol. 153, pp. 12-32, 2018.

[4] P. Kartikeyan and G. Shrivastava, "Review on emerging trends in detection of plant diseases using image processing with machine learning," International Journal of Computer Applications, vol. 975, no. 8887, 2021.

[5] L. C. Ngugi, M. Abelwahab, and M. Abo-Zahhad, "Recent advances in image processing techniques for automated leaf pest and disease recognition–A review," Information processing in agriculture, vol. 8, no. 1, pp. 27-51, 2021.

[6] X. Chao, G. Sun, H. Zhao, M. Li, and D. He, "Identification of apple tree leaf diseases based on deep learning models," Symmetry, vol. 12, no. 7, p. 1065, 2020.

[7] A. Rastogi, R. Arora, and S. Sharma, "Leaf disease detection and grading using computer vision technology & fuzzy logic," in 2015 2nd international conference on signal processing and integrated networks (SPIN), 2015: IEEE, pp. 500-505.

[8] H. Azgomi, F. R. Haredasht, and M. R. S. Motlagh, "Diagnosis of some apple fruit diseases by using image processing and artificial neural network," Food Control, vol. 145, p. 109484, 2023.

[9] E. Omrani, B. Khoshnevisan, S. Shamshirband, H. Saboohi, N. B. Anuar, and M. H. N. M. Nasir, "Potential of radial basis function-based support vector regression for apple disease detection," Measurement, vol. 55, pp. 512-519, 2014.

[10] S. Zhang, X. Wu, Z. You, and L. Zhang, "Leaf image based cucumber disease recognition using sparse representation classification," Computers and electronics in agriculture, vol. 134, pp. 135-141, 2017.

[11] A. Loutfi, S. Coradeschi, G. K. Mani, P. Shankar, and J. B. B. Rayappan, "Electronic noses for food quality: A review," Journal of Food Engineering, vol. 144, pp. 103-111, 2015.

[12] A. Tomar, G. Gupta, W. Salehi, C. H. Vanipriya, N. Kumar, and B. Sharma, "A Review on Leaf-Based Plant Disease Detection Systems Using Machine Learning," Recent Innovations in Computing: Proceedings of ICRIC 2021, Volume 1, pp. 297-303, 2022.

[13] Z. Chuanlei, Z. Shanwen, Y. Jucheng, S. Yancui, and C. Jia, "Apple leaf disease identification using genetic algorithm and correlation based feature selection method," International Journal of Agricultural and Biological Engineering, vol. 10, no. 2, pp. 74-83, 2017.

[14] A. Rastogi, R. Arora, and S. Sharma, "Leaf disease detection and grading using computer vision technology & fuzzy logic," 2015: IEEE, pp. 500-505.

[15] K. R. Gavhale and U. Gawande, "An overview of the research on plant leaves disease detection using image processing techniques," Iosr journal of computer engineering (iosr-jce), vol. 16, no. 1, pp. 10-16, 2014.

[16] S. S. Harakannanavar, J. M. Rudagi, V. I. Puranikmath, A. Siddiqua, and R. Pramodhini, "Plant leaf disease detection using computer vision and machine learning algorithms," Global Transitions Proceedings, vol. 3, no. 1, pp. 305-310, 2022.

[17] N. Ganatra and A. Patel, "A multiclass plant leaf disease detection using image processing and machine learning techniques," International Journal on Emerging Technologies, vol. 11, no. 2, pp. 1082-1086, 2020.

[18] M. Prabu and B. J. Chelliah, "An intelligent approach using boosted support vector machine based arithmetic optimization algorithm for accurate detection of plant leaf disease," Pattern Analysis and Applications, vol. 26, no. 1, pp. 367-379, 2023.

[19] Z. Ghasemi Varjani, S. S. Mohtasebi, H. Ghasemi, and E. Omrani, "Developing a new hybrid system for detection of apple tree leaves diseases," Iranian Journal of Biosystems Engineering, vol. 49, no. 2, pp. 215-225, 2018.

[20] M. Ang, E. Sundararajan, K. Ng, A. Aghamohammadi, and T. Lim, "Investigation of Threading Building Blocks Framework on Real Time Visual Object Tracking Algorithm," Applied Mechanics and Materials, vol. 666, pp. 240-244, 2014.

The Use of Hand Gestures as a Tool for Presentation

Hope Orovwode[1], John Amanesi Abubakar[2], Onuora Chidera Gaius[3], Ademola Abdullkareem[4]
Department of Electrical and Information Engineering, Covenant University, Ota, Ogun State, Nigeria[1, 3, 4]
Department of Computer Science and Engineering, University of Bologna, Bologna, BO, Italy[2]

***Abstract*—Our hands play a crucial role in daily activities, serving as a primary tool for interacting with technology. This paper explores using hand gestures to control presentations, offering a dynamic alternative to traditional devices like mice or keyboards. These conventional methods often limit presenters to a fixed position and depend on the device's proximity. In contrast, hand gesture controls promise a more fluid and engaging presentation style. This study utilizes the HaGRID dataset, supplemented by custom-recorded data, divided into 80% for training, and 10% each for validation and testing. The data undergoes preprocessing and a linear classifier with four dense layers and a SoftMax activation layer is employed. The model, optimized with the Adam optimizer and a learning rate of 1e-1, incorporates a motion classifier (LSTM) with two dense layers and an LSTM layer, tailored for long-distance body pose estimation. The resulting application, a local desktop tool independent of internet connectivity, uses tkinter for its user interface. It demonstrates high accuracy in classifying gestures, achieving 90.1%, 89%, and 90% in training, validation, and testing, respectively, for the linear classifier. The motion classifier records 79.8%, 72%, and 70.1%. The model effectively recognizes and categorizes dataset gestures, capturing live camera feeds to manage presentations. Users benefit from various features, including PowerPoint selection, distance mode, gesture toggling and assignment, and appearance mode. This study illustrates how hand gesture control can enhance presentation experiences, merging technology with natural human movement for a more seamless interaction.**

***Keywords*—Hand gesture; linear classifier; motion classifier; LSTM; interface**

I. Introduction

Our hands are more flexible when it comes to the usage of our body and when it comes to movement, so they are involved greatly in our day-to-day activities. It is a natural occurrence that someone uses their hands immediately when they wake up or want to get something. Researchers have been going over ways of interacting with computers and other devices and the aim is trying to increase human-computer interaction (HCI) through a communication channel from the user to the device. One of the ways we can communicate with a computer is with our "hands" making movements which are said to be hand gestures. The use of hand gestures as a means of interacting with technology has become increasingly popular and allows users to interact more naturally and intuitively and it has started entering the field of presentation [1]. Hand gestures have been used as a form of major communication in presentations for ages. Hand gestures have been effectively used in presentations by renowned presenters and public figures across past times, including Martin Luther King Jr., Winston Churchill, and Steve Jobs, to captivate audiences and convey compelling messages [2]. This historical setting highlights the importance of hand gestures as an important part of effective communication. Hand gestures in the field of presentation enhance communication performance, especially for presenters who may experience stage fright. While slides are often seen as the focal point, effective hand gestures can captivate the audience and draw their attention into the presentation. It then aids in emphasizing essential points, drawing focus on significant details, and illustrating ideas or operations. Presenters can improve listener understanding and recollection of information by adopting gestures that correspond to the topic [3]. Non-verbal indicators, such as hand gestures, are an important part of communicating. These aid in the general comprehension and understanding of the information. Presentations get better when presenters enhance their verbal communication, express emotions, and include richness in their presentations by integrating intentional and well-timed gestures [4].

In recent years, many eyes have been on various technologies to offer improved user interfaces that enable computer interactions as intuitive as human interactions. Furthermore, standard gadgets such as a keyboard and mouse cannot entirely fulfil human interaction needs in presentations [5]. The study on controlling presentations using hand gestures is significant as it aims to enhance user experience, increase interactivity, improve accessibility, enable seamless integration, and drive innovation in presentation delivery. By developing a presentation package that allows presenters to control their presentations using hand gestures, the study seeks to provide a more intuitive and natural interaction method. This approach eliminates the need for physical devices or complex keyboard shortcuts, making the presentation process seamless and enjoyable. Incorporating hand gestures in presentations enhances interactivity, captures the audience's attention, and promotes engagement. The study also emphasizes the importance of seamless integration with popular presentation software and contributes to the advancement of interaction techniques and innovative presentation delivery methods [6].

In summary, this paper investigates the gap focusing on the practical implementation of hand gesture recognition. Section II delves into the existing literature on the subject, establishing the foundation for our study. Section III outlines the research methodology employed, followed by Section IV which presents the results and discussion. Section V presents the user implementation. Finally, Section VI concludes the paper.

II. Related Work

When a 15th-century author describes an individual as 'cute of gesture', they are not referring to mere clapping or shaking performed gracefully. Instead, the author is highlighting the

person's graceful movements and posture [7]. This broader bodily movement was referred to as 'gesture'. The study of gestures is not a recent endeavor. Various physiognomists, such as G. B. Della Porta, Charles Le Brun, and J. C. Lavater, have explored the representation of gestures since the Medieval period. In the 17th Century, Francis Bacon emphasized the importance of gestures as a primary form of symbolism. Giovanni Bonifacio and John Bulwer discussed a universal language of gestures that could facilitate international trade and interactions [8] [9] [10]. Charles Darwin's work on the expression of emotions in humans and animals in the 18th century provided further evidence for the biological inheritance of expressions [11]. Present-day ethologists emphasize the similarities between human and animal bodily movements used to convey emotions such as resentment, superiority, or possessiveness. Certain emotional facial expressions like laughter, tears, yawning, and giggling are nearly universal, transcending linguistic and geographical boundaries [12]. However, contemporary views suggest that a global language of gestures is not feasible, as cultural, and social differences play a significant role. There are also various types of gestures and languages [12].

In the early 19th century, Andrea de Jorio attempted to recreate the mimicry of classical antiquity using Neapolitan gestures. Anthropologist Marcel Mauss, in his work 'The Techniques of the Body' (1935), highlighted the vast cultural variations in bodily activities. These differences became apparent when people from different cultures watched foreign films or encountered unfamiliar gestures. Marcel Mauss' insights led to cross-cultural studies of body language and facial expressions. David Efron's research explained how Italians and Eastern Europeans adapted to the gesture culture of the United States [13] [14]. After World War II, interest in communication surged once again. New development theories introduced powerful models for enhancing communication through analogue and digital codes. The language was seen as an example of digital code, while elements of changing natural behaviour were also considered as forms of language. Gestures, although frequently mentioned, received limited attention in this context [15]. This distinction gave rise to the concept of "nonverbal communication," distinct from oral language, focusing on interpersonal relationship building [16] [17]. Ray L. Birdwhistell introduced the concept of kinesics, aiming to study body movement communication systematically, although this didn't lead to extensive studies of gesture [18] [19]. The coalescence of nonverbal communication studies was dedicated to aspects of behaviour not directly linked to spoken language. Ray L. Birdwhistell's contribution was notable, but his definition of kinesics overlooked the analysis of gestures. Some authors, like Morton Wiener and Paul Ekman, attempted to incorporate gestures into nonverbal communication theory, but their efforts remained somewhat isolated [20] [21].

Gordon Hewes' influential study reignited interest in gestures as a key topic of discussion. Hewes argued that original language might have been gestural and cited Gardner's discovery of sign language in juvenile monkeys as a significant foundation. The study of sign languages progressed rapidly after William Stokoe's work on American Sign Language. Although sign language differs from spoken language, it poses a challenge to linguistic models and necessitates its incorporation into linguistic perspectives [22][23]. Gesture and sign language are intertwined, and the resurgence of interest in sign language has contributed to the recognition of gesture as a significant research area once again. A lot of techniques have been used in advancing hand gestures. The research in [24] used CNN and LSTM for gesture recognition. The authors used the CNN approach followed by the CNN+LSTM approach with a skeleton model (hand points), and they empirically showcased the capability of extracting pertinent attributes from 3D skeleton data. This extraction serves as a precursor to effectively address activity recognition through the utilization of LSTM. The result shows a high performance. Gesture identification, tracking and classification were carried out. K. V. Eshitha et al., [25] firstly, segment the hand gestures by using the skin colour model and AdaBoost classifier based on the type of skin colour, which also segments the hand from the background. It is monitored by the Cam Shift algorithm and classified by CNN to recognize 10 common digits. The result shows a 98.3% accuracy. A. Ikram et al., [26] carried out gesture acquisition, segmentation, feature extraction and classification by compiling different gesture datasets, and these gestures are distinguished from an input image using colour-based segmentation then the Histogram of Gradient technique is used to extract features in this case. After segmentation, the gestures are classified into left-hand and right-hand gestures, together with labels, and then sent to an artificial neural network for training. The result shows a high accuracy. The authors combined CNN and RNN to increase the accuracy of gesture classification using dynamic hand gestures. The result obtained is 85.46% H. Y. Chung et al., [27] I. Dhall et al.,[28] carried out hand gesture classification of people with stroke. The data set contains 140 gestures and is trained using CNN. Accuracy gotten is 99%. P. Parvathy et al., [29] used CNN to improve the accuracy of gesture recognition using the OpenCV library and the result shows a 99.13% accuracy. A. Mujahid et al. [30] proposed a lightweight model based on YOLO and DArkNet-53. The gestures are obtained from the Cambridge hand gesture dataset and are pro-processed and segmented using various techniques and then classified using deep CNN. The result shows a 96.66% accuracy. However, despite extensive research in gesture recognition and classification, deployment has been limited.

III. System Analysis and Design

This section outlines a comprehensive methodology employed in this research. It covers the step-by-step procedure and methods utilized for using hand gestures as a tool for presentation. Additionally, it discusses the process of data acquisition, feature extraction and deployment of the model.

Fig. 1 illustrates a comprehensive block diagram of the system starting from data gathering, followed by Mediapipe hands for preprocessing and extraction then a hybrid model comprising of linear classifier and LSTM will be used to develop the model. This hybrid architecture has been thoughtfully designed to combine the strengths of both components, thereby enhancing the model's overall performance and predictive capabilities. The utilization of the LSTM component holds significant promise in capturing

temporal dependencies within the data, further elevating the model's predictive prowess. The training environment is the Jupital Notebook and finally, the model will be deployed in a model as a package.

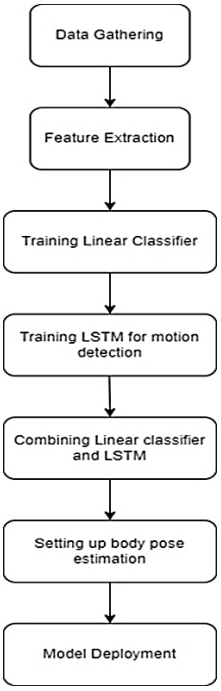


Fig. 1. Block diagram of the system.

A. Data Gathering

This is the first stage in the development of the model. There are two datasets used in the model. The first dataset is called HaGRID (Hand Gesture Recognition Image Dataset) for hand gesture recognition (HGR) systems. HaGRID size is 716GB and the dataset contains 552,992 (1920p × 1080p) RGB images divided into 15 classes of gestures as seen in Fig. 2. The data were split into training 80% for the training set 10% for the validation set and 10% for the test set.

The set contains at least 34,730 distinct individuals and scenes. The subjects range in age from 18 to 65. The dataset was primarily gathered at home, with significant variance in lighting, encompassing both artificial and natural light. Furthermore, the collection contains photographs captured under severe situations, such as facing and backing away from a window. In addition, the subjects were instructed to perform motions between 0.5 and four meters away from the camera.

The second dataset is a custom-made dataset for finger swipe motion which contains five seconds for each video and four classes of gestures.

Fig. 2. Sample of HaGRID dataset.

B. Data Preprocessing

This process utilizes functions provided by the OpenCV library:

1) Color channel conversion: The initial step involves converting the image captured by the webcam from the BGR (blue, green, red) colour channel mode to the RGB mode. This transformation is vital because the models we are using expect images in the RGB format for accurate processing and analysis.

2) Horizontal flipping: To ensure that the image is properly oriented for analysis, a horizontal flip is applied. Since webcams typically capture images facing the subject, this step prevents the image from being displayed in an inverted manner. It ensures that the hand gestures appear as intended when processed by the subsequent stages.

3) NumPy array conversion: Once the colour and orientation adjustments are made, the image is converted into a NumPy array. A NumPy array is a fundamental data structure used in Python for efficient numerical computations. In this case, the image is transformed into a format that can be readily understood and manipulated by deep learning models.

4) Adjusting array shape: The shape of the NumPy array is then modified to adhere to TensorFlow's specific requirements for input data. TensorFlow, a popular deep learning framework, expects the arrangement of data in a specific order. Therefore, the shape of the array is adjusted to have the colour channels come first, followed by the width and height dimensions. This ensures that the input data conforms to TensorFlow's expectations, enabling smooth processing and analysis by the deep learning models.

C. Feature Extraction

This is done using the media pipe hands model and is divided into two phases:

1) Hand detection model: In this phase, a Convolutional Neural Network (CNN) is employed to analyze images or video frames. The primary goal is to detect the presence and location of one or more hands within the visual data. This involves generating a bounding box or a region of interest (ROI) that encapsulates the hand(s). Before entering the CNN, the input image or frame undergoes preprocessing to ensure it is suitable for analysis. This might involve actions such as scaling, normalization, and data transformation into a format compatible with the neural network's requirements.

The architecture of the CNN is meticulously designed to learn patterns and features associated with hands. Comprising multiple layers, including convolutional, pooling, and activation functions, CNN focuses on extracting spatial features from the input data. Throughout this feature extraction process, the CNN discerns important elements like edges, textures, shapes, and spatial relationships that distinguish hands from the background or other objects.

The acquired features are then utilized for classification. Each portion of the input data is assigned a probability or

confidence score, indicating the likelihood of containing a hand.

2) Landmark detection model: After successfully identifying the hand's location using the bounding box, the focus shifts to the more precise estimation of hand landmarks. To achieve this, a specialized CNN is employed, fine-tuned to predict the positions of landmarks on the hand. This CNN operates on the hand ROI extracted from the previous phase.

Similar to the hand detection model, the landmark estimation network comprises layers dedicated to feature extraction. These layers meticulously analyze the hand ROI to capture vital visual patterns, encompassing edges, textures, and other defining characteristics that aid in pinpointing landmarks.

Drawing from these extracted features, the landmark estimation network generates predictions for the coordinates of each hand landmark. These landmarks, totalling 21 key points as seen in Fig. 3, represent precise positions within the image or video coordinate system. These points correspond to crucial locations on the hand, aiding in accurate identification and analysis.

These two phases work collaboratively to provide an integrated solution for hand gesture analysis. The hand detection phase establishes the presence and location of hands, while the landmark detection phase refines this information, pinpointing specific landmarks on the hand. This dynamic process enables precise and nuanced hand gesture recognition, forming the foundation for effective communication and interaction.

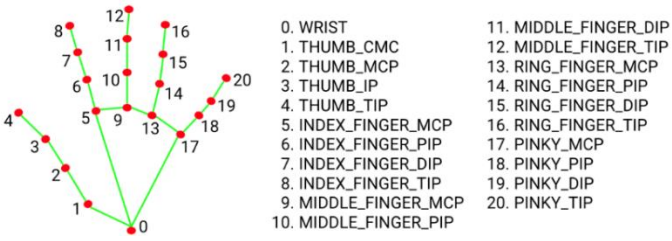


Fig. 3. The 21 key landmarks.

D. The Linear Classifier Model

To commence, the landmark features obtained from hands within the HaGRID dataset using the Mediapipe hands model serve as the input for training the linear classifier. This classifier is structured with four dense layers, also known as fully connected layers. In this arrangement, each neuron is connected to every neuron in the preceding layer and Rectified Linear Activation Units (ReLU) are placed between these neurons. The final layer employs a SoftMax activation, effectively converting logits into probabilities that facilitate classification. The architecture of the linear classifier model can be visualized in Fig. 4, offering a visual summary of its components and connectivity.

During the optimization process, the model is fine-tuned using the Adam optimizer, which utilizes a learning rate of 1e-1. This optimizer plays a pivotal role by iteratively adjusting the weights and biases of the network throughout the training process. It effectively manipulates these parameters to minimize the loss function, thereby enhancing the model's predictive accuracy.

The Adam optimizer operates with the inclusion of two beta parameters: β1 and β2. These parameters govern the decay rates of moving averages of gradients and squared gradients, respectively. In this context, β1 is set to 0.9, and β2 is set to 0.99. These values essentially dictate the extent to which past gradients and squared gradients influence the current weight adjustments during training.

To improve training accuracy and mitigate overfitting, a dropout mechanism is applied across all layers with a rate of 0.2. Dropout involves randomly deactivating a portion of neurons during each training iteration. This deliberate deactivation reduces interdependencies between neurons, encouraging the network to acquire more generalized and robust representations, ultimately enhancing its ability to perform well on new, unseen data.

E. The Motion Classifier Model

To achieve motion detection, we employed a specialized neural network called LSTM, which stands for Long Short-Term Memory. This LSTM network is designed to analyze sequences of landmarks captured throughout 30 consecutive frames or steps. To facilitate training, we utilized a dataset specifically recorded to capture different types of motion. Using the Mediapipe model, we extracted landmark features from this custom motion dataset. These extracted landmark features were then used as input to train the LSTM network. The motion classifier layer diagram is shown in Fig. 5 and the summary of the motion classifier model is shown in Fig. 6.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 64)                2752

 dense_1 (Dense)             (None, 128)               8320

 dense_2 (Dense)             (None, 64)                8256

 dense_3 (Dense)             (None, 32)                2080

 dense_4 (Dense)             (None, 18)                594

=================================================================
Total params: 22,002
Trainable params: 22,002
Non-trainable params: 0
_________________________________________________________________
```

Fig. 4. Linear classifier model architecture.

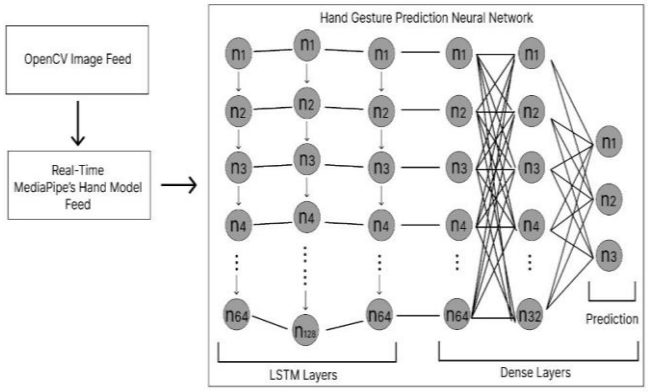
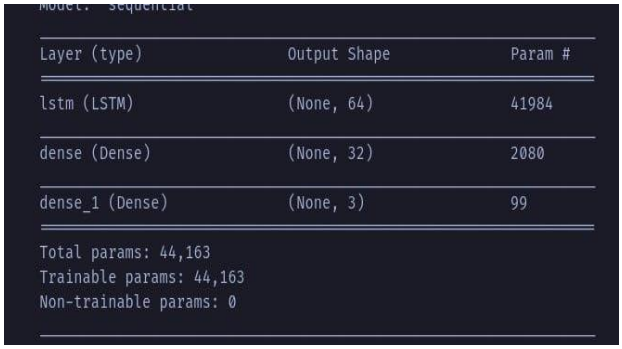


Fig. 5. Model classifier layer diagram.



Model: "sequential"

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 64)	41984
dense (Dense)	(None, 32)	2080
dense_1 (Dense)	(None, 3)	99

Total params: 44,163
Trainable params: 44,163
Non-trainable params: 0

Fig. 6. Summary of the motion classifier model.

The key strength of the LSTM network lies in its ability to recognize and comprehend patterns within motion data over an extended period. It achieves this by capturing both short-term and long-term dependencies present in the sequential landmark information. This capability enables the LSTM network to understand the intricate nuances of motion sequences, making it adept at detecting and interpreting motion patterns. Throughout the training process, the LSTM network continually adjusts its internal parameters to minimize the disparity between the predicted motion labels and the actual, true motion labels present in the dataset. By iteratively fine-tuning its parameters, the LSTM network effectively learns to classify and differentiate various types of motions. This process involves learning the distinctive characteristics and patterns associated with different motions, enhancing its ability to accurately detect and classify motions within new, unseen data.

F. Combining the Linear Classifier and the LSTM Model

Employing an LSTM (Long Short-Term Memory) model to analyze every individual frame in real-time for prediction is computationally demanding. Due to its high computational needs, we adopt a different strategy by integrating the two previously described models.

First, the linear classifier is applied to each frame captured in real time. The linear classifier, known for its computational efficiency, operates on the landmarks or features extracted from individual frames. It predicts the corresponding gesture or action for each frame efficiently. If the linear classifier confidently detects a meaningful gesture or action in a frame, indicating the potential presence of significant motion, a sequence of frames is then directed to the LSTM model for further motion analysis. The linear classifier's outcome serves as a trigger for the LSTM model's involvement. Specifically, the sequence of frames encompasses the present frame along with a set of previous frames. This sequence is inputted to the LSTM model, which has been designed to excel at capturing temporal patterns and dependencies within motion data.

By analyzing this sequence of frames over time, the LSTM model comprehends the evolving motion pattern. The LSTM's ability to account for sequential information empowers it to provide more precise predictions regarding the detected motion. This innovative approach involves the strategic fusion of the linear classifier and the LSTM model, thereby enhancing computational efficiency. The linear classifier rapidly processes individual frames, identifying potential gestures swiftly. However, the more computationally intensive LSTM model is only engaged when the linear classifier detects a gesture or action, ensuring resource-intensive calculations are focused precisely on moments of interest.

This approach effectively enables real-time motion detection while mitigating the overall computational demands. By leveraging the strengths of both models, we strike a balance between efficiency and accuracy, resulting in a streamlined and effective system for motion analysis.

G. Setting up Body Pose Estimation

To cater to situations in which users are situated at a significant distance from the camera, a specialized feature known as Long-Distance Mode is introduced for Body Pose Estimation. While the standard Mediapipe feature extraction excels when users are in proximity (<2m), it might not accurately capture hand landmarks for individuals located far away.

The concept is illustrated in Fig. 7, which presents a body pose diagram. In this context, we focus on specific points of interest, specifically points 13 to 22 that correspond to the location of the hands. Instead of relying on the precise detection of hand landmarks, the approach involves utilizing an estimated bounding box that encapsulates the hands' spatial extent. This bounding box, serving as input for the classifier layers, becomes the basis for making predictions. In this scenario, the classifier operates on the information derived from the bounding box rather than individual hand landmarks. This adjustment allows the classifier to predict and classify gestures or actions based on the overall position and region of the hands within the bounding box.

By incorporating the Long-Distance Mode and leveraging the bounding box estimation from the Mediapipe pose extraction, the system adeptly handles situations where users are positioned at a considerable distance from the camera. This innovation ensures that the system remains versatile and capable of accurately analyzing hand gestures, even when users are far away.

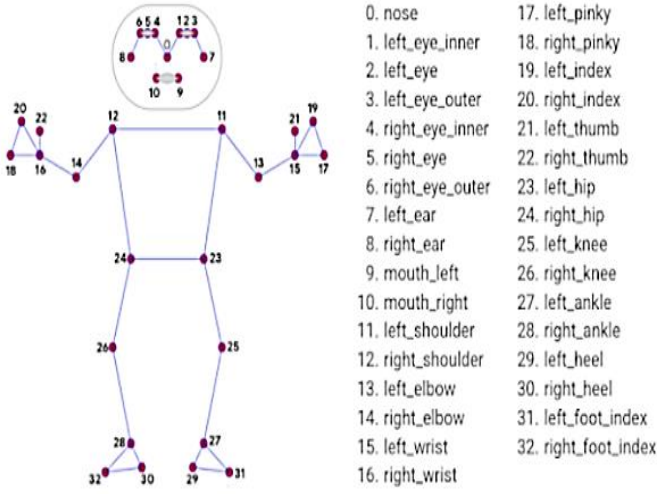


Fig. 7. Body pose diagram.

H. Flow Chart of the Model

The flow chart of the working of the model can be visualized in Fig. 8. The algorithm for the flow chart is as follows.

Step 1: Start

Step 2: Gather the data

Step 3: Carry out Preprocess and Data extraction

Step 4: Select the neural network to use for the model

Step 5: Set up body pose estimation

Step 6: Train, test and validate the model

Step 7: Is Performance Satisfactory?

- If Yes, Go to Step 8
- If No, Go to Step 4

Step 8: Save Model

Step 9: Stop

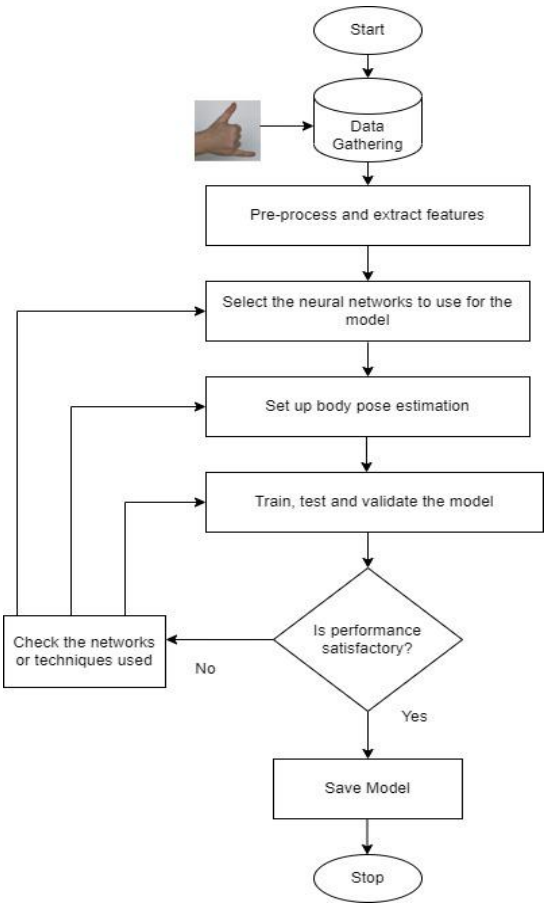


Fig. 8. Flow chart of the system.

I. Model Deployment

The model is built as a local desktop application that does not require any internet connection during usage. The user interface for the application was made using custom tkinter – a python library which uses the tkinter as its baseline but adds customizations to make the interface more user-friendly; also the models are deployed using TensorFlow lite (tflite) models to speed up inference.

Various diagrams as seen in Fig. 9 and Fig. 10 are used to illustrate the working of the deployed model as software. The diagrams are the use case diagram and the flow chart respectively. Fig. 9 shows the Unified Modeling Language (UML) of the system functions and responses, and Fig. 10 shows the flow of the deployed system.

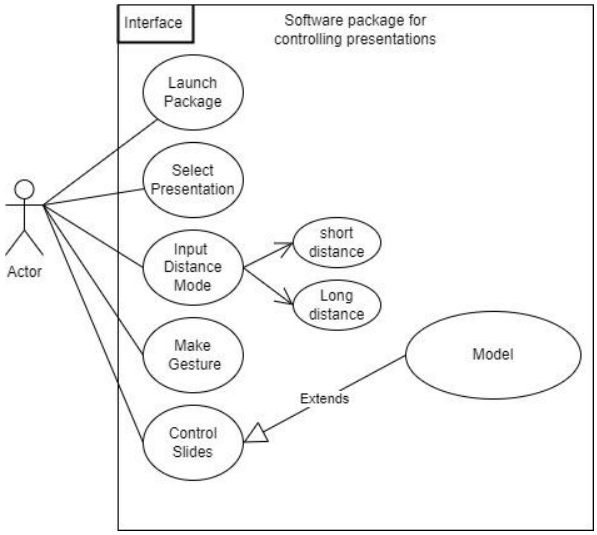


Fig. 9. Use case diagram.

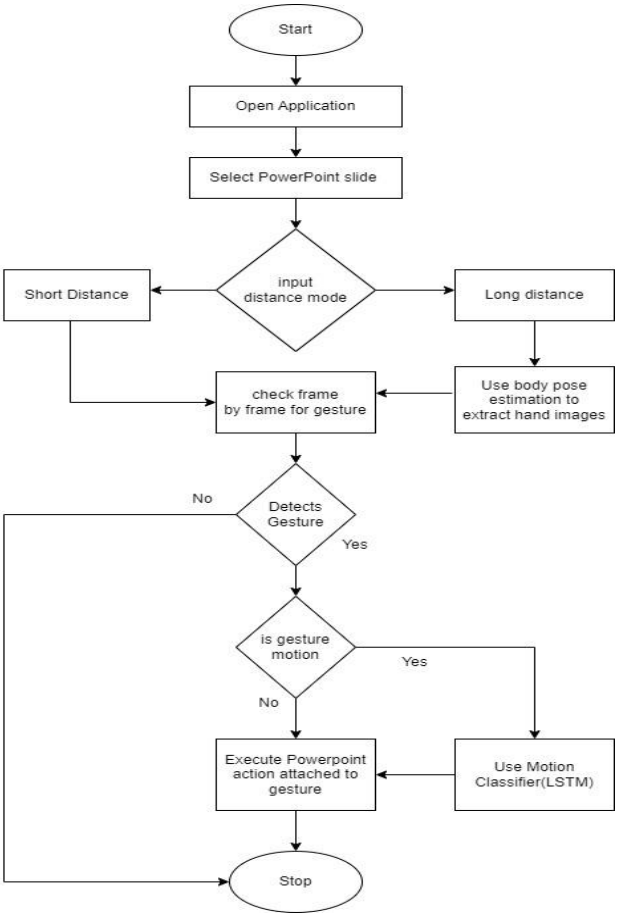


Fig. 10. Model deployed flow chart.

IV. Result and Discussion

This section discusses the results and execution of the model from the previous section and the performance of the system when evaluated. It defines the system constraints and tells us a bit more about the effectiveness of the system.

A. Model Training and Validation

The linear classifier model was trained using a strategy of 10 epochs, which means the entire dataset was processed 10 times. In each epoch, a batch of 128 data points was used for training. The entire training process took approximately three minutes, with each epoch lasting about 16 seconds. Notably, as the epochs progressed, the accuracy of the model consistently improved while the loss decreased. This is a positive sign that the model was learning effectively without becoming overfitted.

Throughout the training phase, the model achieved high levels of accuracy across different datasets: the training dataset, the validation dataset, and the testing dataset. Specifically, after the completion of 10 epochs, the model achieved accuracy rates of 90.1%, 89%, and 90% on the training, validation, and testing datasets, respectively. This indicates that the model had become well-acquainted with the dataset and was capable of accurately classifying the gestures within it.

Fig. 11 illustrates the model's training progress over each epoch. This visualization focuses on key metrics: loss, accuracy, val_loss (loss on the validation dataset), and val_accuracy (accuracy on the validation dataset). These metrics are tracked and displayed after every epoch to provide insight into the model's performance.

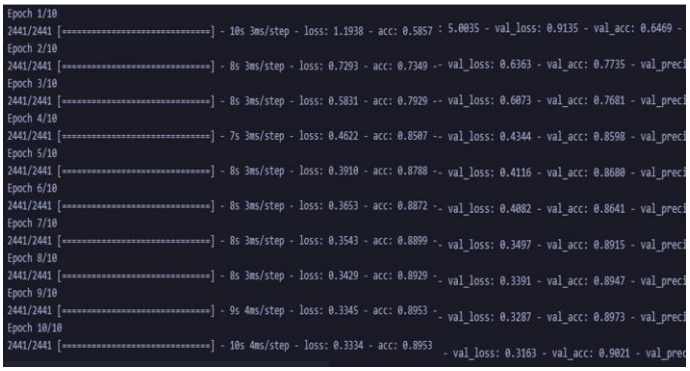


Fig. 11. Output of the linear classifier model's training process.

Here's a breakdown of what these metrics represent:

1) Loss: This metric indicates how much the model's predictions deviate from the actual values (labels) in the training dataset. It quantifies the error between predicted and actual values.

2) Accuracy: This metric represents the proportion of correctly classified data points in the training dataset. It shows how well the model is performing in terms of correctly predicting gestures.

3) Val_loss: Similar to the loss metric, val_loss measures the error between predicted and actual values, but it specifically pertains to the validation dataset.

4) Val_accuracy: Like accuracy, val_accuracy indicates the proportion of correctly classified data points, but on the validation dataset. It provides insight into how well the model generalizes to unseen data.

Fig. 12 and Fig. 13 provide a visual representation of the trends highlighted in Fig. 11, offering a closer look at the training and validation accuracy of the model as well as the corresponding training and validation loss throughout the 10 epochs. The x-axis on these figures denotes the number of epochs, while the y-axis illustrates the model's accuracy and loss.

From Fig. 12, it becomes evident that during the initial stages of training, the model's training accuracy starts at a relatively low point. At this early phase, the model lacks the understanding to accurately identify static hand gestures and is essentially making random guesses. However, as training progresses through each epoch, there is a steady and gradual improvement in accuracy. This trend signifies that the model is learning and becoming more adept at recognizing and distinguishing various hand gestures over time.

Examining Fig. 13, the graph portrays the trajectory of the model's loss on both the training and validation datasets across the epochs. Initially, during the early epochs, the training loss was notably high. This is attributed to the model's nascent stage, where it generates random and unrefined predictions, resulting in higher discrepancies between predicted and actual values. However, as the number of epochs advances, the training loss consistently diminishes. This decline indicates that the model is progressively honing its ability to accurately identify hand gestures within the training dataset.

Furthermore, the graph reveals the validation loss, which pertains to the model's accuracy in recognizing hand gestures on unseen validation data. As the epochs unfold, the validation loss follows a similar downward trajectory, signifying the model's enhanced capacity to generalize its learning from the training dataset to effectively identify hand gestures in new, previously unseen data.

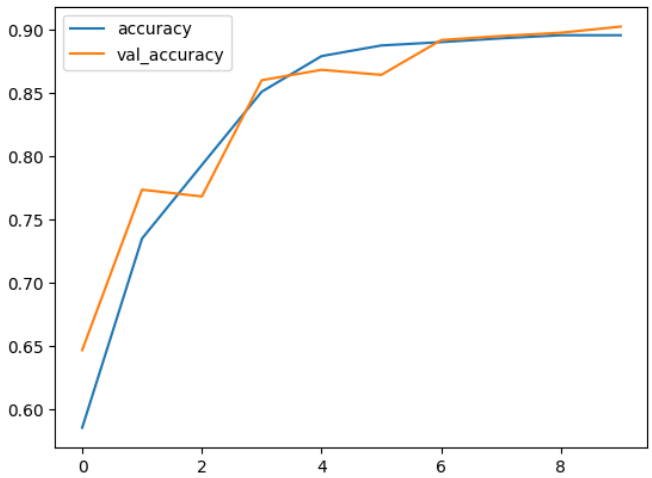


Fig. 12. Training and validation accuracy curves of the linear classifier model.

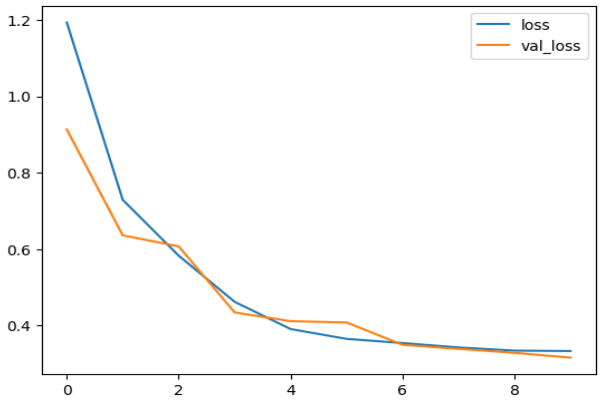


Fig. 13. Training and validation loss curve of the linear classifier model.

In essence, these graphical representations within Fig. 12 and Fig. 13 align with the narrative conveyed in Fig. 11, collectively providing a comprehensive insight into the model's iterative learning process. The figures underscore the model's journey from initial uncertainty and randomness to increasingly accurate and refined hand gesture recognition, ultimately attesting to its growing proficiency in this specific task.

The motion classifier model (LSTM) underwent training with a total of 200 epochs, where each epoch involved processing a batch of 64 data points. The entire training process took approximately 33 minutes, with each epoch lasting about 10 seconds. Similar to the linear classifier, the accuracy of the motion classifier increased while the loss decreased consistently with each successive epoch.

Throughout the training phase, the motion classifier model demonstrated high accuracy levels across different datasets: the training dataset, the validation dataset, and the testing dataset. Specifically, after the completion of the training process, the model achieved accuracy rates of 79.8%, 72%, and 70.1% on the training, validation, and testing datasets, respectively. This performance indicates that the model had become well-acquainted with the dataset and was proficient in correctly classifying different types of motions.

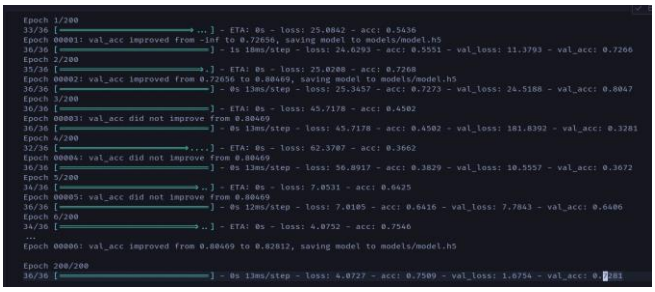


Fig. 14. Output of the motion classifier (LSTM) model's training process.

The training progress and performance of the motion classifier model are visualized in Fig. 14, 15, and 16. Fig. 14 provides an overview of the model's output during the training process, highlighting key metrics such as accuracy and loss across epochs. Fig. 15 and 16 depict the training and validation accuracy curves, as well as training and validation loss curves respectively, providing a graphical representation of the model's learning journey.

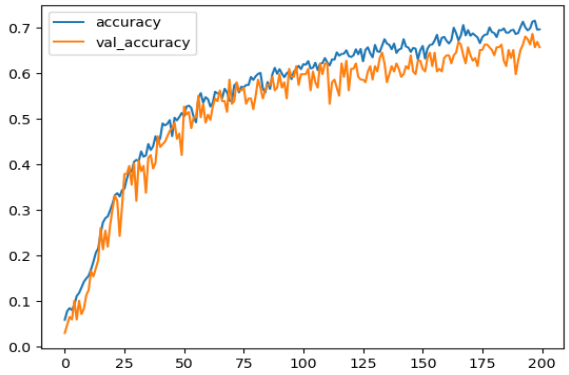


Fig. 15. Training and validation accuracy curve of the LSTM model.

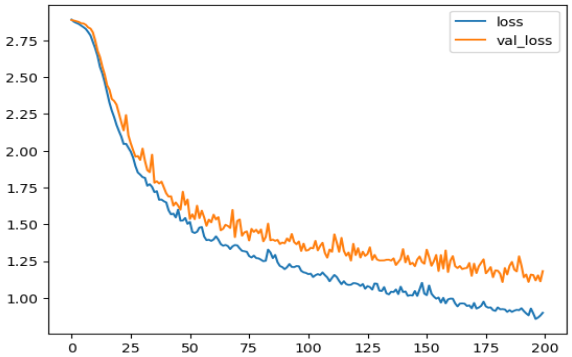


Fig. 16. Training and validation loss curves of the LSTM model.

In Fig. 15, it is observable that the accuracy initially starts at a lower point, similar to the linear classifier model. As training progresses over the epochs, the accuracy gradually improves, reflecting the model's increasing ability to recognize and classify motion patterns.

Fig. 16 displays the training and validation loss curves, illustrating a similar pattern observed in the linear classifier. At the start of training, the training loss is relatively high due to the model's random guesses. However, as the epochs advance, the training loss consistently decreases, indicating improved accuracy in recognizing motion patterns. The validation loss also follows a similar trajectory, indicating the model's enhanced ability to generalize its learning to unseen data.

B. Performance Index

Table I presents a comprehensive summary of how well the linear classifier model performed using various evaluation metrics. These metrics provide insights into the model's ability to accurately classify hand gestures.

- Accuracy: This metric measures the overall correctness of the model's predictions. The reported accuracy of 90% indicates that the model correctly classified 90% of the gestures in the dataset.
- Precision: Precision refers to the ratio of true positive predictions to the total number of positive predictions made by the model. In this case, the precision score of 92% implies that when the model predicts a gesture as positive (i.e., a certain hand gesture), it is correct 92% of the time.

- Recall: Recall, also known as sensitivity or true positive rate, calculates the ratio of true positive predictions to the total number of actual positive instances in the dataset. A recall of 90% indicates that the model correctly identified 90% of the actual hand gestures.
- F1-score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. An F1 score of 91% suggests that the model achieves a balanced trade-off between precision and recall.

TABLE I. PERFORMANCE OF THE LINEAR CLASSIFIER MODEL

Accuracy	*Precision*	*Recall*	*F1-score*
90%	92%	90%	91%

Fig. 17 visually represents the model's performance in terms of accuracy, precision, recall, and F1 score using a bar chart. This chart offers a clear overview of the model's strengths in terms of accuracy and precision, indicating its ability to make accurate predictions, particularly with high precision. However, it may miss some positive cases, as indicated by the relatively lower recall score.

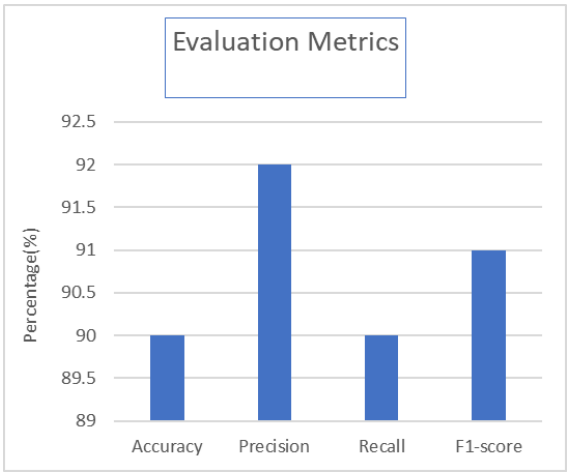


Fig. 17. Bar chart showing the linear classifier's performance.

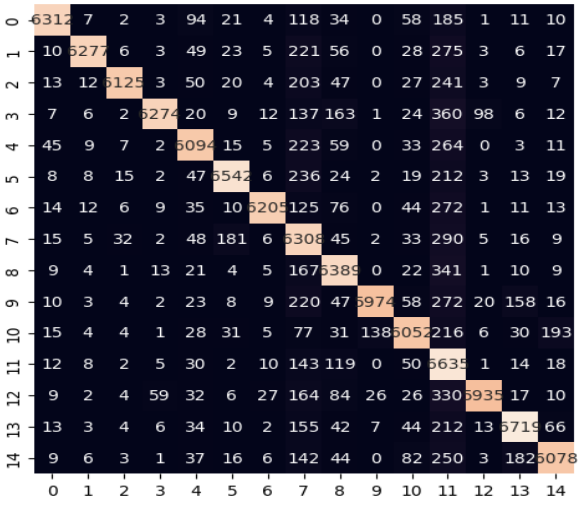


Fig. 18. The confusion matrix of the linear classifier model.

Fig. 18 and Fig. 19 present confusion matrices for the linear classifier model and the LSTM model, respectively. These matrices provide a visual depiction of the model's classification performance across different classes. They allow us to see how well the model correctly predicted each class and where it might have made errors.

Table II delves into the performance of individual classes. It provides a breakdown of precision, recall, and F1-score metrics for each class. For instance, for class 0, the precision score of 0.97 means that when the model predicted class 0, it was correct 97% of the time. The recall score of 0.92 indicates that the model correctly identified 92% of instances belonging to class 0. The F1-score of 0.94 represents a balanced measure of precision and recall for class 0.

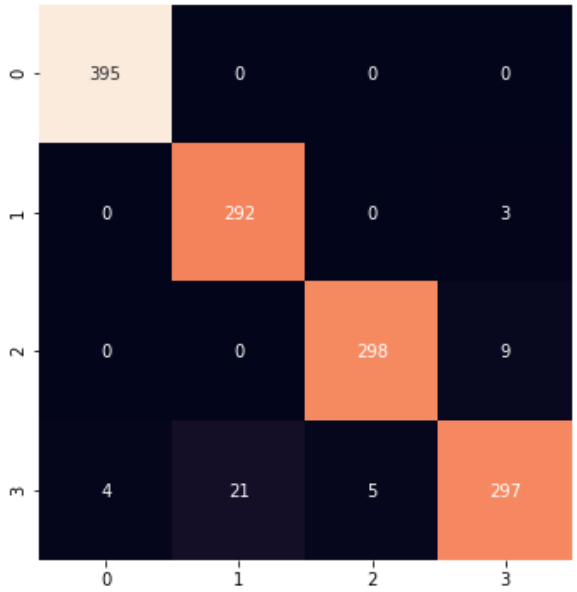


Fig. 19. The confusion matrix of the LSTM model.

TABLE II. PRECISION, RECALL AND F1-SCORE OF THE CLASSES

Class	*Precision*	*Recall*	*F1-Score*
0	0.97	0.92	0.94
1	0.99	0.90	0.94
2	0.99	0.91	0.94
3	0.98	0.88	0.93
4	0.92	0.90	0.91
5	0.95	0.91	0.93
6	0.98	0.91	0.94
7	0.73	0.90	0.81
8	0.88	0.91	0.90
9	0.97	0.88	0.92
10	0.92	0.89	0.90
11	0.64	0.94	0.76
12	0.97	0.88	0.93
13	0.93	0.92	0.92
14	0.94	0.89	0.91

V. User Implementation

A. The PowerPoint Selection

The interface permits users to choose which presentation file to open. The file browser approach is used to do this. The interfaces include a file browser which enables users to search the file system of their device for the desired PowerPoint presentation file. When selected, the slides pop up in the background as shown in Fig. 20.

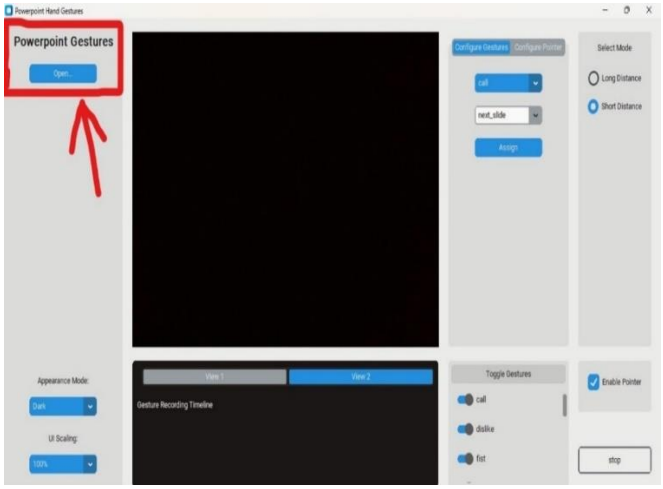


Fig. 20. Power point selection.

B. Distance Mode Selection

The user can select which mode is needed based on his/her distance from the system. A short distance is used in this operation. This can be visualized in Fig. 21.

C. Gesture Selection and Assignment

The interface allows the user to select gestures, and their functions and then assign them. In Fig. 22, the gesture used in this case is "call" and it is assigned to a function called "next_slide" which moves the slides page by page, the other 17 gestures in this model can also be assigned to other functions.

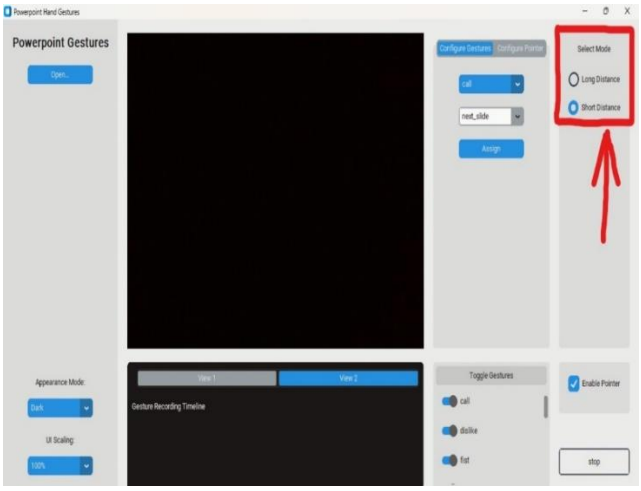


Fig. 21. Input distance mode.

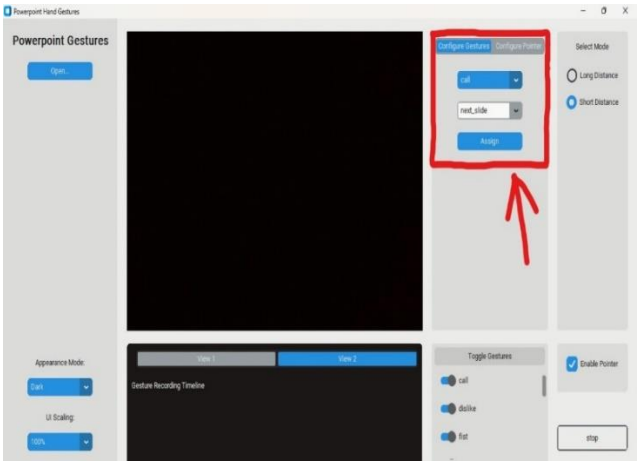


Fig. 22. Gesture selection and assignment.

D. Gesture Toggler

After assigning the gesture needed by the user, the user can turn on or off gestures based on their preferences as shown in Fig. 23, Fig. 24 And Fig. 25 shows that when the user toggles off a gesture, the gesture can no longer be recognized by the system and cannot perform its assigned function on the slides. When the gesture is toggled on, the system recognizes the gesture and the action assigned to it on the slides.

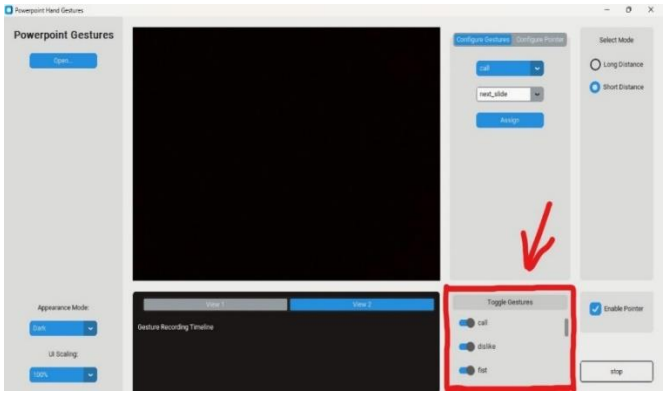


Fig. 23. Gesture toggler.

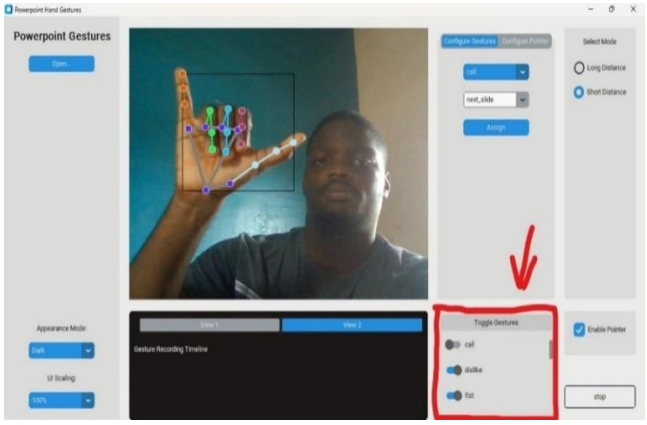


Fig. 24. Toggling OFF a gesture.

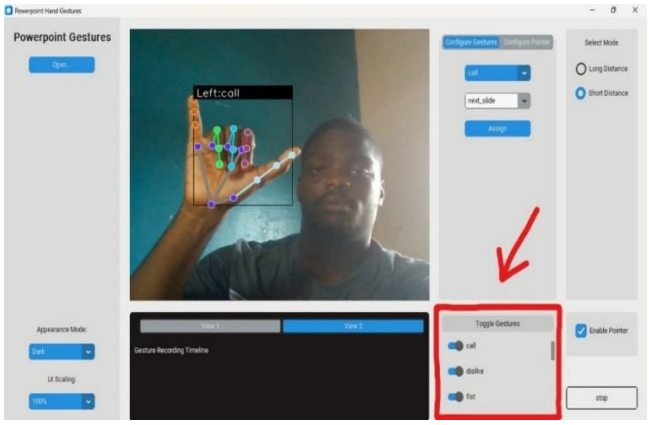


Fig. 25. Toggling ON a gesture.

E. Flow Chart

Fig. 26 shows the flow of the working of the package from the first stage to the end. This shows how the user should operate the software package.

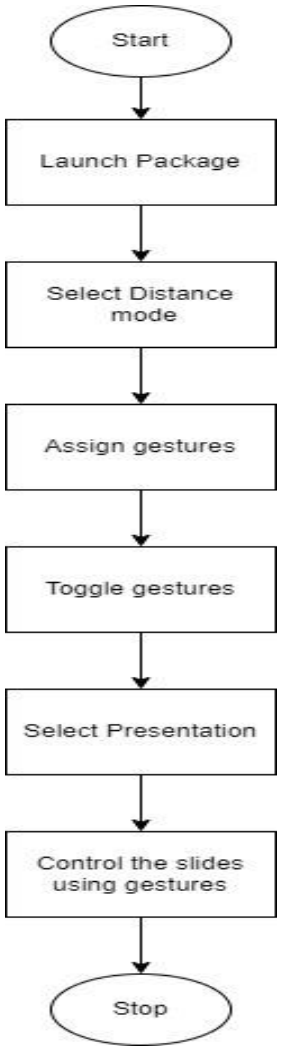


Fig. 26. Interface flowchart.

VI. Conclusion

In this study, we developed a software package that controls presentations using hand gestures. The system's performance was thoroughly examined through various evaluation metrics and visualization tools. Our findings highlight the effectiveness of the developed models in accurately identifying and classifying hand gestures.

The linear classifier model demonstrated impressive results, achieving an overall accuracy of 90%. This model's strong precision and recall scores further emphasize its ability to make accurate predictions while effectively identifying positive cases. The bar chart depicting the model's performance underscored its proficiency in precision, shedding light on its potential for minimizing false positive predictions. The confusion matrix provided valuable insights into the model's classification patterns. By analyzing true positives, true negatives, false positives, and false negatives, we gained a clearer understanding of where the model excelled and where improvements could be made. This granular analysis helps guide future refinements and optimizations of the system. Additionally, the LSTM model, designed to capture temporal patterns in motion data, demonstrated its efficacy in motion detection. While achieving lower accuracy compared to the linear classifier model, the LSTM model's performance remained solid, with an accuracy of 70.1% on the testing dataset. This model's potential for capturing long-term dependencies and recognizing complex motion patterns positions it as an asset for more nuanced applications. The study also addressed challenges related to distance by implementing Long-Distance Mode, which enhanced the system's adaptability to users situated far from the camera. This innovation showcases the system's robustness in accommodating varying scenarios and environments.

In conclusion, the developed system, consisting of a linear classifier and LSTM model, exhibits strong potential for real-time and accurate hand gesture identification. The combination of efficient classification and temporal analysis provides a comprehensive approach to gesture recognition. As a result, this system holds promise for a wide range of applications, from interactive presentations to virtual reality interfaces, enhancing user experience and interaction. Further research and optimization can propel this system towards even greater accuracy and utility in real-world settings.

Acknowledgment

The authors would like to acknowledge Covenant University for the funding of this paper.

References

[1] Y. Zhu, Z. Yang, and B. Yuan, “Vision-based hand gesture recognition,” in Proceedings of International Conference on Service Science, ICSS, 2013, pp. 260–265. doi: 10.1109/ICSS.2013.40.

[2] D. Ballotta, “Public speaking and presentations a critical review: The caring speaker,” 2010, [Online]. Available: http://hdl.handle.net/10071/2068.

[3] Alausa, D. W., Adetiba, E., Badejo, J. A., Davidson, I. E., Obiyemi, O., Buraimoh, E., ... & Oshin, O. (2022). Contactless palmprint recognition system: a survey. IEEE Access, 10, 132483-132505.

[4] S. E. Hollingsworth, K. Weinland, S. Hanrahan, M. Walker, E. Elwood, and M. Linsenmeyer, "Introduction to speech communication." Oklahoma State University Libraries, 2021.

[5] C. Roberto, "Development of a Hand-Gesture Recognition System for Human-Computer Interaction," 2014.

[6] Alausa, D. W., Adetiba, E., Badejo, J. A., Davidson, I. E., Akindeji, K. T., Obiyemi, O., & Abayomi, A. (2023, January). PalmMatchDB: An On-Device Contactless Palmprint Recognition Corpus. In 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA) (pp. 318-325). IEEE.

[7] A. Daly, J. Bremmer, and H. Roodenburg, A Cultural History of Gesture, vol. 39, no. 1. Cornell University Press Ithaca, NY, 1995. doi: 10.2307/1146410.

[8] Okokpujie, K., & Apeh, S. (2020). Predictive modeling of trait-aging invariant face recognition system using machine learning. In Information Science and Applications: ICISA 2019 (pp. 431-440). Springer Singapore.

[9] T. D'Orazio, R. Marani, V. Renò, and G. Cicirelli, "Recent trends in gesture recognition: How depth data has improved classical approaches," Image Vis. Comput., vol. 52, pp. 56–72, 2016, doi: 10.1016/j.imavis.2016.05.007.

[10] L. Benke, "Gesture, intimacy, and violence in contemporary artists' books," Reconstr. Stud. Contemp. Cult., vol. 16, no. 1, 2016.

[11] C. Darwin and F. Darwin, The expression of the emotions in man and animals. Oxford University Press, USA, 2009. doi: 10.1017/CBO9780511694110.

[12] A. Kendon, "Pragmatic functions of gestures some observations on the history of their study and their nature," Gesture, vol. 16, no. 2, pp. 157–175, 2017.

[13] J. Y. Mensah and M. J. Nabie, "The Effect of PowerPoint Instruction on High School Students' Achievement and Motivation to Learn Geometry.," Int. J. Technol. Educ., vol. 4, no. 3, pp. 331–350, 2021.

[14] J. Ruesch and G. Bateson, Communication: The social matrix of psychiatry. Routledge, 2017. doi: 10.4324/9781315080932.

[15] A. Kendon, "Introduction: Current issues in the study of 'nonverbal communication,'" in Nonverbal Communication, Interaction, and Gesture: Selections from SEMIOTICA, De Gruyter Mouton, 2010, pp. 1–53. doi: 10.1515/9783110880021.1.

[16] A. Kendon, "Gesture and anthropology: notes for a historical essay," Gesture, vol. 18, no. 2/3, pp. 142–172, 2019.

[17] A. Mehrabian, "Communication without words," in Communication Theory: Second Edition, Routledge, 2017, pp. 193–200. doi: 10.4324/9781315080918-15.

[18] P. Ekman and W. V. Friesen, "Hand movements," in Communication Theory: Second Edition, Routledge, 2017, pp. 273–292. doi: 10.4324/9781315080918-21.

[19] C. F. Hockett, "Language, mathematics, and linguistics," in Language, mathematics, and linguistics, De Gruyter Mouton, 2019.

[20] M. Weinert, "On the Contemporary Theories of the Development of Human Language," Acad. J. Mod. Philol., no. 12, pp. 229–238, 2021.

[21] J. M. Power, "Historical Linguistics of Sign Languages: Progress and Problems," Front. Psychol., vol. 13, 2022, doi: 10.3389/fpsyg.2022.818753.

[22] A. M. Borghi, F. Binkofski, C. Castelfranchi, F. Cimatti, C. Scorolli, and L. Tummolini, "The challenge of abstract concepts," Psychol. Bull., vol. 143, no. 3, pp. 263–292, 2017, doi: 10.1037/bul0000089.

[23] J. C. Nunez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Velez, "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition," Pattern Recognit., vol. 76, pp. 80–94, 2018.

[24] J. H. Sun, T. T. Ji, S. Bin Zhang, J. K. Yang, and G. R. Ji, "Research on the Hand Gesture Recognition Based on Deep Learning," in 2018 12th International Symposium on Antennas, Propagation and EM Theory, ISAPE 2018 - Proceedings, 2019, pp. 1–4. doi: 10.1109/ISAPE.2018.8634348.

[25] K. V. Eshitha and S. Jose, "Hand Gesture Recognition Using Artificial Neural Network," in 2018 International Conference on Circuits and Systems in Digital Enterprise Technology, ICCSDET 2018, 2018, pp. 1–5. doi: 10.1109/ICCSDET.2018.8821076.

[26] A. Ikram and Y. Liu, "Skeleton-based dynamic hand gesture recognition using LSTM and CNN," in ACM International Conference Proceeding Series, 2020, pp. 63–68. doi: 10.1145/3421558.3421568.

[27] H. Y. Chung, Y. L. Chung, and W. F. Tsai, "An efficient hand gesture recognition system based on deep CNN," in Proceedings of the IEEE International Conference on Industrial Technology, 2019, vol. 2019-Febru, pp. 853–858. doi: 10.1109/ICIT.2019.8755038.

[28] I. Dhall, S. Vashisth, and G. Aggarwal, "Automated hand gesture recognition using a deep convolutional neural network model," in Proceedings of the Confluence 2020 - 10th International Conference on Cloud Computing, Data Science and Engineering, 2020, pp. 811–816. doi: 10.1109/Confluence47617.2020.9057853.

[29] P. Parvathy, K. Subramaniam, G. K. D. Prasanna Venkatesan, P. Karthikaikumar, J. Varghese, and T. Jayasankar, "Development of hand gesture recognition system using machine learning," J. Ambient Intell. Humans. Comput., vol. 12, no. 6, pp. 6793–6800, 2021.

[30] A. Mujahid et al., "Real-time hand gesture recognition based on deep learning YOLOv3 model," Appl. Sci., vol. 11, no. 9, p. 4164, 2021.

SmishGuard: Leveraging Machine Learning and Natural Language Processing for Smishing Detection

Saleem Raja Abdul Samad[1], Pradeepa Ganesan[2], Justin Rajasekaran[3],
Madhubala Radhakrishnan[4], Hariraman Ammaippan[5], Vinodhini Ramamurthy[6]
College of Computing and Information Sciences, Information Technology Department,
University of Technology and Applied Sciences-Shinas, Sultanate of Oman[1,2,3.4,5]
Little Angel Institute, Karur, Tamil Nadu. India[6]

***Abstract*—SMS facilitates the transmission of concise text messages between mobile phone users, serving a range of functions in personal and business domains such as appointment confirmation, authentication, alerts, notifications, and banking updates. It plays a vital role in daily communication due to its accessibility, reliability, and compatibility. However, SMS unintentionally generates an environment where smishing can occur. This is because SMS is extensively available and reliable. Smishing attackers exploit this trust to trick victims into divulging sensitive information or performing malicious actions. Early detection saves users from being victimized. Researchers introduced different methods for accurately detecting smishing attacks. Machine Learning models, coupled with Language Processing methods, are promising approaches for combating the escalating menace of SMS phishing attacks by analyzing large datasets of SMS messages to differentiate between legitimate and fraudulent messages. This paper presents two methods (SmishGaurd) to detect smishing attacks that leverage machine learning models and language processing techniques. The results indicate that TF-IDF with the LDA method outperforms Weight Average Word2Vec in precision and F1-Score, and Random Forest and Extreme Gradient Boosting demonstrate higher accuracy.**

***Keywords*—Smishing; phishing; SMS; machine learning; natural language processing; TF-IDF**

I. INTRODUCTION

SMS (Short Message Service) is a pervasive mode of communication that spans geographic boundaries and device types in the digital world. It delivers information, warnings, and notifications quickly and reliably, making it essential for personal and professional interactions. However, this convenience raises the risk of smishing. Smishing, a malicious combination of "SMS" and "phishing," represents a grievous and pervasive threat to cybersecurity [1]. It is a particularly effective vector for cybercriminals due to the prevalence of mobile phones and the inherent trust in text messages, that causes serious harm to both individuals and organizations. Many victims are tricked into giving over personal information, clicking on harmful links, or unintentionally installing malware on their mobile devices, compromising their financial security, privacy, and digital identity [2]. The compromise of sensitive personal information can lead to identity theft, unauthorized account access, and additional intrusions. Smishing is especially dangerous because attackers are constantly changing their strategies, and with the low technical barrier to entry, attackers of all ability levels can participate, making this danger widespread. They use social engineering and psychological tricks to make text messages that look real and trustworthy. Even the most cautious people can be tricked by urgent messages posing as from banks, governments, or well-known businesses, forcing them to take steps that would benefit the attackers [3]. Additionally, when personal and financial information is stolen, it makes people more likely to be victims of hacking in the future, since attackers can use this information for damaging purposes.

In this dynamic environment, machine learning (ML) arises as a formidable ally in the defense against smishing attacks [4]. By their very nature, machine learning algorithms excel at pattern recognition, allowing for the detection of nuanced clues within text messages that reveal fraudulent intent. Most importantly, machine learning systems always learn from new data and adapt to the changing strategies used by smishing attackers. This adaptability is crucial because smishing attacks evolve to avoid detection and capitalize on emergent trends. Furthermore, machine learning scales well, enabling real-time analysis of enormous amounts of SMS messages across large mobile networks, enabling proactive detection and prevention. One of the main goals of machine learning in smishing detection is to reduce false positives, which prevents valid SMS messages from being inadvertently tagged as suspicious and protects the integrity of communication channels. Machine Learning with Natural Language Processing (NLP) methods enhance the detection of smishing attacks. ML-NLP models excel at identifying linguistic anomalies and patterns within SMS messages, distinguishing smishing-specific keywords, phrases, and grammatical inconsistencies [5].

Researchers introduced several smishing detection strategies using deep and machine learning techniques. Most methodologies use machine learning or deep learning models to extract features from textual content for classification. Generating features from textual data, such as counting the number of words or special characters in a document, provides structural insights and facilitates data representation. However, these features are not concerned with the context or semantics of the text but rather with its structural characteristics. Sometimes, the generated features may not optimize the potential of the dataset. Particularly in smishing detection, the context of the textual content is more critical than structural characteristics. By capturing the relationships

between words and contextual meaning, NLP methods for fully vectorizing text, such as TF-IDF (Term Frequency Inverse Document Frequency) or word embeddings, provide a richer grasp of the text's semantic content and make complicated analysis and machine learning tasks possible [6]. This paper introduces two novel methods (SmishGuard), which use TF-IDF with LDA (Latent Dirichlet Allocation) topic proportion score and Average Word2vec with TF-IDF score as weight for smishing detection and compares the performances. The results show that algorithms using random forest and extreme gradient boosting attain higher accuracy.

Contribution of the work:

- Two novel methods (SmishGuard) for smishing detection.
- The first method combines TF-IDF with LDA topic proportion scores to improve smishing detection by providing a comprehensive view of SMS messages that captures both term-level and underlying topics.
- The second method uses Average Word2Vec with TF-IDF scores as weights for smishing detection capitalizing on the strengths of both approaches, capturing semantic information and term importance.
- Proposed methods decrease false positives and enhance cybersecurity by enabling more precise and context-aware detection.
- Experimental evaluation and Performance Comparisons.

The research work's sections are arranged as follows. Section I delineates the problem's significance and outlines the proposed research works. The background is in Section II. In Section III, related works are highlighted. Section IV describes the proposed methodologies in detail. Experimental results are explained in Section V. The paper concludes with Section VI.

II. BACKGROUND

Smishing attacks, a form of phishing that utilizes SMS or text messaging, pose a serious cybersecurity risk. In these attacks, scammers typically send false communications to victims to coerce them into disclosing personal information, opening malicious links, or downloading dangerous files. Machine learning analyses content, sender information, and message context to detect smishing using natural language processing methods. Through training models on labeled datasets that include examples of both smishing and real messages, machine learning systems can identify patterns, language indications, and typical behavioral oddities. This proactive approach enhances the capability to automatically identify and flag suspicious messages, thereby protecting users from smishing frauds and boosting the security of mobile communication channels. This section describes the key technologies utilized in the proposed methods.

- TF-IDF Vectorizer: Using vectorizers, the unstructured text data are transformed into a numerical representation that machine learning algorithms can process. They convert text to numerical vectors. TF-IDF vectorizer creates a matrix where each document is a row, and each distinct word is a column from a set of text documents. A term's TF-IDF score is calculated by multiplying TF and IDF. TF counts how frequently a term appears in a document, while IDF measures a word's rarity across the corpus [7].
- Latent Dirichlet Allocation (LDA): LDA plays a key role in detecting smishing (SMS phishing) by revealing hidden themes and topics in SMS communications. It enables the identification of underlying linguistic patterns frequently observed in fraudulent messages, enabling the classification of incoming texts as either legitimate or potentially malevolent. By leveraging LDA, smishing detection systems acquire the ability to distinguish context and content nuances, making them more adept at recognizing attacker's deceptive techniques [8] [9].
- Average Wor2Vec Word Embedding: It expands on Word2Vec word embeddings by providing a method for representing entire sentences or documents as dense vectors. Average Word2Vec computes the vector representation of a sentence by calculating the average of the word vectors in it rather than considering each word separately. The approach involves converting each sentence word to its Word2Vec form and determining the mean of these vectors. Thus, the whole phrase is a short, fixed-length vector that captures its meaning. Average Word2Vec efficiently represents sentences while preserving semantic links and context from the Word2Vec model. It helps machine learning algorithms understand sentences and documents of different lengths by providing a constant vector dimension [10].
- Machine Learning Algorithms (ML): In detecting smishing messages, machine learning models are quite effective [11]. They are essential in recognizing and classifying smishing messages with dangerous or harmful information, which helps to improve cyber security and shield users from potential risks. Most of the research work utilizes different machine learning algorithms for smishing detection using SMS text, such as Logistic Regression (LogR), Support Vector Machine (SVM), Multinomial Naive Bayes (MNB) Gaussian Naive Bayes (GNB), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Ada Boosting (AB), Extra Tree (ET) and Extreme Gradient Boosting (XGB).

III. RELATED WORKS

Detecting and preventing smishing attacks is essential for safeguarding individuals and organizations from potential harm. Effective methods for detecting smishing attempts have been developed using machine learning and deep learning (DL) approaches. The most recent findings from research conducted in this area are presented in this section.

Boukari et al. [12] proposed a fraud detection system that utilizes the TF-IDF method to convert SMS text into a vectorized representation. A dataset of 5000 emails was used

for smishing. The importance of each word in the SMS is determined using the TF-IDF algorithm. The experiment's findings indicate a high level of accuracy, with the RF algorithm achieving an accuracy of 98.15% and the NB approach achieving an accuracy of 90.59%. Mishra et al. [13] developed an efficient smishing detection system using an artificial neural network. The dataset utilized in the experiment consisted of 5858 messages, with 538 classifieds as smishing messages and 5320 as valid messages. Seven distinctive features were extracted from the dataset. The neural network was implemented with seven features, including email, URL, phone number, etc. The performance results for NB, DT, and Neural Network were 96.29 %, 93.40 %, and 97.40 %, respectively. Ulfath et al. [14] employed the N-grams and TF-IDF techniques for feature extraction and the statistical feature selection method. The vocabulary only includes the top 10,000 terms across the SMS dataset. The dataset was obtained from UCI's machine learning library. Five machine learning algorithms, XGB, DT, RF, SVM, and AB were employed. The outcome reveals that the SVM classifier obtains a higher accuracy of 98.39%. Smishing messages from various Internet sources were collected by Mishra et al. [15]. There are 638 smishing, 489 spam, and 4844 ham messages in the 5971 samples. Finally, five features were extracted and used in the experiment. ML algorithms such as NB, RF, and DT were employed. The performance outcomes for NB were 94.06%, RF 94.64%, and DT 93.96%. Maqsood et al. [16] used both ML and DL classifiers to detect spam SMS. The dataset was obtained via Kaggle, (SMS Spam Collection Dataset). The process of extracting features was conducted via the Bag-of-Words and TF-IDF methodologies. The gathered features served as input for three distinct machine learning models and CNN in deep learning. The SVM demonstrated a high level of accuracy, outperforming the other models with a performance of 99.6% for spam SMS. Ramanujam et al. [17] compiled a multiple languages SMS dataset, which encompasses both spam and non-spam messages in English and four distinct Indian languages. The SMS data contains 2,757 ham and 525 spam messages. A deep learning hybrid model (CNN-LSTM) has been proposed without feature engineering. The model exhibits a 97.7% accuracy rate. Jain et al. [18] presented a method to classify messages as smishing messages. The model was trained using multiple machine learning techniques with the Almeida spam data collection of 5169 messages, 362 of which are smishing and 4817 non-smishing. In addition, the model's accuracy has been determined for across datasets. The voting classifier that integrates KNN, RF, and ET Classifier (ETC) achieves an accuracy of 99.03% and a precision of 98.94%. Using machine learning, Mishra et al. [19] extracted the five most effective text message features for smishing detection. The dataset used for smishing detection contains 5858 text messages, 538 of which are smishing and 5320 valid. For the experimental results, RF, NB, DT, and Backpropagation Algorithm were used. The Backpropagation Algorithm outperformed the competition with a 97.93% accuracy rate. Sjarif et al. [20] used an algorithm incorporating TF-IDF and machine learning algorithms to identify SMS spam messages. The UCI Repository provided the dataset. The collection includes 5,574 English raw text messages categorized as ham or spam. Of these messages, 4,827 are classified as ham and the remaining 747 as spam. Five ML algorithms were used combined with the TF-IDF method for experimental purposes. The TF-IDF and RF combination produced an impressive 97.50% accuracy rate.

In existing research, feature extraction and vectorization from the SMS dataset are prioritized. Understanding the content of the message is essential for improved classification. The existing related works summary is shown in Table I.

TABLE I. RELATED WORKS SUMMARY

Author	Dataset	Feature generation method	Accuracy	Issues
Boukari et al. [12]	Smishing dataset 5000 messages	TF-IDF Vectorizer	Naïve Bayes: 90.59% Random Forest: 98.15%	Contextualization of text is not included.
Mishra et al. [13]	Smishing dataset Smishing messages=538 Legitimate messages=5320	Feature Extraction	Naïve Bayes: 96.29 Decision Tree: 93.40 Neural Network:97.40	A limited number of features. The system does not fully utilize the dataset.
Ulfath et al. [14]	UCI Dataset (SMS-Spam Collection) Spam messages=1197 Legitimate messages=4377	TF-IDF Vectorizer	Support vector machine: 98.39	Contextualization of text is not included.
Mishra et al. [15]	Smishing dataset Smishing messages=638, Spam messages=489 Legitimate messages=4844	Feature Extraction	Nave Bayes: 94.06, Random Forest: 94.64, Decision Tree: 93.96	The system is based on 5 features. The system does not fully utilize the dataset.
Maqsood et al. [16]	Kaggle- SMS Spam Collection Dataset. Spam messages=750, Legitimate messages=4250	Count (BoW) and TF-IDF Vectorizer	Support vector machine: 99.6	Contextualization of text is not included.
Ramanujam et al. [17]	Multilingual SMS Dataset. Legitimate messages=2,757, Spam messages=525	--	CNN-LSTM model: 97.7	Ne specific feature engineering was adopted.
Jain et al. [18]	Almeida spam data set. Spam Messages=362, Legitimate Messages=4817	TF-IDF Vectorizer	Voting classifier with KNN, RF, and ET Classifier (ETC) =99.03	Time-consuming method
Mishra et al. [19]	Smishing dataset Smishing messages=538 Legitimate messages=5320	Feature Extraction	Backpropagation Algorithm: 97.93	A limited number of features. The system does not fully utilize the dataset.
Sjarif et al. [20]	UCI Dataset Legitimate message=4,827 Spam messages= 747	TF-IDF Vectorizer	RF:97.50	Contextualization of text is not included.

IV. PROPOSED METHODS

Our proposed methods (SmishGuard) include two different systems for smishing (SMS phishing) detection that employ ML and NLP methods to improve mobile communications security. Most of the existing methodologies extract features from textual content for classification using ML or DL models. Generating features from textual data, such as counting the number of words or special characters in a document, provides structural insights and facilitates the representation of data. However, these features are not concerned with the context or semantics of the text, but rather with its structural characteristics. Sometimes, the generated features may not optimize the potential of the dataset. The primary goal of the proposed system is to maximize the utilization of the dataset through the implementation of several natural language processing methods, including TF-IDF with LDA and average Word2Vec with TF-IDF. NLP methods fully vectorizing text, such as TF-IDF or word embeddings, provide a richer grasp of the text's semantic content and make complicated analysis and machine learning tasks feasible. The outcomes of both methods exhibit a higher level of performance in comparison to the currently available methods. The process flow of the proposed system is shown visually in Fig. 1.

A. Dataset

For experiments, two distinct data sources are combined. Table II provides information about the data sources. Combining data from multiple sources necessitates the elimination of duplicate entries from the dataset. After removing duplicates, the final dataset used for our experiment is presented in Table III.

TABLE II. DATA SOURCES

Data Set	Features
SMS Phishing Dataset [21]	Smishing =1127 Legitimate messages=4844.
SMS Spam Collection [22]	Legitimate Messages=4825 Spam messages=747

TABLE III. COMBINED DATASET

Data Set	Features
SMS_PHISH	Smishing Messages =1627 Legitimate messages=5634

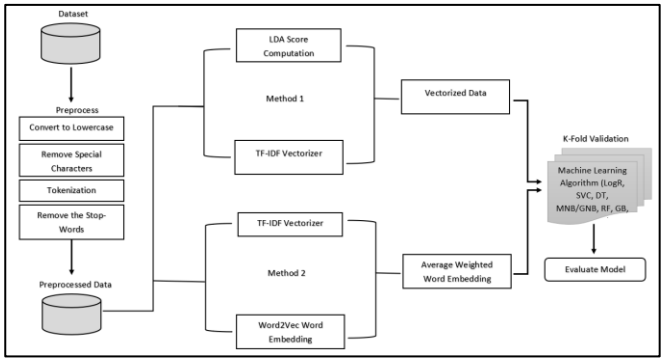


Fig. 1. Process flow of the proposed system.

B. Preprocessing

Text preprocessing is a crucial initial stage for Natural Language Processing and Machine Learning tasks that involve textual data. Effective text preprocessing is essential for maintaining data integrity and eliminating irrelevant information, enabling NLP and ML models to concentrate on extracting valuable insights and discerning patterns from textual data [23] [24]. Common procedures include lowercasing, tokenization, stop word elimination, and special character handling, etc. Algorithm 1 describes the preprocessing procedures.

Algorithm 1: Processing

1. Input: Raw Text
2. Output: Preprocessed Text
3. Import required libraries (nltk for natural language processing and re for regular expressions).
4. Load the raw text data.
5. Convert text to lowercase for consistency.
6. Remove any special characters, punctuation, and numerical values using regular expressions.
7. Break the text down into individual words or tokens.
8. Eliminate stop words as they do not contribute much to the meaning.
9. Join the tokens back into a preprocessed text string.

C. Method 1: TF-IDF Vectorization with LDA Topic Proportion Score (TF-IDF-LDA)

Combining TF-IDF vectorization and LDA text proportion scores is an effective approach. Through TF-IDF, text messages are transformed into numerical feature vectors that highlight the significance of phrases in each message, allowing machine learning algorithms to identify patterns suggestive of attempted smishing. Along with LDA topic proportion scores (The number of topics is set to 2 for our experiment) improve the identification accuracy of fraudulent messages. These scores provide a numerical representation of the subjects present in each message by applying LDA to SMS content. Smishing attempts often use certain language patterns or themes. LDA can reveal hidden motifs in smishing efforts, helping models distinguish between authentic and suspect communications. High proportions of topics relating to fraud, urgency, or deceptive content can serve as strong indicators of smishing, enabling the development of robust detection systems that protect users from phishing threats concealed within text messages.

Term Frequency (TF) for a word in a document [25]

$$\text{TF}(w, d) = \frac{n_{w,d}}{N_d}$$

$n_{w,d}$ = The frequency of occurrences of the term w in document d.

N_d = Total word count in document d.

Inverse Document Frequency (IDF) for a word in the corpus:

$$\text{IDF}(w) = log\left(\frac{\text{N}}{n_w}\right)$$

N= Number of documents contained in the corpus

n_w = Count of documents that contain the word w

TF-IDF Score for a word in the document is obtained by multiplying TF and IDF.

Topic Proportion Calculation for a Document [26]

For a given document D and topic T:

$$P(T|D) = \frac{P(T).P(D|T)}{P(D)}$$

where:

P(T|D) is the probability (proportion) of topic T in document D.

P(T) is the topic's prior probability over the entire corpus. It represents the proportion of documents within the entire dataset that are assigned to topic T.

P(D|T) is the probability of document D being generated from topic T. It determines the probability that the words in document D are generated by topic T.

P(D) is the probability of observing document D in the corpus.

D. Method 2: Weighted Average Word Embedding Generation

After preparing the raw textual contents, preprocessed data is prepared for the experiment. To determine the weight of words within a corpus through the utilization of TF-IDF vectorization, the initial step involves the computation of TF-IDF scores for every individual word encompassed within the entirety of the corpus. The TF-IDF metric quantifies the significance of a word within a particular document by considering its frequency in that document and inversely comparing it to its frequency across all documents. After obtaining these TF-IDF scores, the next step is to compute the weighted average Word2Vec embedding for machine learning using these scores as weights. The Word2Vec model is utilized to represent words as continuous vectors that capture semantic relationships. To calculate the weighted average Word2Vec embedding for a document, the vector of Word2Vec for each word is multiplied by its corresponding TF-IDF score. The result is a weighted vector. Then the sum of all weighted vectors of the document is divided by the sum of all weights of a document. Every document in the corpus finally gets a single weighted average word vector. This weighted average Word2Vec representation includes word importance and semantic context for downstream machine-learning tasks. The entire process is presented in Algorithm 2.

Weighted Average Word2Vec Computation for a document 'd' [27]

Multiply each word's Word2Vec vector by its corresponding TF-IDF score. The weighted average word embedding is calculated using the equation below:

Weighted Average Word2Vec (d) =

$$\frac{\sum_{w \in d} \left(TF - IDF_{(w,d)} \times Word2Vec\,(w) \right)}{\sum_{w \in d} \left(TF - IDF_{(w,d)} \right)}$$

where, w represents the word and d represents the document.

Where n is the number of words in a document.

Algorithm 2: Weighted Average Word2Vec

Input: TF-IDF matrix, Word2Vec embeddings

Output: Weighted Average Word2Vec representation for a document.

1. Multiply each word's Word2Vec vector by its TF-IDF score in the matrix.
2. Calculate the weighted average word embeddings for each of the document's words.

E. Training and Testing

Finally, K-fold validation trains and tests machine learning models using vectorized data/ word embedding. K-fold cross-validation is a widely employed technique utilized for the assessment of a model's performance and the mitigation of overfitting [28]. In this technique, the dataset is partitioned into K subsets, sometimes referred to as "folds," which are approximately of similar size. K-fold cross-validation trains and evaluates the model K times. For each iteration, one-fold is marked as the validation set, and K-1 folds are used for training. The performance metric is determined by calculating the average of K evaluation results.

V. Experiments and Results

The experiments employ Jupyter Notebook and Python's sklearn package on Windows 10. Accuracy, precision, recall, and the f1-score are performance parameters that are considered. Table IV presents the experimental result of the TF-IDF Vectorization with the LDA Topic Proportion Score (TF-IDF-LDA) method. Fig. 2 shows the ROC AUC curve.

The outcome demonstrates that the methods used in Random Forest and Extreme Gradient Boosting attains accuracy levels of 98.42% and 98.47%, respectively. Additionally, they obtained respective F1-Scores of 96.52% and 96.30%.

Table V presents the parameter details of the Word2Vec model. Table VI presents the experimental result of the weighted average word2vec (WAW2Vec) method. Fig. 3 shows the ROC AUC curve for the Random Forest and Extreme Gradient Boosting Algorithm.

TABLE IV. Performance of TF-IDF-LDA Method

ML Algorithm	Accuracy %	Precision %	Recall %	F1-Score %
LogR	96.97	95.26	91.03	93.07
SVC	98.33	98.33	94.16	96.17
MNB	97.52	98.58	90.22	94.19
DT	97.42	94.60	93.85	94.20
RF	**98.42**	**99.53**	**93.36**	**96.30**
GB	97.08	95.85	90.90	93.27
AB	97.22	94.97	92.50	93.70
XGB	**98.47**	**97.91**	**95.20**	**96.52**

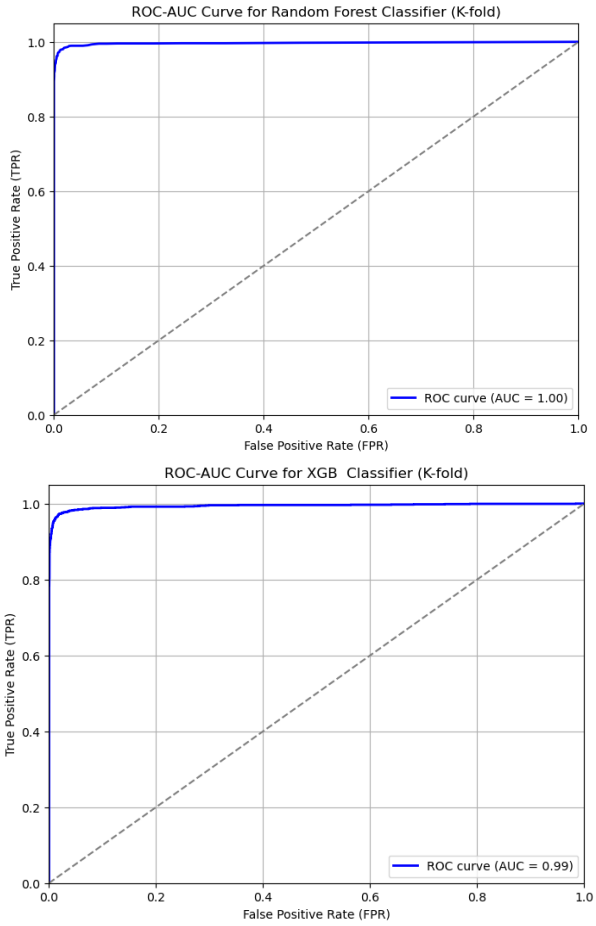


Fig. 2. ROC-AUC curve for TF-IDF-LDA method.

TABLE V. PARAMETER OF WORD2VEC MODEL

Parameter	Value
Vector Size	100
Window	5
Min_count	1
Sg	0

TABLE VI. PERFORMANCE OF WAW2VEC METHOD

ML Algorithm	Accuracy	Precision	Recall	F1-Score
LogR	86.08	85.35	90.07	87.34
SVC	91.23	91.23	92.55	91.78
GNB	82.83	87.58	79.18	82.53
DT	94.13	85.61	88.62	87.05
RF	**96.09**	**92.50**	**89.79**	**91.07**
GB	94.33	88.67	85.61	87.09
AB	93.04	84.94	83.83	84.34
XGB	**96.21**	**92.23**	**90.71**	**91.42**

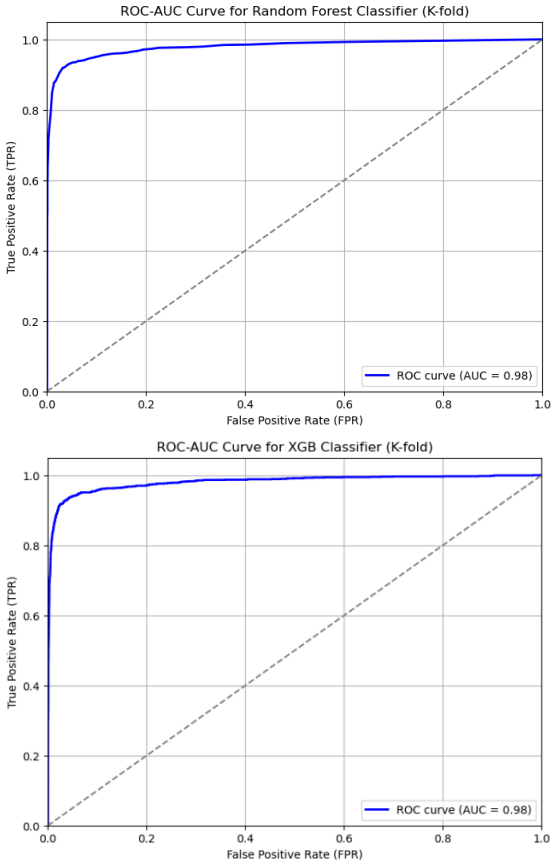
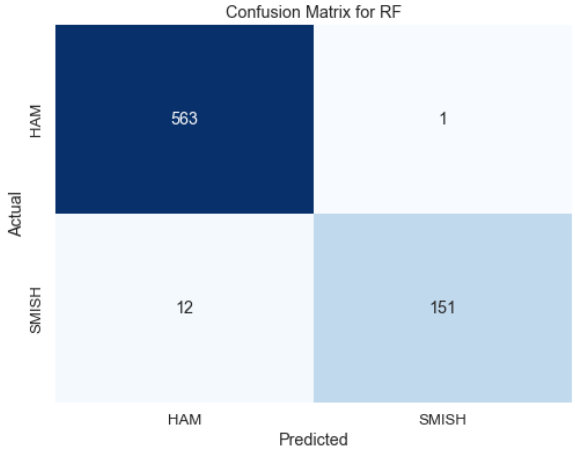


Fig. 3. ROC-AUC curve for WAW2Vec method.

The results indicate that Random Forest and Extreme Gradient Boosting obtained an accuracy of 96.09% and 96.22%, respectively. They also obtained respective F1-Scores of 91.07% and 91.42%.

Confusion matrix is crucial to classification model evaluation as it provides extensive insights into model performance. It divides the predictions and actual outcomes of the model into four categories: true positives, true negatives, false positives, and false negatives. Fig. 4 and Fig. 5 illustrate the performance of the proposed model by displaying fewer false positives and false negatives.



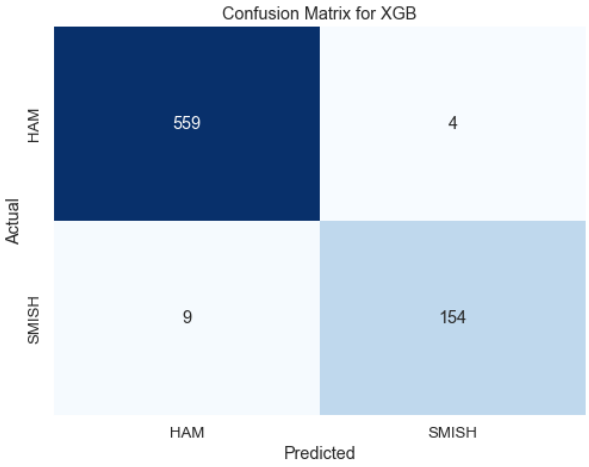


Fig. 4. Confusion matrix for TF-IDF-LDA method.

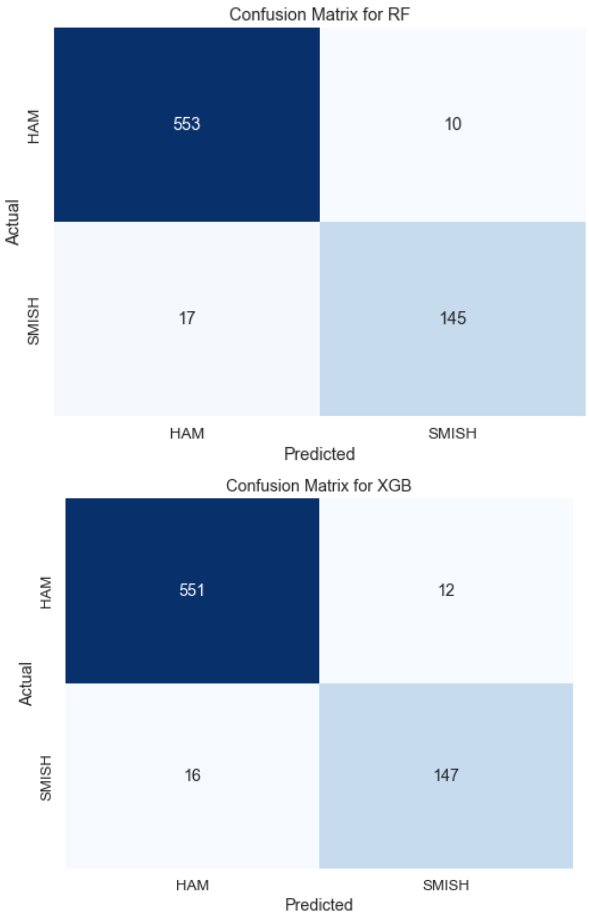


Fig. 5. Confusion matrix for WAW2Vec method.

The results demonstrate that the TF-IDF with LDA approach (Method-1) exhibits superior accuracy and F1-Score compared to the Weight Average Word2Vec method (Method-2). However, both methods produce better outcomes than existing methods as shown in Fig. 6.

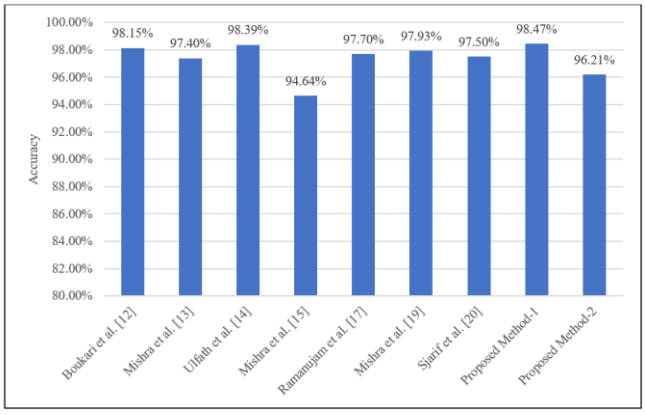


Fig. 6. Performance comparison (accuracy).

VI. Conclusion

Cybercriminals are increasingly using smishing, also known as SMS phishing, as a sneaky way to trick and deceive people and businesses. The expanding usage of mobile devices and text messaging for communication and transactions has increased the risk of smishing attacks. Attackers typically imitate reputable organizations to trick victims into disclosing critical information or committing crimes. So, smishing detection systems are vital for maintaining the security of vital infrastructure, protecting private and financial information, and maintaining public confidence in digital communication. Using ML and NLP methods for Smishing Detection is a promising way to counter the growing menace of SMS phishing attempts. This approach improves smishing detection by analyzing message content and contextual information.

This paper presents two methods based on ML algorithm and NLP methods for smishing detection. The results indicate that the TF-IDF with LDA approach (Method-1) achieves greater precision and F1-Score than the Weight Average Word2Vec technique (Method-2). Nevertheless, both methodologies yield outcomes that exhibit enhancements compared to current approaches. This empirical evidence is crucial to cybersecurity research since it refines methods and guides future study. Proposed method enhances the overall security of digital communication and transactions by contributing to the development of effective tools to counter evolving cyber threats.

Despite the potential of the proposed smishing detection system to mitigate the exponential growth of the smishing threat, the current system's capabilities are restricted to smishing messages in the English language. The task of managing diverse languages is progressively becoming more difficult.

References

[1] A.Kanaoka and T.Isohara, “Beyond Mobile Devices: A Cross-Device Solution for Smishing Detection and Prevention”, USENIX Symposium on Usable Privacy and Security, pp. 6–8, 2023.

[2] Smishing: https://dgc.org/en/smishing/ . (Last access: 20/9/2023).

[3] Malware: https://www.malwarebytes.com/what-is-smishing. (Last access: 20/9/2023).

[4] A.Mahmood and S.Hameed, "Review of Smishing Detection Via Machine Learning," Iraqi Journal of Science, 64(8), pp. 4244–4259, 2023. https://doi.org/10.24996/ijs.2023.64.8.42.

[5] A.Alhogail and A.Alsabih, "Applying machine learning and natural language processing to detect phishing email," Computers and Security, vol.1,10, 2021. https://doi.org/10.1016/j.cose.2021.102414.

[6] B.Sharma and P.Singh, "An improved anti-phishing model utilizing TF-IDF and AdaBoost," Concurrency Computation Practice and Experience. 34(26):e7287, 2022. https://doi.org/10.1002/cpe.7287.

[7] Mukesh, "TF-IDF Vectorizer scikit-learn," https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a. (Last access: 20/9/2023).

[8] E.Gualberto, R.Sousa, T.Vieira, J.Costa and C.Duque, "From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection," in IEEE Access, vol. 8, pp. 76368-76385, 2020. doi: 10.1109/ACCESS.2020.2989126.

[9] S.Lee, S.Kim, S.Lee, J.Choi, H.Yoon, D.Lee and J.Lee, "LARGen: Automatic Signature Generation for Malwares Using Latent Dirichlet Allocation," in IEEE Transactions on Dependable and Secure Computing, vol. 15, no. 5, pp. 771-783, 2018. doi: 10.1109/TDSC.2016.2609907.

[10] Word embedding: https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/. (Last access: 20/9/2023).

[11] J.Nabi, "Machine Learning Fundamentals," 2018 https://towardsdatascience.com/machine-learning-basics-part-1-a36d38c7916. (Last access: 20/9/2023).

[12] B.Boukari, A.Ravi and M.Msahli, "Machine Learning detection for SMiShing frauds," IEEE 18th Annual Consumer Communications & Networking Conference (CCNC), 2021.

[13] S.Mishra and D.Soni, "Implementation of 'Smishing Detector: An Efficient Model for Smishing Detection Using Neural Network," SN Computer Science, Springer. 2022. https://doi.org/10.1007/s42979-022-01078-0.

[14] R.Ulfath, I.Sarker, M.Chowdhury and M.Hammoudeh, "Detecting Smishing Attacks Using Feature Extraction and Classification Techniques," Proceedings of the International Conference on Big Data,IoT, and Machine Learning, Lecture Notes on Data Engineering and Communications Technologies, 95, 2022. https://doi.org/10.1007/978-981-16-6636-0_51

[15] S.Mishra and D.Soni, "SMS Phishing Dataset for Machine Learning and Pattern Recognition," Proceedings of 14th International Conference on Soft Computing and Pattern Recognition, Lecture Notes in Networks and Systems, 648, pp. 597–604, 2023. https://doi.org/10.1007/978-3-031-27524-1_57

[16] U.Maqsood, S.Rehman, T.Ali, K.Mahmood , T.Alsaedi and M.Kundi, "An Intelligent Framework Based on Deep Learning for SMS and e-mail Spam Detection," Applied Computational Intelligence and Soft Computing, 2023. https://doi.org/10.1155/2023/6648970.

[17] E.Ramanujam, K.Shankar and A.Sharma, "Multi-lingual Spam SMS detection using a hybrid deep learning technique," IEEE Silchar Subsection Conference (SILCON), pp. 1-6, 2022. doi: https://doi.org/10.1109/SILCON55242.2022.10028936

[18] A.K.Jain , B.B Gupta, K. Kaur, P.Bhutani, A.Almomani and W.Alhalabi, "A content and URL analysis-based efficient approach to detect smishing SMS in intelligent systems," International Journal of Intelligent Systems. 2022. https://doi.org/10.1002/int.23035

[19] S.Mishra and D.Soni, "DSmishSMS-A System to Detect Smishing SMS," Neural Computing and Applications. Springer. 35, pp. 4975–4992, 2023. https://doi.org/10.1007/s00521-021-06305-y.

[20] N.Sjarif, N.Azmi, S.Chuprat, H.Sarkan, Y.Yahya and S.Sam, "SMS Spam Message Detection using Term Frequency-Inverse Document Frequency and Random Forest Algorithm," Procedia Computer Science, vol. 161, pp. 509-515, 2019. https://doi.org/10.1016/j.procs.2019.11.150.

[21] S.Mishra and D.Soni, https://data.mendeley.com/datasets/f45bkkt8pr/1. (Last access: 20/9/2023).

[22] UCI Dataset: https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset. (Last access: 20/9/2023).

[23] Y.Kerner, D.Miller and Y.Yigal, " The influence of preprocessing on text classification using a bag-of-words representation," PLoS One. 1;15(5):e0232525, 2020. doi: 10.1371/journal.pone.0232525.

[24] A.Saleem, B.Sundarvadivazhagan, R.Vijayarangan and S.Veeramani, "Malicious Webpage Classification Based on Web Content Features using Machine Learning and Deep Learning, International Conference on Green Energy," Computing and Sustainable Technology (GECOST), Miri Sarawak, Malaysia, pp. 314-319, 2022. doi: 10.1109/GECOST55694.2022.10010386. [Included in IEEE Xplore]

[25] S.W.Kim and J.M.Gil, "Research paper classification systems based on TF-IDF and LDA schemes," Human-centric Computing and Information Sciences, vol. 9, no. 1, pp 1–21, 2019. https://doi.org/10.1186/s13673-019-0192-7

[26] J.Gan and Y.Qi, "Selection of the Optimal Number of Topics for LDA Topic Model-Taking Patent Policy Analysis as an Example," Entropy (Basel). 3;23(10):1301. 2021, doi: 10.3390/e23101301.

[27] A.Elsaadawy, M.Torki and N.Makky, "A Text Classifier Using Weighted Average Word Embedding," 2018 International Japan-Africa Conference on Electronics, Communications and Computations (JAC-ECC), Alexandria, Egypt, pp. 151-154, 2018. doi: 10.1109/JEC-ECC.2018.8679539.

[28] J.Brownlee, A Gentle Introduction to k-fold Cross-Validation, https://machinelearningmastery.com/k-fold-cross-validation/ (Last access: 20/9/2023).

Sleep Apnea Detection Method Based on Improved Random Forest

Xiangkui Wan, Yang Liu, Liuwang Yang, Chunyan Zeng, Danni Hao*
Hubei Key Laboratory for High-efficiency Utilization of Solar Energy and Operation Control of Energy Storage System, Hubei University of Technology, Wuhan, 430068, China

***Abstract*—Random forest (RF) helps to solve problems such as the detection of sleep apnea (SA) by constructing multiple decision trees, but there is no definite rule for the selection of input features in the model. In this paper, we propose a SA detection method based on fuzzy C-mean clustering (FCM) and backward feature rejection method, which improves the sensitivity and accuracy of SA detection by selecting the optimal set of features to input to the random forest model. Firstly, FCM clustering is performed on the RR interval features of ECG signals, and then the backward feature rejection method is used to combine the intra-cluster tightness, inter-cluster separation and contour coefficient metrics to eliminate redundant features to determine the optimal feature set, which is then inputted into the RF to detect SA. The experimental results of this method on Apnea-ECG database data show that the SA detection accuracy is 88.6%, sensitivity is 90.5%, and specificity is 85.5%, and the algorithm can adaptively select a smaller number of more discriminative features through FCM to reduce the input dimensions and improve the accuracy and sensitivity of the RF model for sleep apnea detection.**

***Keywords*—Sleep apnea; fuzzy c-means; backward feature elimination method; random forest**

I. INTRODUCTION

Sleep apnea (SA) leads to nocturnal hypoxia and hypercapnia, which makes elevated blood pressure and heart rate [1], and may lead to cardiovascular diseases such as hypertension, coronary artery disease, and cardiac arrhythmia, and even cause serious consequences such as heart failure and sudden death [2-3]. Studies have shown that SA is also associated with neurological disorders such as Alzheimer's, Parkinson's disease and depression. Owen, et al. first identified Alzheimer's-like amyloid plaques in the brains of people who are clinically proven to have obstructive SA [4]. Patients with SA in middle age are more likely to develop Alzheimer's disease in old age [5]. Parkinson's disease patients have degeneration in the brain stem area that controls breathing, which can cause reduced respiratory muscle function, and sleep-disordered breathing. Yang, et al. propose a method to detect Parkinson's disease and predict disease severity by breathing at night [6]. SA can also cause neurasthenia, which affects the classification of autistic patients using electroencephalography [7]. Therefore, timely, accurate and convenient detection of SA is of great significance.

Scholars have carried out a lot of exploration and research in SA monitoring. Tagluk, et al. proposed a SA detection method based on wavelet transform and artificial neural network [8], which utilizes multi-resolution wavelet transform to decompose the abdominal breathing signal into multiple spectral components, and these spectral components are inputted into the artificial neural network to classify the SA condition of patients. However, detecting abdominal breathing often requires the use of larger devices, which impacts the patient's daily life. Mendez, et al. investigated a method for detecting SA based on empirical mode decomposition (EMD) and wavelet analysis (WA) of ECG signals [9], whereby features are extracted from the decomposition results, a heart rate variability time-domain measure and three additional nonlinear measures are used as inputs to a linear discriminant classifier. However, this method requires lot input feature parameters and complex computational model, thus the robustness needs to be improved. Iwasaki, et al. analyzed the adjacent R-wave intervals in the ECG signals and used a long and short-term memory model to detect SA [10]. Urtnasan, et al. proposed a deep learning architecture based on a convolutional neural network using a single-lead ECG signal for SA classification [11]. However, such deep learning models usually require multiple experiments to get the algorithm parameters and lack of interpretability.

Among the commonly used machine learning algorithms, the random forest (RF) algorithm has been widely used in biomedical signal processing due to its ability to handle high-dimensional data, capture nonlinear relationships, reduce the risk of overfitting, and possess the advantages of noise immunity [12-13]. The RF algorithm can also help to solve the problems of signal classification and anomaly detection [14-15]. However, the RF algorithm has no definite rules for input features selection. Although the algorithm itself measures the contribution of each feature by means of feature importance assessment [16], this is only a relative metric, which does not reflect which features are absolutely important and should be selected. Selecting appropriate features is a relatively subjective process that depends on the specific dataset and problem domain [17].

To address the problem of how to select appropriate input features, this paper proposes a SA detection method based on Fuzzy C-means clustering (FCM) and backward feature rejection method, which selects appropriate input features through FCM clustering index, reduces the dimensionality of the input features, reduces the complexity of model training and prediction, and avoids too much noise and irrelevant information from causing model Interference. Feature selection allows the model to focus more on important features, thus better capturing patterns and relationships in the data and

*Corresponding Author.

improving the accuracy and sensitivity of the RF model for SA detection.

The focus of this paper is to improve the random selection process of input features in the traditional RF method. FCM was applied to the input features, and the index of intra-cluster tightness, inter-cluster separation and contour coefficient of the samples were calculated. Combined with the method of reverse feature elimination, redundant features are eliminated to determine the optimal selection. Better classification accuracy of SA can be obtained by using only a small number of distinct features. The structure of this paper is as follows: Section II describes the data set and the method used in this paper, Section III is the experimental results and analysis, Section IV is the discussion, and Section V is the conclusion.

II. Materials and Methods

A. Datasets

Apnea-ECG database [18]: the database consists of 70 records divided into a training set containing 35 records (a01 to a20, b01 to b05 and c01 to c10) and a test set of 35 records (x01 to x35). The individual records in the database range in length from seven to ten hours and are sampled at a frequency of 100 Hz. Each record consists of the ECG signal, a set of manual apnea annotations and a set of machine-generated QRS annotations. In addition, eight records (a01 to a04, b01, and c01 to c03) were accompanied by four additional signals: chest and abdominal respiratory signals obtained using respiratory inductive plethysmography, oronasal airflow and oxygen saturation recorded with nasal thermistor.

B. Extracting ECG Signal Features

SA causes changes in the autonomic nervous system and cardiovascular regulation, leading to prolongation or shortening of the RR interval [19]. In this paper, the Pan-Tompkins algorithm was adapted to identify QRS wave clusters [20], and the RR interval was determined by the time difference between adjacent R peaks, and the feature information was extracted from the RR interval sequence.

QRS wave detection: the raw ECG signals were processed in one-minute segments according to the annotation file. Per-segment SA detection determines whether each one-minute segment is SA or normal, and it is an important basis for SA diagnosis in suspected patients [21]. Band-pass filtering, differential amplification, squaring operation, moving window integration and threshold detection were performed on each segment using the Pan-Tompkins algorithm to locate the R-wave, as shown in Fig. 1.

Feature extraction: combined with the results of literature [22-23], we per-formed feature extraction on the obtained RR intervals as shown in Table I.

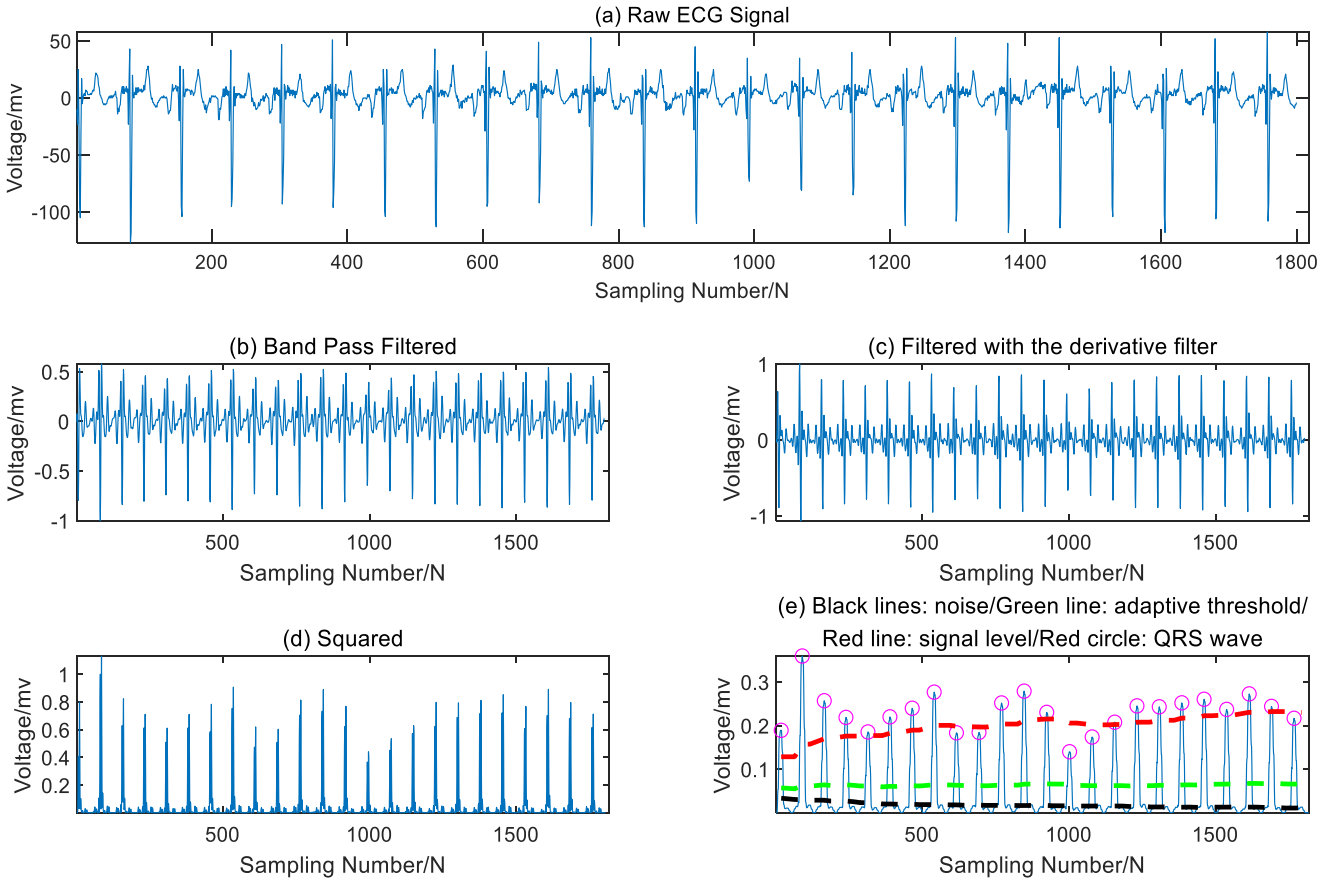


Fig. 1. QRS wave detection of ECG signals.

TABLE I. LIST OF RR INTERVAL FEATURES

No	Feature	Description
1	AVRR	$AVRR = \frac{1}{N}\sum_{j=1}^{N} RR_j$ (1) The average of all RR intervals.
2	SDRR	$SDRR = \sqrt{\frac{1}{N-1}\cdot\sum_{j=1}^{N}\left(RR_j - \overline{RR}\right)^2}$ (2) The standard deviation of all RR intervals.
3	RMSSD	$RMSSD = \sqrt{\frac{1}{N-1}\cdot\sum_{j=1}^{N-1}\left(RR_{j+1} - RR_j\right)^2}$ (3) The root mean square value of the difference between all neighboring RR intervals.
4	PRR	$PRR = max(RR_1、RR_2、\ldots、RR_n)$ (4) The peak value of all RR Intervals.
5	PNN50	$PNN50 = \frac{NN50}{N}\cdot 100$ (5) Percentage of the number of heartbeats where the difference between two neighboring RR intervals is greater than 50ms.
6	KRR	$KRR = \frac{diff(AVRR)}{diff(SDRR)}$ (6) The degree of bias and kurtosis of the RR interval signal.
7	SDHR	$SDHR = \frac{1}{N}\cdot\sqrt{\sum_{j=1}^{N}\left(60/RR_j - \overline{HR}\right)^2}$ (7) The standard deviation of heart rate.
8	LHFHRatio	$LFHFRatio = \frac{LF}{HF}$ (8) LF：Total power of 0.04 to 0.15Hz; HF：Total power of 0.15 to 0.4Hz.

C. Determining the Best Subset of Features

There is redundancy among the RR interval features, which affects the accuracy of classification [24-25], therefore feature selection is needed before classification. The backward feature elimination method used in this paper is a kind of greedy algorithm [26], which and obtains a feature set that has the smallest number of features and the highest correct classification rate. The specific process of the method is described as follows:

- Initialization: Determine the complete feature set containing all features.
- Random feature elimination: Randomly eliminate one feature from the feature set to form a new feature set.
- FCM feature clustering [27]: It make the new feature sample into n fuzzy clusters $X = \{x_1, x_2, \ldots, x_n\}$. Utilizing the fuzzy cluster's degrees of membership ranging from 0 to 1, iteratively optimize the objective function S to find the minimum value.

$$S(X, u_1, u_2, \ldots, u_c) = \sum_{k=1}^{n}\sum_{i=1}^{c}(v_{ik})^m (d_{ik})^2 \quad (9)$$

$$\begin{cases} u_i = \sum_{k=1}^{n}(v_{ik})^m x_k \,/\, \sum_{k=1}^{n}(v_{ik})^m \\ v_{ik} = 1/\sum_{i=1}^{n}(\frac{d_{ik}}{d_{jk}})^{\frac{2}{m-1}} \\ \sum_{i=1}^{n} v_{ik} = 1 \\ 0 \le v_{ik} \le 1 \\ d_{ik} = \|x_k - u_i\| \end{cases} \quad (10)$$

- u_i denotes the clustering centroid of the ith class; v_{ik} denotes the degree of membership between the kth sample and the ith class; d_{ik} denotes the euclidean distance between the center of the ith class and sample k. Compare the degree of membership $v_{ik}^{(b)}$ and $v_{ik}^{(b+1)}$, if $\|v_{ik}^{(b+1)} - v_{ik}^{(b)}\| \le \varepsilon$ (the given sensitivity threshold), it means that the objective function S has reached the minimal value, and the final clustering result has been obtained; otherwise, continue iterating until the convergence condition is satisfied.
- Calculate the clustering metrics: calculate the average intra-cluster compactness (AIC), average inter-cluster separation (AIS) and average silhouette coefficient (ASC) for each cluster.
- AIC indicates the degree of compactness of the sample points within the FCM clusters. The lower the value, the more compact the sample points within the clusters.

$$\begin{cases} AIC = \frac{1}{K} * \sum_{n=1}^{K} (IC)_n \\ IC = \frac{1}{N-1} * \sum_{i \ne j}^{N} d(i,j) \end{cases} \quad (11)$$

- Where N is the number of samples in the cluster, K is the number of clusters, and $d(i,j)$ is the distance between the ith sample and the jth sample, and IC denotes the average distance of each data point to the clustering center.
- AIS measures the separation between different clusters, and higher values indicate higher separation between different clusters, clearer boundary between clusters.

$$\begin{cases} AIS = \frac{1}{K} * \sum_{n=1}^{K} (IS)_n \\ IS = \frac{2}{K(K-1)} * \sum_{i \ne j}^{N} d(Ci, Cj) \end{cases} \quad (12)$$

- Here, $d(Ci, Cj)$ is the distance between the center of the ith cluster and the center of the jth cluster, and IS denotes the average distance between two clustering centers.
- AIS measures the separation between different clusters, and higher values indicate higher separation between different clusters, clearer boundary between clusters.

$$ASC = \frac{1}{M} * \sum_{i=1}^{M} \frac{IS(i) - IC(i)}{max\{IS(i), IC(i)\}} \quad (13)$$

- Here, M is the total number of samples, $IS(i)$ average distance between sample point i and all the samples in the nearest cluster, and $IC(i)$ denotes the average

distance between sample point i and the other samples within the same cluster.

- Determine the best feature subset: It compare the clustering metrics before and after the removal of features, if the metrics after the removal of features are increased, then continue to perform steps (2) ~ (4); otherwise, keep the feature. Go through remaining features until all the corresponding metrics are reduced, then stop the search process and determine the set of features before removal as the best feature subset.
- The overall flow chart of the backward feature elimination method is shown in Fig. 2.

D. Detecting Sleep Apnea

The best feature subset determined from the last step was used as input to the RF classifier, and the classification accuracy of SA was calculated using the 10-fold cross validation method and the confusion matrix.

- Extraction of the best feature subset. From the best feature subset D, n samples (sub-training set) are drawn randomly with put-back to form a new training set as samples at the root node of the decision tree. The remaining samples form the out-of-bag dataset (OOB) as the final test set.
- Attribute selection. Assume each sample has M attributes, m attributes are randomly selected from these M attributes when each node of the decision tree needs to be split (where m is less than M).
- Calculate the Gini index and node splitting. Assume A is an attribute of data set D. Attribute A has k different values $\{a_1, a_2, \dots a_k\}$. Under the condition $A = a_j (j = 1,2, \dots, k)$, the dataset D is partitioned into two parts D_1 and D_2, and the Gini index of this partition is:

$$Gini(D, A = a_j) = \frac{|D_1|}{n}\left(1 - \sum_{i=1}^{m}\left(\frac{|D_1|}{n}\right)^2\right) + \frac{|D_2|}{n}\left(1 - \sum_{i=1}^{m}\left(\frac{|D_2|}{n}\right)^2\right) \quad (14)$$

- Select the attribute with the smallest Gini index and its corresponding splitting node as the optimal attribute and optimal splitting node, generate two child nodes, and distribute the remaining training data into the two child nodes.
- Construct a random forest. Repeat step (3) in the sample subset of each child node, and recursively perform node splitting until all leaf nodes are generated; repeat (2) to (4) to obtain k different decision trees.
- Sleep apnea detection. Each decision tree performs a 10-fold cross validation calculation for each piece of data in the test set, and the category with the most votes in k classification results is the final category for that sample.

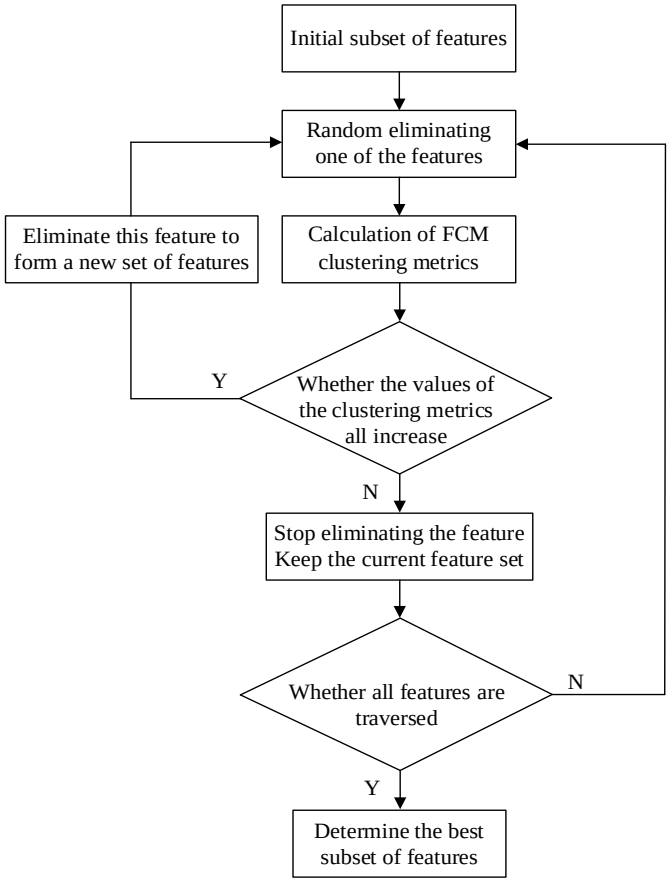


Fig. 2. Flowchart of backward feature elimination method.

E. Evaluation Criteria

The performance of the method was evaluated by calculating the metrics of accuracy, sensitivity and specificity for SA detection through a confusion matrix.

$$Accracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (15)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (16)$$

$$Specificity = \frac{TN}{TN+FP} \quad (17)$$

Where TP denotes the number of true samples classified as positive (true positive); TN denotes the number of true samples classified as negative (true negative); FP denotes the number of false samples classified as positive (false positive); and FN denotes the number of false samples classified as negative (false negative).

III. Results

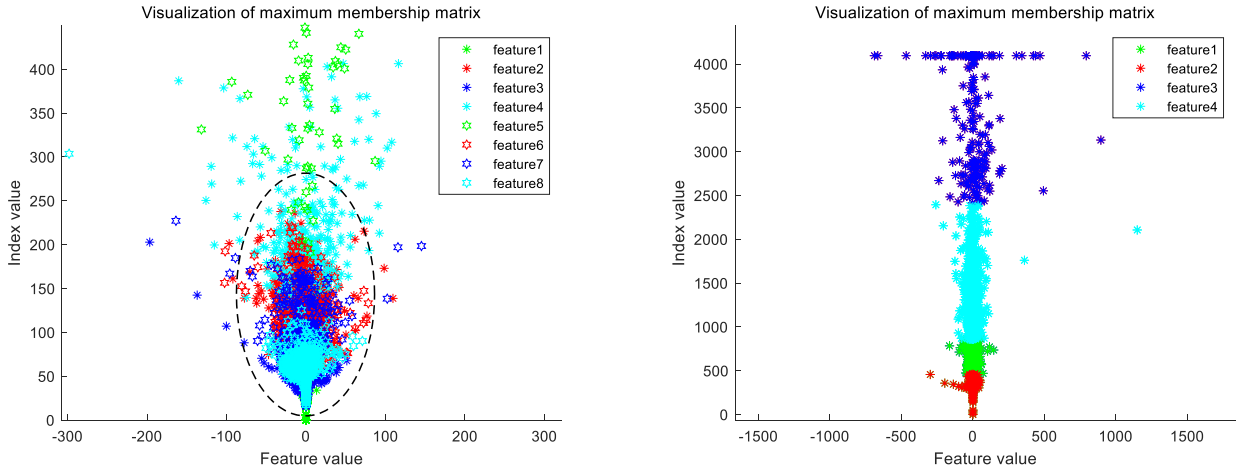
In this paper, 70 records from the Apnea-ECG database are used as experimental samples, and each ECG signal of these records is segmented into one-minute segment by annotation file. The RR interval features of ECG signals are extracted by Pan-Tompkins algorithm. The best subsets of features are selected using the backward feature elimination method as: AVRR, RMSSD, PRR and KRR.

In Fig. 3, by visualizing the membership matrix U, the membership distribution of the data points between the clusters before and after the removal of features are illustrated. The data membership distribution in Fig. 3(a) is more centralized, and the feature points have fuzzy attribution relationships among multiple clusters, which make it difficult to be clearly classified into specific clusters. In contrast, the distribution of data membership in Fig. 3(b) is more dispersed, and the attribution of feature points is more explicit and differentiated.

Features with greater divergence are more favorable, since they allow the classifier capture and represent the characteristic patterns of different types of SA events with higher accuracy.

The average intra-cluster tightness, average inter-cluster separation, and average contour coefficient were calculated before and after the removal of the features, and the two sets of features were input into RF classifier to evaluate the SA detection accuracy, as shown in Table II, respectively:

The above four RR period features were matched with apnea labels to reconstruct the database as input to the RF classifier, and the classification accuracy of SA detection is calculated using the 10-fold cross validation method and the confusion matrix. The performance comparison with existing studies using the Apnea-ECG dataset is given in Table III.



(a) Visualization of the affiliation matrix for 8 features

(b) Visualization of the affiliation matrix for 4 features

Fig. 3. Visualization results of the affiliation matrix of RR interval features.

TABLE II. Comparison of Clustering Metrics and SA Classification Performance before and after Feature Removal

Subset of features	Feature number	AIC	AIS	ASC	Acc (%)	Sen (%)	Spe (%)
Features before elimination	8	98.42	232.65	0.58	87.7	89.1	85.3
Features after elimination	4	190.44	816.17	0.77	88.6	90.5	85.5

TABLE III. Comparison of SA Detection Accuracy of Different Methods

Method	Feature number	Acc (%)	Sen (%)	Spe (%)
WA+HT [9]	40	90.5	84.9	93.7
EMD+RAS [9]	20	88.9	80.6	94.4
MLP [28]	18	81.4	74.3	85.7
SVM [29]	11	85.6	79.1	88.9
Proposed method	4	88.6	90.5	85.5

IV. DISCUSSION

As shown in Table II, the feature clustering metrics are all increased by using backward feature elimination, indicating that the optimal feature subset has greater variability and differentiation, which is consistent with the visualized membership matrix in Fig. 3 Inputting the features after elimination to the RF classifier yields a classification with higher accuracy, sensitivity, and specificity, it is shown that the method adaptively removes irrelevant features and reduces the dimensionality of the input features so that the RF classifier can better capture patterns and relationships in the data and improve the accuracy of the detection of SA.

Table III compares the SA detection results of different methods. Although the method based on WA and Hilbert transform can achieve the detection accuracy of up to 90.5%, the number of features used in it is as high as 40, which undoubtedly increases signal pre-processing process and the overall algorithm complexity. The methods based on EMD and redistributed spectra require half the number of features to achieve similar results, indicating that the nonlinear features calculated by the former do not play an important role in classification [9]. The number of hidden layers and neurons in the Multilayer Perceptron classifier depends on experience, and the detection accuracy of sleep apnea is only 81.4%. Support Vector Machine also has problems with optimization of regularization coefficient and kernel function parameters. Compared with the above method, the number of features used by SA is the least, but the detection performance is not ideal.

Under the same dataset, the accuracy of our method for SA detection is 88.6%, sensitivity is 90.5% and specificity is 85.5%. It outperforms other existing methods in terms of sensitivity, high sensitivity means that the model can detect as many true apnea events as possible, reducing the possibility of underreporting, and is comparable to WA and EMD methods in terms of accuracy. This method reduces the computation and storage requirements by selecting fewer and more discriminative features using the FCM method. Moreover, it improves the model's sensitivity to key information and enhances the robustness of the model.

The discussion and evaluation of proposed algorithm on different data sets will be carried out in the future. Driving by the need of integrating the algorithm into wearable devices, how to further improve the running speed and efficiency of the algorithm is our improving direction.

V. CONCLUSION

In this paper, an improved RF-based SA detection method is proposed. FCM and backward feature elimination method are used to select the RR interval features of ECG signals. And a small number of the best feature subsets with obvious differences are obtained as inputs to the RF classifier, which improve the sensitivity of the RF model to key information and obtain a better SA detection accuracy.

REFERENCES

[1] Q. M. Sun, L. Xing, C. Wang, W. Liang, “Cardiopulmonary coupling analysis predicts early treatment response in depressed patients: A pilot study”, Psychiatry. Res, 2019, 276, 6-11, DOI: 10.1016/j.psychres.2019.04.002.

[2] A. B. Newman, F. J. Nieto, U. Guidry, “Relation of sleep-disordered breathing to cardiovascular disease risk factors: the Sleep Heart Health Study”, Am. J. Epidemiol, 2001, 154(1), 50-59, DOI: 10.1164/ajrccm.163.1.2001008.

[3] A. Abdullah, S. Nithya, M. Mary Shanthi Rani, S. Vijayalakshmi, B. Balusamy, “Stacked LSTM and Kernel-PCA-based Ensemble Learning for Cardiac Arrhythmia Classification”, Int. J. Adv. Comput. Sci. Appl, 2023, 14, 9, DOI:10.14569/IJACSA.2023.0140905.

[4] O. M. Bubu, A. G. Andrade, O. Q. Umasabor-bubu, “Obstructive sleep apnea, cognition and Alzheimer's disease: A systematic review integrating three decades of multidisciplinary research”, Sleep. Med. Rev, 2020, 50, 101250, DOI: 10.1016/j.smrv.2019.101250.

[5] J. E.Owen, B. Benediktsdottir, E. Cook, I. Olafsson, T. Gislason, S.R. Robinson, “Alzheimer's disease neuropathology in the hippocampus and brainstem of people with obstructive sleep apnea”, Sleep. 2021, 44(3), zsaa195, DOI: 10.1093/sleep/zsaa195.

[6] Y. Z. Yang, Y. Yuan, G. Zhang, “Artificial intelligence-enabled detection and assessment of Parkinson’s disease using nocturnal breathing signals”, Nat. Med. 2022, 28, 2207-2215, DOI: 10.1038/s41591-022-01932-x.

[7] L. Y. Wu, “Classification of Coherence Indices Extracted from EEG Signals of Mild and Severe Autism” , Int. J. Adv. Comput. Sci. Appl, 2023, 14, 9, DOI: 10.14569/IJACSA.2023.0140903.

[8] M. E. Tagluk, N. Sezgin, “Classification of sleep apnea through sub-band energy of abdominal effort signal using Wavelets + Neural Networks”, J. Med. Syst, 2010, 34(6), 1111-1119, DOI: 10.1007/s10916-009-9330-5.

[9] M. O. Mendez, J. Corthout, S. V. Huffel, T. Penzel, S. Cerutti, A. M. Bianchi, “Automatic screening of obstructive sleep apnea from the ECG based on empirical mode decomposition and wavelet analysis”, Physiol. Meas, 2010, 31(3), 273-289, DOI: 10.1088/0967-3334/31/3/001.

[10] A. Iwasaki, C. Nakayama, K. Fujiwara, Y. Sumi, M. Matsuo, M. Kano, H. Kadotani, “Screening of sleep apnea based on heart rate variability and long short-term memory”, Sleep Breath, 2021, 25(4), 1821-1829, DOI: 10.1007/s11325-020-02249-0.

[11] E. Urtnasan, J. U. Park, K. J. Lee, “Multiclass classification of obstructive sleep apnea/hypopnea based on a convolutional neural network from a single-lead electrocardiogram”, Physiol Meas, 2018, 39(6), 065003, DOI: 10.1088/1361-6579/aac7b7.

[12] L. Breima, “Random Forests”, Mach. Learn, 2001, 45, 5-32, DOI: 10.1023/A:1010933404324.

[13] X. F. Lv, J. B. Li, “A Method of Detecting Apnea Using Random Forest”, Beijing Youdian Daxue Xuebao, 2020, 43(5), 64-70, DOI: 10.13190/j.jbupt.2019-255.

[14] Y. L. Niu, J. Liu, Y. D. Zhu, X. Li, J. Li, F. L. Feng, “Sleep apnea detection method based on heart rate variability signal”, Shaanxi Shifan Daxue Xuebao (Ziran Kexueban), 2020, 48(6), 26-32, DOI: 10.15983/j.cnki.jsnu.2020.02.033.

[15] F. Hajipour, M.J. Jozani, Z. Moussavi, “A comparison of regularized logistic regression and random forest machine learning models for daytime diagnosis of obstructive sleep apnea”, Med. Biol. Eng. Comput, 2020, 58(10), 2517-2529, DOI: 10.1007/s11517-020-02206-9.

[16] T. Y. Cao, “Study on the Importance of Variables Based On Random Forest”, Stat. decis, 2022, 38(4), 60-63, DOI: 10.13546/j.cnki.tjyjc.2022.04.011.

[17] D. J. Yao, J. Yang, X. J. Zhan, “Feature selection algorithm based on random forest”, Jilin Daxue Xuebao (Gongxueban), 2014, 44(1), 137-141, DOI: 10.13229/j.cnki.jdxbgxb201401024.

[18] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, J. H. Peter, “The Apnea-ECG Database”, Comp. cardiol., 2000, 27, 255-258, DOI: 10.1109/CIC.2000.898505.

[19] R. W. Logan, C. Mcclung, “A Rhythms of life: circadian disruption and brain disorders across the lifespan”, Nat. Rev. Neurosci, 2019, 20(1), 49-65, DOI: 10.1038/s41583-018-0088-y.

[20] J. Pan, W. J. Tompkins, “A real-time QRS detection algorithm”, IEEE Trans Biomed Eng, 1985, 32(3), 230-236, DOI: 10.1109/TBME.1985.325532.

[21] M. Sakuna, K. Mekhora, W. Jalajondeja, C. Jalajondeja, "Breathing retraining with chest wall mobilization improves respiratory reserve and decreases hyperactivity of accessory breathing muscles during respiratory excursions: A randomized controlled trial", Acta. Bioeng. Biomech., 2020, 22, 153-159, DOI: 10.37190/ABB-01641-2020-03.

[22] T. Wang, C. Lu, G. Shen, "Detection of Sleep Apnea from Single-Lead ECG Signal Using a Time Window Artificial Neural Network", Biomed. Res. Int., 2019, 2019, 9768072, DOI: 10.1155/2019/9768072.

[23] A. I. Sharaf, "Sleep Apnea Detection Using Wavelet Scattering Transformation and Random Forest Classifier", Entropy, 2023, 25(3), 399, DOI: 10.3390/e25030399.

[24] H. K. Zhang, Y. Z. Cheng, T. Y. Zhang, "Research on ambulatory arterial stiffness index estimation using random forest model", J. Biomed. Eng. Res., 2022, 41(1), 55-61, DOI: 10.19529/j.cnki.1672-6278.2022.01.09.

[25] J. J. Lu, "Classifying Model of Ancient Glass Products Based on Ensemble Feature Selection and Random Forest", J. Chin. Ceram. Soc, 2023, 51(4), 1060-1065, DOI : 10.14062/j.issn.0454-5648.20220790.

[26] D. Wang, C. N. Liu, Y. Zeng, T. Tian, Z. Sun, "Dryland Crop Classification Combining Multitype Features and Multitemporal Quad-Polarimetric RADARSAT-2 Imagery in Hebei Plain, China", Sens., 2023, 44(3), 222-232, DOI : 10.3390/s21020332.

[27] E. H. Ruspini, "Numerical methods for fuzzy clustering", Inf. Sci., 1970, 2(3), 319-350, DOI: 10.1016/S0020-0255(70)80056-1.

[28] Y. D. Zhang, Y. Sun, P. Phillips, G. Liu, X. X. Zhou, S. H. Wang, "A "Multilayer Perceptron Based Smart Pathological Brain Detection System by Fractional Fourier Entropy", J. Med. Syst., 2016, 40(7), 173, DOI: 10.1007/s10916-016-0525-2.

[29] L. Almazaydeh, K. Elleithy, M. Faezipour, "Obstructive sleep apnea detection using SVM-based classification of ECG signal features", Annu. Int. Conf. IEEE. Eng. Med. Biol. Soc, 2012, 4938-4941, DOI: 10.1109/EMBC.2012.6347100.

Graph Anomaly Detection with Graph Convolutional Networks

Aabid A. Mir, Megat F. Zuhairi, Shahrulniza Musa
Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Kuala Lumpur, Malaysia

***Abstract*—Anomaly detection in network data is a critical task in various domains, and graph-based approaches, particularly Graph Convolutional Networks (GCNs), have gained significant attention in recent years. This paper provides a comprehensive analysis of anomaly detection techniques, focusing on the importance and challenges of network anomaly detection. It introduces the fundamentals of GCNs, including graph representation, graph convolutional operations, and the graph convolutional layer. The paper explores the applications of GCNs in anomaly detection, discussing the graph convolutional layer, hierarchical representation learning, and the overall process of anomaly detection using GCNs. A thorough review of the literature is presented, with a comparative analysis of GCN-based approaches. The findings highlight the significance of graph-based techniques, deep learning, and various aspects of graph representation in anomaly detection. The paper concludes with a discussion on key insights, challenges, and potential advancements, such as the integration of deep learning models and dynamic graph analysis.**

***Keywords*—Anomaly detection; deep learning; dynamic graphs; Graph Convolutional Networks (GCNs); Graph Neural Networks (GNNs); network data**

I. Introduction

Graph anomaly detection has gained significant attention in various domains, including insider threat detection, fraud detection, and network security [1]. The increasing prevalence of complex network data, such as social networks, financial transactions, and blockchain networks, has posed challenges for traditional anomaly detection approaches in capturing the inherent structural dependencies and contextual information encoded in graph-structured data [2]. As a result, there has been a growing interest in leveraging deep learning techniques, particularly graph convolutional networks (GCNs), to address these limitations and achieve more effective anomaly detection [3]. The primary objective of graph anomaly detection is to identify abnormal patterns, behaviors, or entities within a given graph. This involves analyzing the connections, relationships, and attributes of the nodes and edges in the graph to distinguish between normal and anomalous instances [4]. Deep learning-based approaches, especially GCNs, have shown promising results in capturing the complex dependencies and learning meaningful representations from graph-structured data [2].

In this paper, we aim to provide a comprehensive introduction to graph anomaly detection, with a particular focus on GCN-based methods. We will discuss the foundational concepts and techniques in anomaly detection, highlighting the unique challenges posed by graph-structured data. Additionally, we will look into the advancements made in the field, with a specific focus on the utilization of GCNs for modeling and analyzing graphs. We will explore how GCNs leverage graph convolutions to propagate information between nodes, enabling them to capture both local and global structural information.

This study analyzes the field of graph-based anomaly detection in network data in the subsequent sections, beginning with the background and motivation in Section II. Section III explores the details of anomaly detection in network data, establishing the foundation for an in-depth comprehension of its complexities. Section IV then expands on the fundamental backbone of this study, Graph Convolutional Networks (GCNs), providing insights into its structure and pivotal role in anomaly detection. Section V, describes the methodology used for analysis. Section VI then presents a comprehensive overview of the existing literature, focusing on GCN-based techniques and their comparative analysis. Section VII summarizes the findings and conclusions obtained from the review and the discussion section VIII discusses the essential insights, limitations, and issues encountered in graph-based anomaly detection, answering the three research questions of this study. Section IX provides a concise summary of the findings and future research prospects. Finally, in Section X, the conclusion summarizes the findings and implications established during the study.

II. Background and Motivation

Graph anomaly detection has emerged as a critical task in various domains, including network security, fraud detection, and anomaly monitoring in dynamic systems. Traditional methods often rely on handcrafted features or statistical techniques, which may lack the ability to capture complex patterns and hidden anomalies. In recent years, the advent of deep learning and graph convolutional networks (GCNs) has provided new opportunities for more effective and automated graph anomaly detection [1] [5].

The motivation behind this research stems from the growing need to develop advanced techniques that can effectively detect anomalies in complex graph-structured data. With the increasing scale and complexity of real-world networks, there is a pressing demand for anomaly detection methods that can handle large-scale graphs and capture intricate relationships between entities [4] [45]. By leveraging the power of GCNs and deep learning, it is possible to extract high-level representations from graphs and capture both local and global patterns, leading to more accurate and robust anomaly detection [6].

III. Anomaly Detection in Network Data

A. Overview of Anomaly Detection Techniques

Anomaly detection is a critical task in network data analysis, aiming to identify abnormal patterns or behaviors that deviate from the expected norm. Various techniques have been proposed to tackle this problem. Traditional approaches include statistical methods, clustering algorithms, and rule-based systems [1] [2]. However, these methods often struggle to capture complex dependencies and subtle anomalies in large-scale network data. Recent advancements in deep learning and graph theory have led to the emergence of novel anomaly detection techniques that leverage the structural information of networks. These techniques have shown promising results in detecting anomalies in diverse domains, including cybersecurity, fraud detection, and insider threat detection [5].

B. Importance and Challenges of Network Anomaly Detection

Network anomaly detection plays a vital role in maintaining the security and integrity of network systems. Anomalies in network data can indicate malicious activities, system failures, or emerging threats. However, detecting anomalies in complex networks poses several challenges. First, networks often exhibit dynamic behavior, making it difficult to distinguish between normal fluctuations and anomalous events [6]. Second, network data is high-dimensional and heterogeneous, containing various attributes and interdependencies. Third, anomalies can manifest in different forms, such as structural changes, attribute deviations, or unusual patterns. These challenges highlight the need for advanced anomaly detection techniques that can effectively capture the complex characteristics of network data and adapt to evolving network dynamics [7] [8] [9]. Furthermore, the work of [11] proposes new approaches to address these challenges.

C. Graph-based Anomaly Detection Methods

Graph-based anomaly detection methods have gained significant attention due to their ability to model and exploit the inherent structure of network data. These methods represent network data as graphs, where nodes represent entities (e.g., users, devices) and edges capture relationships or interactions. By leveraging graph theory and network analysis techniques, graph-based anomaly detection methods can effectively capture local and global dependencies, identify abnormal patterns, and distinguish between different types of anomalies [12]. These methods often utilize graph-based features, such as node degrees, clustering coefficients, and centrality measures, to detect anomalies [5]. Furthermore, [21] introduced a novel graph-based anomaly detection algorithm that incorporates additional attributes and contextual information associated with nodes and edges.

D. Introduction to Graph Convolutional Networks (GCNs)

Graph Convolutional Networks (GCNs) have emerged as a powerful deep learning technique for graph-based anomaly detection [13]. GCNs extend convolutional neural networks (CNNs) to operate directly on graph-structured data. They leverage a localized aggregation scheme, where each node aggregates information from its neighboring nodes, capturing the graph's structural properties. By stacking multiple graph convolutional layers, GCNs can capture hierarchical representations of the network data [14]. This hierarchical representation allows GCNs to learn discriminative features and identify anomalous patterns in the network. Moreover, GCNs can handle attributed graphs, where additional features are associated with nodes or edges, enabling the integration of both structural and attribute information for more accurate anomaly detection [11] [16]. The researchers in [41] propose an enhanced version of GCNs for graph-based anomaly detection.

IV. Graph Convolutional Networks (GCNs)

A. Fundamentals of Graph Convolutional Networks (GCNs)

Graph Convolutional Networks (GCNs) have emerged as a powerful deep learning technique for analyzing graph-structured data. GCNs extend the convolutional operation from regular grids, such as images, to irregular graph structures, enabling effective representation learning and analysis of complex relational data [48]. Graph Convolutional Networks provide a powerful framework for learning representations and analyzing graph-structured data. Through graph convolutional layers, spectral-based or spatial-based approaches, and effective training techniques, GCNs enable deep learning on complex relational data, offering promising solutions in diverse application domains [47] [49].

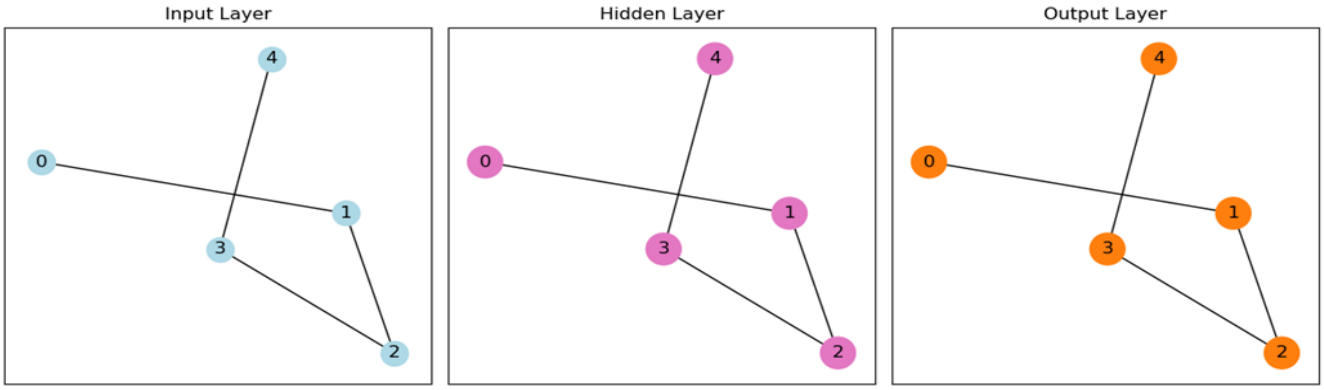


Fig. 1. Graph Convolutional Network (GCN) applied to a sample graph.

B. Graph Representation

A graph is a mathematical representation that consists of nodes and edges, where nodes represent entities or elements, and edges capture relationships or connections between nodes. Graphs are widely used to model various real-world systems, such as social networks, citation networks, and biological networks. Formally, a graph can be represented as G = (V, E), where V denotes the set of nodes and E represents the set of edges connecting pairs of nodes [47].

C. Graph Convolutional Operations

The core operation in GCNs is the graph convolution, which generalizes the convolutional operation to graph-structured data. In traditional convolutional neural networks (CNNs), convolutions are performed on regular grids using fixed-size filters. In contrast, GCNs leverage the graph structure to define a neighborhood aggregation scheme. Given a graph G = (V, E) with node features $X \in R^{N \times D}$, where N is the number of nodes and D is the feature dimension, the graph convolution operation aims to update the node representations by aggregating information from their neighboring nodes [48].

D. Graph Convolutional Layer

The graph convolutional layer is the building block of GCNs. It combines the graph convolution operation with non-linear transformations to learn expressive node representations. The output of a graph convolutional layer can be computed as $H = \sigma(AXW)$, where $H \in R^{N \times F}$ is the output matrix of node representations, A is the adjacency matrix that encodes the graph structure, X is the input node feature matrix, W is the learnable weight matrix, and σ denotes the activation function. By iteratively stacking multiple graph convolutional layers, GCNs can capture increasingly complex and abstract features [48].

E. Spectral-based and Spatial-based Approaches

GCNs can be categorized into spectral-based and spatial-based approaches based on the underlying mathematical framework. Spectral-based GCNs leverage the graph Laplacian matrix to transform the graph convolution operation into the spectral domain, where the eigenvectors of the Laplacian matrix serve as the basis for filtering the node features. Spatial-based GCNs, on the other hand, operate directly on the spatial relationships between nodes without relying on the eigenvalue decomposition. They typically employ local neighborhood aggregation schemes to capture information propagation on the graph [45] [46].

F. Training and Learning

GCNs are trained using labeled data through a supervised learning process. The training objective typically involves minimizing a loss function that measures the discrepancy between the predicted labels and the ground truth labels. To mitigate overfitting and enhance generalization, regularization techniques such as dropout and weight decay can be applied. Moreover, the backpropagation algorithm, coupled with gradient descent optimization, is employed to update the parameters of the GCN model iteratively [48].

G. GCNs in Anomaly Detection

Graph Convolutional Networks (GCNs) have gained significant attention in anomaly detection due to their ability to capture complex relational information and extract meaningful features from graph-structured data. GCNs leverage graph convolution operations to propagate information among nodes and learn node representations that encode both local and global structural characteristics of the graph. By exploiting the relational dependencies encoded in the graph, GCNs can effectively capture complex patterns and identify anomalies that would be challenging to detect using traditional methods. Several studies have demonstrated the effectiveness of GCNs in anomaly detection across various domains, such as insider threat and fraud detection [1], Ethereum blockchain network [5], and network anomaly detection [7]. Researchers have explored techniques like adaptive graph convolutional layers, local and global aggregation strategies [4], data augmentation [8], and community detection [9] to enhance the performance of GCNs in anomaly detection. Continued research aims to develop novel GCN architectures and techniques to further improve accuracy and robustness in diverse application domains.

Fig. 1 represents a Graph Convolutional Network (GCN) model applied to a sample graph. The visualization consists of three subplots: the input layer, the hidden layer, and the output layer. In the input layer, nodes are shown as circles with light blue colors representing the input features. In the hidden layer, nodes are colored based on the features extracted by the GCN model using the 'coolwarm' colormap. Finally, in the output layer, nodes are colored according to the predicted class labels using the 'Set1' colormap. This visualization helps understand how the GCN model transforms the input features, captures meaningful representations in the hidden layer, and makes predictions in the output layer, providing insights into the model's inner workings. Anomaly detection using GCNs involves computing anomaly scores, which quantify the likelihood of an anomaly, by comparing predicted representations with reconstructed representations. Training on labeled data allows GCNs to learn normal patterns and discriminate between normal instances and anomalies. The mathematical equations that underpin GCN-based anomaly detection provide a formal framework for understanding the key components of the approach.

1) Graph convolutional layer: Node Representation Update: The update rule for a single graph convolutional layer can be defined as:

$$h_v^{(l+1)} = \sigma\left(\sum_u \quad h_u^{(l)} W^{(l)}\right) \qquad (1)$$

Here,

$h_u^{(l)}$ represents the representation of node *v* at layer *l*,

$W^{(l)}$ denotes the learnable weight matrix at layer *l*,

σ is an activation function, and the sum is taken over the neighboring nodes *u* of *v*.

2) Hierarchical representation learning: Stacked Graph Convolutional Layers: Multiple layers of graph convolutions

can be stacked to capture increasingly complex patterns and higher-order relationships. The hierarchical representation learning can be expressed as,

$$h_v^{(L)} = GCN(h_v^{(L-1)}, A) \quad (2)$$

Here,

$h_v^{(L)}$ represents the final representation of node *v* after *L* layers of graph convolutions,

GCN denotes the graph convolutional operation, and

A is the adjacency matrix representing the graph structure.

3) Anomaly detection anomaly score calculation: Anomaly detection can be performed by computing anomaly scores for nodes based on their predicted representations and reconstructed representations. The anomaly score can be calculated as:

$$s_v = f(h_v^{(L)}, h_v^{(L")}) \quad (3)$$

Here,

s_v represents the anomaly score assigned to node *v*,

$h_v^{(L)}$ is the predicted representation of node *v* using the GCN,

$h_v^{(L")}$ is the reconstructed representation of node *v*, and

f is a function that measures the difference between the predicted and reconstructed representations.

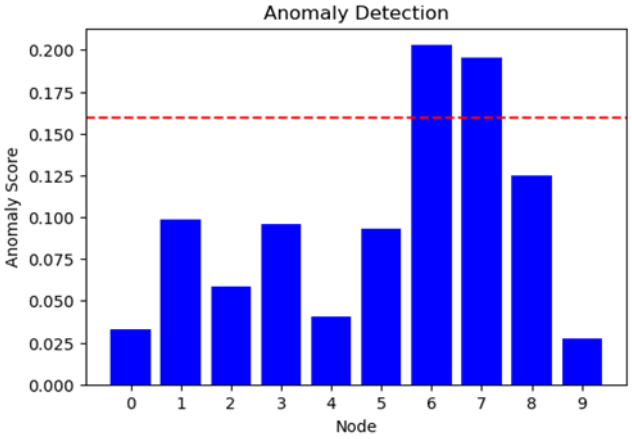


Fig. 2. Bar chart visualization of the anomaly detection results, allowing for quick identification of nodes with higher anomaly scores and potential outliers in the graph.

Fig. 2 provides a visualization of anomaly scores for each node in the graph. The x-axis represents the nodes, numbered from 0 to 9, and the y-axis represents the anomaly scores. The height of each bar corresponds to the anomaly score of the respective node.

The blue bars in the chart represent the anomaly scores of the nodes, indicating the level of deviation from the expected pattern. Higher bars indicate higher anomaly scores, suggesting nodes with more significant deviations. The red dashed line represents the anomaly threshold, separating the nodes into normal and anomalous categories. Nodes above the threshold are considered anomalous, while nodes below the threshold are considered normal.

4) Training and evaluation training on labeled data: GCNs can be trained on a labeled dataset containing normal and anomalous instances to learn to distinguish between them. The training objective can be defined as:

$$min \sum \quad Loss(y_v, GCN(h_v^{(L)}, A)) \quad (4)$$

Here, Loss is a loss function that compares the predicted labels y_v^ with the ground truth labels, and the sum is taken over all nodes in the training dataset.

V. Methodology

This systematic literature review (SLR) follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [10] guidelines to investigate the topic of "Graph Anomaly Detection with Graph Convolutional Networks." Fig. 3 shows the PRISMA flow diagram. The purpose of this SLR is to provide a comprehensive analysis and synthesis of the existing literature on the application of graph convolutional networks (GCNs) for detecting anomalies in graph-structured data.

A. Research Questions

The following research questions guide this review:

RQ1. What are the current approaches and techniques for graph anomaly detection using GCNs?

RQ2. What are the challenges and limitations of existing GCN-based graph anomaly detection methods?

RQ3. What are the emerging trends and future directions in this field?

B. Search Strategy

Our search strategy, which initially relied on automated techniques using logical operators, was followed by rigorous manual curation. We have exercised our expertise to select articles that met our stringent criteria for quality and relevance. This dual approach ensures the inclusion of the most relevant and high-quality sources. A systematic search was conducted in major academic databases, including IEEE Xplore, ACM Digital Library, Springer, Elsevier, Scopus, and Google Scholar. The search terms used included variations of "graph anomaly detection," "graph convolutional networks," "graph neural networks," and "anomaly detection in graph data." The search was limited to articles published between 2019 and 2023.

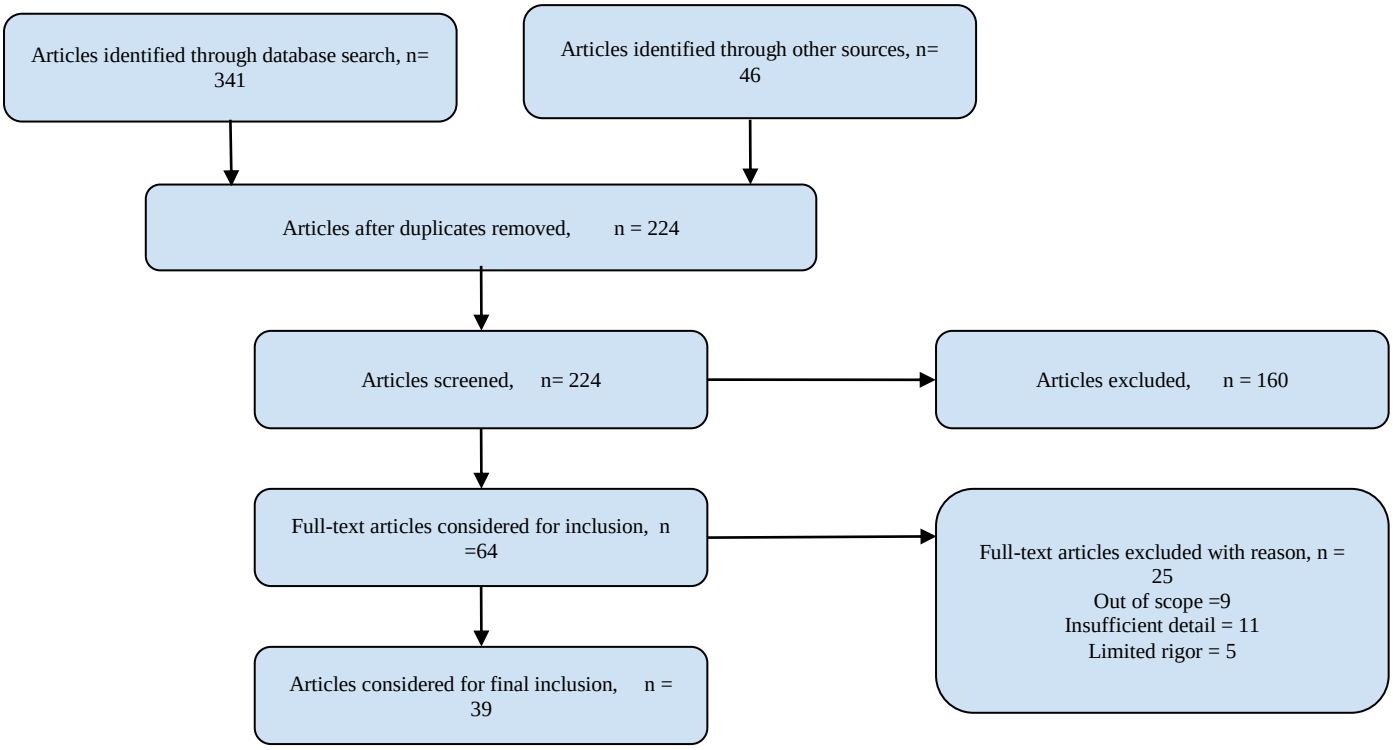


Fig. 3. PRISMA flow diagram.

C. Study Selection

The inclusion and exclusion criteria were defined to ensure the relevance and quality of the studies. Included studies were required to:

- Focus on the application of GCNs for graph anomaly detection.
- Present novel methodologies, techniques, or frameworks.
- Include evaluation metrics and datasets.
- Be published in peer-reviewed journals or conference proceedings.

Studies that did not meet the inclusion criteria, such as review articles, tutorials, or studies unrelated to graph anomaly detection, were excluded.

D. Data Extraction

Data from the selected studies were extracted using a standardized form. The extracted information included:

- Author(s) and publication details.
- Key contributions.
- Methodologies, techniques, and algorithms used.
- Datasets employed for evaluation.
- Evaluation metrics and performance results.
- Limitations and future research.

E. Data Synthesis

A qualitative synthesis was performed to analyze the findings from the selected studies. The key themes, methodologies, challenges, and trends in graph anomaly detection with GCNs were identified. The studies were analyzed to identify commonalities, differences, and gaps in the existing literature.

F. Quality Assessment

The quality and rigor of the selected studies were assessed using predefined criteria. The criteria included aspects such as research design, clarity of methodology, use of appropriate datasets, and statistical analysis.

G. Results Presentation

The findings of this SLR will be presented in a narrative format, organized thematically based on the identified research areas, methodologies, challenges, and trends. The results will be accompanied by tables, figures, and visual representations to enhance understanding and facilitate comparisons.

H. Limitations

The limitations of this SLR include potential publication bias, language limitations, and the possibility of missing relevant studies despite the comprehensive search strategy.

VI. Review of Literature

A. Overview of Selected Studies

Anomaly detection has become a critical task in various domains, such as cybersecurity, finance, healthcare, and industrial systems. Researchers have been investigating the

use of graph-based methods for anomaly detection, leveraging the power of graph neural networks (GNNs) and deep learning techniques. A comprehensive survey by [2] provides an extensive overview of graph anomaly detection with deep learning, highlighting the advancements and challenges in this field. Several studies have focused on leveraging GNNs for anomaly detection, such as the work by [1], who proposed using graph convolutional networks (GCNs) for insider threat and fraud detection. The researchers in [3] also explored graph anomaly detection with graph neural networks, discussing the current state and challenges in this area. Various approaches have been proposed to enhance the performance of graph anomaly detection. For instance, Ding and Li (2022) presented AnoGLA, an efficient scheme for improving network anomaly detection, while [5] developed a graph deep learning-based anomaly detection model specifically for Ethereum blockchain networks. The researchers in [6] introduced graph fairing convolutional networks for anomaly detection, aiming to address the fairness issue in graph-based models. The researchers in [7] proposed a rethinking of graph neural networks for anomaly detection, exploring novel architectures and techniques. The researchers in [8] developed DAGAD, a data augmentation method for graph anomaly detection, to enhance the performance of anomaly detection models. Furthermore, researchers have focused on incorporating domain-specific features and knowledge into graph anomaly detection. The researchers in [9] proposed COMGA, a community-aware attributed graph anomaly detection method that considers community structures in graphs. The researchers in [11] introduced GCCAD, a graph contrastive learning approach for anomaly detection, which leverages the contrastive learning framework. The researchers in [13] developed a high accuracy and adaptive anomaly detection model using a dual-domain graph convolutional network for insider threat detection. The researchers in [18] proposed Guard Health, a secure data management system that combines blockchain technology with graph convolutional networks for anomaly detection in smart healthcare. The literature also includes studies focusing on temporal aspects of graph anomaly detection. For example, the paper [19] addressed motif-level anomaly detection in dynamic graphs, while [21] proposed structural temporal graph neural networks for anomaly detection in dynamic graphs. The researchers in [41] introduced a multi-scale contrastive learning network with augmented view for graph anomaly detection. The researchers in [39] presented a synergistic approach that combines pattern mining and feature learning for graph anomaly detection. The researchers in [44] explored addressing heterophily in graph anomaly detection by considering the graph spectrum.

These selected studies highlight the diverse approaches and advancements in graph anomaly detection using deep learning techniques. From leveraging GNNs and GCNs to incorporating domain-specific knowledge and addressing temporal aspects, researchers are continuously striving to improve the accuracy and effectiveness of graph-based anomaly detection methods.

B. Comparative Analysis of GCN-based Approaches

A comprehensive comparative analysis of various Graph Convolutional Network GCN-based approaches for anomaly detection is presented in this section. The analysis is conducted in a tabular format, considering several key parameters to evaluate and compare the different approaches. The parameters include the reference source, graph type, method, category, objective function employed, measurement metrics used for evaluation, and the outputs of the approaches. This comparative analysis provides insights into the similarities, differences, and effectiveness of different GCN-based approaches in addressing anomaly detection tasks. By examining these parameters, we aim to identify the strengths and limitations of each approach, facilitating a better understanding of their performance and applicability in real-world scenarios.

Table I provides an analysis of various studies focused on anomaly detection using graph-based approaches. It includes information about the types of graphs used, specific methods employed, categories of anomaly detection, objective functions used measurement metrics for evaluation, and the outputs of these approaches. The table offers a comprehensive overview of different studies, highlighting their techniques, evaluation criteria, and intended applications. These studies utilize diverse types of graphs, such as attribute graphs, social graphs, blockchain transaction graphs, dynamic graphs, etc., and employ methods like Graph Convolutional Networks (GCNs), graph deep learning, community-aware attributed graph anomaly detection, contrastive learning, among others. The evaluation metrics primarily consist of precision, recall, F1-Score, AUC-ROC, and accuracy, showcasing the effectiveness of these approaches in detecting anomalies across various domains and graph types.

C. Gaps in the Existing Literature

The existing scenario of graph-based anomaly detection has advanced significantly, employing machine learning, specifically Graph Convolutional Networks (GCNs), to identify complex relationships and patterns. However, persistent constraints highlight critical gaps in existing literature. Scalability concerns [39] exist in large-scale graph processing, demanding more efficient techniques for real-time anomaly detection. The reliance on domain expertise [1] to perform feature engineering poses challenges, reducing detection accuracy. Existing approaches are mostly focused on static graphs, making it difficult to capture dynamic patterns adequately [27]. Heterogeneous graph structures [4] present modeling and analysis challenges, requiring advanced integration of various data sources. Furthermore, the interpretability [9] of graph-based models, such as GCNs, is still a challenge. These challenges and limitations will be examined in detail in the subsequent section. Addressing these limitations represents a significant gap in current understanding, necessitating the development of novel methodologies to improve scalability, reduce dependability on expertise, manage dynamic graphs effectively, accommodate heterogeneous structures, and enhance model interpretability, establishing the possibility of robust anomaly detection in real-world scenarios.

TABLE I. ANALYSIS OF INCLUDED STUDIES

Reference	Graph Type	Method	Category	Objective Function	Measurement	Outputs
Jiang et al. (2019) [1]	Attribute Graph	Graph Convolutional Networks (GCNs)	Insider Threat, Fraud	Binary Cross-Entropy Loss	Accuracy, Precision, Recall, F1-Score	Anomaly Detection
Ding & Li (2022) [4]	Social Graph	AnoGLA: Graph Link Anomaly Detection	Anomaly Detection	Link Anomaly Detection, Modularity Maximization	AUC-ROC, F1-Score	Link Anomaly Detection
Patel et al. (2020) [5]	Blockchain Transaction Graph	Graph Deep Learning	Anomaly Detection	Reconstruction Error	Precision, Recall, F1-Score	Anomaly Detection
Mesgaran & Hamza (2020) [6]	Attribute Graph	Graph Fairing Convolutional Networks (GFCNs)	Anomaly Detection	Reconstruction Error	Precision, Recall, F1-Score	Anomaly Detection
Liu et al. (2022) [8]	Attributed Graph, Temporal Graph	Data Augmentation for Graph Anomaly Detection (DAGAD)	Anomaly Detection	Reconstruction Loss, Discriminative Loss, Triplet Loss	Precision, Recall, F1-Score	Anomaly Detection
Luo et al. (2022) [9]	Attributed Graph	Community-aware attributed graph anomaly detection (COMGA)	Anomaly Detection	Reconstruction Loss, Discriminative Loss, Entropy Regularization	Precision, Recall, F1-Score	Anomaly Detection
Chen et al. (2022) [11]	Attributed Graph	Graph Contrastive Learning (GCCAD)	Anomaly Detection	Contrastive Loss, Reconstruction Loss	Precision, Recall, F1-Score	Anomaly Detection
Ding et al. (2019) [12]	Attributed Graph	Deep Anomaly Detection on Attributed Networks	Anomaly Detection	Reconstruction Loss	Precision, Recall, F1-Score	Anomaly Detection
Li et al. (2023) [13]	Dual-Domain Graph	Dual-Domain Graph Convolutional Network (DD-GCN)	Insider Threat Detection	Binary Cross-Entropy Loss	Accuracy, Precision, Recall, F1-Score	Anomaly Detection
Wu et al. (2022) [14]	Industrial IoT Graph	Graph Neural Networks (GNNs)	Anomaly Detection	Reconstruction Loss	Accuracy, Precision, Recall, F1-Score	Anomaly Detection
Cao et al. (2022) [15]	Video Graph	Adaptive Graph Convolutional Networks (AGCN)	Anomaly Detection	Reconstruction Loss	Accuracy, Precision, Recall, F1-Score	Anomaly Detection
Ma et al. (2022) [2]	Graph Level	Glocal Knowledge Distillation (GKD)	Anomaly Detection	Graph-Level Reconstruction Loss	Precision, Recall, F1-Score	Anomaly Detection
Zhang et al. (2022) [16]	Graph Level	Dual-Discriminative Graph Neural Network (D2GNN)	Anomaly Detection	Binary Cross-Entropy Loss, Margin Loss	AUC-ROC, Precision, Recall, F1-Score	Anomaly Detection
Huang et al. (2022) [17]	Financial Transaction Graph	Dgraph: Large-Scale Financial Dataset	Anomaly Detection	Reconstruction Loss, Classification Loss	AUC-ROC, Precision, Recall, F1-Score	Anomaly Detection
Wang et al. (2020) [18]	Blockchain-based Healthcare Transaction Graph	Graph Convolutional Network (GCN)	Anomaly Detection	Reconstruction Loss	Accuracy, Precision, Recall, F1-Score	Anomaly Detection
Yuan et al. (2023) [19]	Dynamic Graph	Motif-Level Anomaly Detection	Anomaly Detection	Reconstruction Loss	AUC-ROC, Precision, Recall, F1-Score	Anomaly Detection
Kisanga et al. (2023) [20]	Social Network Graph	Graph Neural Network (GNN)	Anomaly Detection	Reconstruction Loss, Classification Loss	AUC-ROC, Precision, Recall, F1-Score	Anomaly Detection
Cai et al. (2021) [21]	Dynamic Graph	Structural Temporal Graph Neural Network (STGNN)	Anomaly Detection	Reconstruction Loss, Temporal Loss	AUC-ROC, Precision, Recall, F1-Score	Anomaly Detection
Zhong et al. (2019) [22]	Graph Convolutional Networks	Graph Convolutional Label Noise Cleaner (GCLN)	Anomaly Detection	Reconstruction Loss, Classification Loss	Accuracy, Precision, Recall, F1-Score	Anomaly Detection
Zhao et al. (2022) [23]	Graph Pattern Mining	Synergistic Approach for Graph Anomaly Detection	Anomaly Detection	Pattern Mining, Feature Learning	Precision, Recall, F1-Score	Anomaly Detection
Markovitz et al. (2020) [25]	Pose Graph	Graph Embedded Pose Clustering	Anomaly Detection	Pose Clustering	Precision, Recall, F1-Score	Anomaly Detection
Lin et al. (2022) [26]	Air Quality Graph	Graph Neural Networks (GNNs)	Anomaly Detection	Reconstruction Loss	Accuracy, Precision, Recall, F1-Score	Anomaly Detection
Chen et al. (2023) [27]	Video Graph	Spatial-Temporal Graph Attention Network (ST-GAT)	Anomaly Detection	Reconstruction Loss	Accuracy, Precision, Recall, F1-Score	Anomaly Detection
Patel et al. (2022) [28]	Blockchain Transaction	Evolving Graph Deep Neural Network (EvAnGCN)	Anomaly Detection	Reconstruction Loss	Precision, Recall, F1-Score	Anomaly Detection

	Graph					
Pei et al. (2021) [29]	Attributed Graph	Attention-based Deep Residual Modeling (ResGCN)	Anomaly Detection	Reconstruction Loss	Precision, Recall, F1-Score	Anomaly Detection
You et al. (2020) [30]	Attributed Graph	Graph Attention-based Anomaly Detection (Gatae)	Anomaly Detection	Reconstruction Loss	Precision, Recall, F1-Score	Anomaly Detection
Duan et al. (2022) [41]	Graph	Multi-Scale Contrastive Learning Networks (MS-CLN)	Anomaly Detection	Contrastive Loss, Reconstruction Loss	Precision, Recall, F1-Score	Anomaly Detection
Feng et al. (2022) [42]	Graph	Full Graph Autoencoder (FGA)	Anomaly Detection	Reconstruction Loss	Precision, Recall, F1-Score	Anomaly Detection
Akoglu et al. (2015) [31]	Graph	Graph-based	Anomaly Detection	-	Survey	Anomaly detection and description
Bilgin & Yener (2006) [32]	Dynamic Graph	Network Evolution	Dynamic Network	Link Anomaly Detection, Modularity Maximization	Modularity, Clustering Coefficient, Network Evolution Measures	Models, clustering, anomaly detection
Deng & Hooi (2021) [33]	Multivariate Time-series	Graph Neural Networks	Anomaly Detection	Reconstruction Error	Precision, Recall, F1-score, AUC-ROC	Anomaly detection in multivariate time series
Fan et al. (2020) [34]	Heterogeneous Graph	Graph Neural Networks	Illicit Traded Product Detection	Reconstruction Error	Precision, Recall, F1-score, Accuracy	Identification of illicit traded products
Huang et al. (2021) [35]	Temporal Heterogeneous Graph	Information Network Embedding	Heterogeneous Networks	Reconstruction Loss, Discriminative Loss, Triplet Loss	Precision, Recall, F1-Score, Accuracy, Area Under the ROC Curve (AUC-ROC)	Temporal heterogeneous information network embedding
Wang et al. (2021) [36]	Dynamic Graph	Dynamic Hypergraph Convolution	Passenger Flow Prediction	Reconstruction Loss, Discriminative Loss, Entropy Regularization	Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE)	Metro passenger flow prediction
Wang et al. (2019) [37]	Heterogeneous Graph	Graph Attention Network	Heterogeneous Networks	Contrastive Loss, Reconstruction Loss	Precision, Recall, F1-Score, Area Under the ROC Curve (AUC-ROC)	Heterogeneous graph attention network
Zhang et al. (2019) [38]	Heterogeneous Graph	Graph Neural Network	Heterogeneous Networks	Reconstruction Loss	Precision, Recall, F1-Score, Area Under the ROC Curve (AUC-ROC)	Heterogeneous graph neural network
Zhao et al. (2021) [39]	Heterogeneous Graph	Heterogeneous Graph Structure	Graph Neural Networks	Binary Cross-Entropy Loss	Precision, Recall, F1-Score, Area Under the ROC Curve (AUC-ROC)	Heterogeneous graph structure learning
Zhu et al. (2020) [40]	Heterogeneous Mini-Graph	Neural Network	Fraud Invitation Detection	Reconstruction Loss	Precision, Recall, F1-Score, Accuracy, Area Under the ROC Curve (AUC-ROC)	Fraud invitation detection

VII. RESULTS

A. Summary of Reviewed Studied

An overview of the percentage-wise distribution of literature based on the parameters; anomaly detection, graph-based techniques, machine/deep learning, static graphs, dynamic graphs, and graph representation, is presented in the form of a pie chart (see Fig. 4). This comprehensive analysis offers valuable insights into the common practices and trends in the field.

1) Anomaly detection: Anomaly detection is a crucial aspect addressed in all the included studies, indicating its significance in identifying and flagging unusual patterns or events. This aligns with the primary objective of anomaly detection, which is to distinguish abnormal behavior from normal patterns in various domains such as cybersecurity, fraud detection, and system monitoring

2) Graph-based techniques: Graph-based techniques emerge as a prominent approach across the included studies, illustrating their effectiveness in capturing complex relationships and structures within data. These techniques leverage graph representations to model interconnected entities and interactions, enabling the detection of anomalies based on their deviations from the expected graph patterns.

3) Machine/Deep learning: Machine and deep learning methods are predominantly utilized across the literature,

showcasing their ability to handle large volumes of data and extract meaningful patterns. These techniques leverage neural networks and advanced algorithms to learn complex representations and detect anomalies based on the learned patterns.

4) Static graphs: Static graphs, which represent pre-defined structures, are widely employed in the literature to model relationships and dependencies. These static graph representations enable the analysis of anomalies by comparing observed patterns against expected graph structures.

5) Dynamic graphs: In contrast, the adoption of dynamic graphs, which capture time-dependent interactions, is relatively limited in the included literature. Most studies focus on analyzing static relationships rather than temporal variations, leaving an opportunity for future research and development in this area.

6) Graph representation: The inclusion of graph representation is prevalent across the literature, indicating its significance in organizing and structuring data for efficient anomaly detection. Graph representations facilitate the identification of abnormal patterns by capturing the relationships and dependencies among entities in a structured manner.

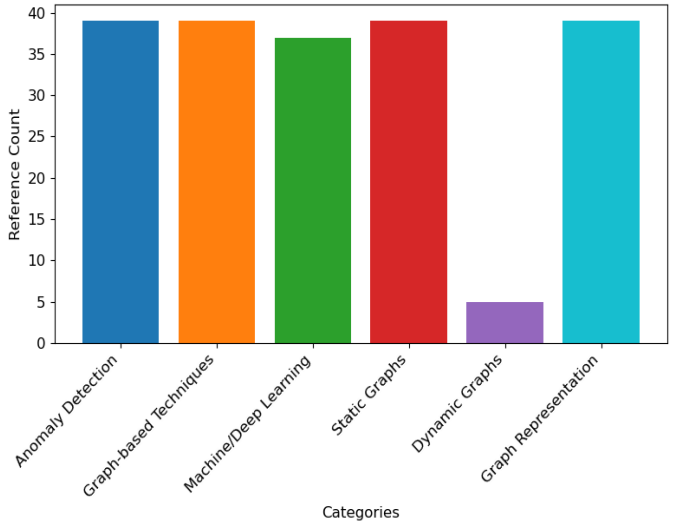


Fig. 4. Distribution of included studies based on given parameters.

B. Publication Trends in Graph-Based Anomaly Detection

The analysis of year-wise publication trends in graph-based anomaly detection, with a focus on Graph Convolutional Networks (GCN), among the included studies since 2019 reveals an interesting pattern as shown in Fig. 5. There has been a steady growth in the number of publications on this topic over the years, indicating the increasing importance and popularity of graph-based anomaly detection techniques. Specifically, since 2019, there has been a surge in research papers incorporating GCNs as a key component in detecting anomalies within graph structures. By the end of 2023, it is expected to surpass the number of publications in the year 2022. This trend suggests that researchers have recognized the power and effectiveness of GCNs in modeling complex relationships and capturing anomalous patterns within graphs. The utilization of GCNs reflects the continuous effort to leverage advanced machine/deep learning techniques to enhance anomaly detection performance in graph-based scenarios. This trend highlights the ongoing interest and active research in the field, with a focus on advancing graph-based anomaly detection methods using GCNs as a key tool.

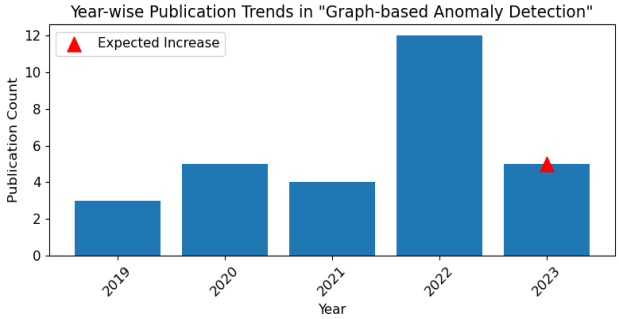


Fig. 5. Year-wise publication trends in graph-based anomaly detection.

VIII. Discussion

A. Key Insights and Observations

Graph-based anomaly detection has seen significant advancements in recent years. Machine learning, particularly deep learning, has emerged as a major approach in graph-based anomaly detection, enabling accurate detection by leveraging neural networks to capture complex relationships and patterns [1] [4]. Graph convolutional networks (GCNs) have been developed to effectively model graph structures and detect anomalies [9]. Graph representations play a crucial role in graph-based anomaly detection, and different approaches utilize various representation techniques [12]. These techniques include adjacency matrices, node features, edge features, or a combination of these [11]. By leveraging these representations, graph-based anomaly detection algorithms can effectively model normal behavior and detect deviations [27]. Graph-based anomaly detection encompasses various techniques, such as centrality-based methods, clustering-based methods, spectral methods, and local anomaly detection methods [32] [27]. Centrality-based methods identify nodes with high centrality measures as potential anomalies [31]. Clustering-based methods group nodes based on similarity and identify anomalies as nodes that do not fit into any cluster [27]. Spectral methods utilize the graph Laplacian matrix to identify anomalous patterns in the eigenvector space [32]. Local anomaly detection methods analyze local patterns and identify anomalies based on their deviation from local structures [11]. While graph-based anomaly detection has shown promise, dynamic graphs pose challenges [27]. Dynamic graphs involve evolving structures, requiring techniques that can handle the temporal dimension and capture evolving patterns [41] [43]. Developing algorithms that can effectively adapt to changing graph structures and identify anomalies in a dynamic setting remains a challenge [27].

Graph-based anomaly detection has witnessed significant advancements through machine/deep learning techniques and the utilization of various graph representations. Different techniques, such as centrality-based, clustering-based,

spectral, and local anomaly detection methods, contribute to the detection of anomalies. However, the challenges posed by dynamic graphs necessitate the development of innovative approaches to handle evolving structures and capture temporal patterns. Further research in this field will contribute to the advancement of graph-based anomaly detection techniques and their application in real-world scenarios.

B. Challenges and Limitations of Graph-Based Approaches

The challenges and limitations of graph-based approaches in anomaly detection include scalability issues, the requirement of domain expertise for manual feature engineering, handling imbalanced datasets, addressing the dynamic nature of graphs, heterogeneity and ensuring interpretability of the models. Overcoming these challenges will enhance the effectiveness and applicability of graph-based anomaly detection techniques in various real-world scenarios.

1) Scalability: One of the challenges in graph-based anomaly detection is the scalability issue when dealing with large-scale graphs [39]. As the size of the graph increases, the computational complexity of graph algorithms grows significantly, making it challenging to detect anomalies efficiently. Efficient algorithms and techniques are required to handle large-scale graphs and maintain real-time anomaly detection.

2) Domain expertise: Another challenge is the requirement of domain expertise and manual feature engineering [1]. Graph-based approaches often rely on feature extraction and selection, which demand expert knowledge and a deep understanding of the underlying graph structure. Manual feature engineering can be time-consuming and may not capture all relevant information, limiting the accuracy of anomaly detection.

3) Imbalanced datasets: Furthermore, graph-based approaches face challenges in handling imbalanced datasets [27]. Anomalies are typically rare events, resulting in imbalanced classes where the number of normal instances outweighs the number of anomalies. Imbalanced datasets can lead to biased models and reduced performance in detecting anomalies. Techniques such as data augmentation, oversampling, or adjusting the anomaly detection threshold are required to address this issue.

4) Dynamic graphs: The dynamic nature of graphs poses another significant challenge [27]. Dynamic graphs involve changing network structures over time, making it essential to develop techniques that can adapt to evolving patterns [43]. Handling temporal dependencies, capturing time-varying behaviors, and maintaining real-time detection in dynamic graphs remain active areas of research [35] [36] [37] [38] [39].

5) Heterogeneity: Heterogeneous graphs pose additional challenges in graph-based anomaly detection [4]. Heterogeneous graphs consist of multiple types of nodes and edges, representing diverse entities and relationships within a system or network. The presence of different node and edge types introduces complexity in modeling and analyzing the graph structure. Heterogeneous graphs often involve diverse data types, such as textual data, numerical attributes, or temporal information associated with different node types. The effective fusion and utilization of these heterogeneous data sources for anomaly detection require careful consideration and feature engineering techniques [37] [38] [39].

6) Interpretability: The interpretability of graph-based approaches is another limitation [9]. Deep learning models, such as graph convolutional networks (GCNs), often act as black boxes, making it challenging to understand the factors contributing to anomaly detection. Interpretability is crucial for building trust in the models and gaining insights into the detected anomalies.

C. Potential Advancements and Future Research Directions

1) Integration of deep learning: The integration of deep learning models has emerged as a promising approach in graph-based anomaly detection, enabling more effective and accurate detection of anomalies in complex graph structures. Several studies in the included literature have explored the integration of deep learning models in this context.

The researchers in [27] highlighted the effectiveness of deep learning models in graph-based anomaly detection. They emphasized that deep learning models, such as Graph Convolutional Networks (GCNs), can capture intricate relationships and patterns in graphs, leading to improved anomaly detection performance. The work carried out in [5] specifically mentioned the use of GCNs in anomaly detection. They highlighted the ability of GCNs to aggregate information from neighboring nodes, enabling the model to capture local graph structures and identify anomalous patterns. The study by [4] also discussed the integration of deep learning models in graph-based anomaly detection. They highlighted the potential of deep learning models, such as autoencoders (AEs) and recurrent neural networks (RNNs), in capturing complex patterns and anomalies in graphs. Moreover, [9] proposed a deep learning-based method for anomaly detection in graphs. They introduced a deep autoencoder model that leverages the expressive power of deep learning to learn robust representations of graph data and detect anomalies based on reconstruction errors. These studies collectively demonstrate the growing interest in leveraging deep learning models, particularly GCNs and autoencoders, for graph-based anomaly detection. The integration of deep learning models provides the capability to effectively capture and analyze complex graph structures, enhancing the detection performance and enabling the detection of subtle anomalies that may be challenging for traditional methods. By harnessing the power of deep learning, these approaches offer the potential to improve the accuracy and scalability of anomaly detection in various domains, including cybersecurity, social networks, biological networks, transportation networks, and other applications where graph-based data is prevalent.

The integration of deep learning models in graph-based anomaly detection represents an exciting research direction that holds promise for future advancements in the field.

2) Dynamic graph analysis: Let $G = (V,E,T)$ be a dynamic graph, where V represents the set of vertices, E represents the set of edges, and T represents the set of timestamps. We aim to detect anomalies in this dynamic graph.

For each timestamp $t \in T$ we denote the graph at time t as $G_t = (V_t, E_t)$, where V_t is the set of vertices at time t and E_t is the set of edges at time t. We assume that the graph evolves over time, and the vertex and edge sets can change at different timestamps. The anomaly detection problem in dynamic graphs can be formulated as finding a function $f : G \rightarrow \{0,1\}$, where $f(G) = 1$ indicates an anomaly in the graph and, $f(G) = 1$ indicates a normal graph. The temporal aspect of the network must be considered, as nodes and edges can appear, disappear, or change over time. Fig. 6 represents a dynamic network with 500 nodes, where each node appears at a different time point. The x and y coordinates of the nodes correspond to their respective node IDs. The color of each node represents its timestamp, with earlier nodes being displayed in cooler colors (e.g., yellow) and later nodes in warmer colors (e.g., red). The graph starts with the first node appearing at time point 0 and gradually increases to the last node appearing at time point 499. The edges between nodes connect consecutive nodes, forming a linear structure. The colorbar on the right side of the plot indicates the mapping between the timestamp values and the corresponding colors. The goal is to design a suitable algorithm or model that can accurately detect anomalies in the dynamic graph based on the evolving vertex [43] and edge sets at different timestamps.

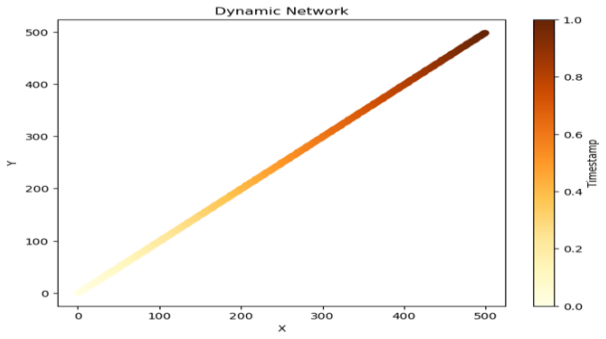


Fig. 6. Graphical representation of a dynamic network where each node appears at a different time point.

The existing research primarily focuses on the detection of anomalies in simple graphs. However, real-world networks are significantly more intricate and exhibit diverse characteristics [24]. These include heterogeneous graphs with multiple node types [38] [39], spatio-temporal graphs that evolve with time [35], and hypergraphs with non-pairwise relations [36]. Detecting and predicting anomalies in such complex graphs pose significant technical challenges [4]. For instance, the dynamic nature of nodes and links in real-world networks means that anomalous entities or relationships can sometimes exhibit normal behaviors similar to other entities in static networks. As a result, the accuracy of anomaly detection methods diminishes [4]. Consequently, key challenges persist in effectively modeling the temporal characteristics of dynamic networks and updating real-time graph embeddings. Additionally, in the context of heterogeneous graph anomaly detection, the incorporation of both attribute and structure information pertaining to various types of nodes and edges into the graph learning model represents an open research problem [11] [27] [37].

Hence, there remains ample scope for further exploration of anomaly detection and prediction on complex graphs as an important avenue for future research highlighting the importance of dynamic graph analysis in understanding and detecting anomalies in evolving systems.

IX. Summary

This systematic literature review focuses on anomaly detection in network data, with a particular emphasis on graph-based approaches, specifically Graph Convolutional Networks (GCNs). The paper discusses the fundamentals of GCNs, including graph representation, graph convolutional operations, and the structure of the graph convolutional layer. The paper also explores the use of GCNs in anomaly detection, discussing the applications of the graph convolutional layer, hierarchical representation learning, and the overall process of anomaly detection using GCNs. To address Research Question 1, a comprehensive review of the relevant literature is presented, comparing various GCN-based approaches. The findings and analysis section summarizes the reviewed studies, highlighting the significance of graph-based techniques, machine/deep learning, static graphs, dynamic graphs, and graph representation in anomaly detection. Further to address Research Question 2, the paper proceeds with a discussion on key insights, challenges, and limitations of graph-based approaches, such as scalability, domain expertise, imbalanced datasets, dynamic graphs, heterogeneity, and interpretability. Finally, to address Research Question 3, potential advancements and future research directions, including the integration of deep learning models and dynamic graph analysis, are identified.

X. Conclusion

This review has provided a comprehensive overview and analysis of anomaly detection in network data, with a focus on graph-based approaches and GCNs. The review of literature highlighted the significance of graph-based techniques, machine/deep learning, and various aspects of graph representation in anomaly detection. The findings suggest that GCNs have shown promising results in detecting anomalies and can effectively capture the complex relationships and patterns present in network data. However, several challenges and limitations, such as scalability, domain expertise, imbalanced datasets, dynamic graphs, heterogeneity, and interpretability, need to be addressed to enhance the practicality and applicability of graph-based approaches. Furthermore, the integration of deep learning models and dynamic graph analysis emerges as potential areas for future research. By leveraging the advancements in deep learning and exploring the temporal dynamics of networks, further improvements can be made in anomaly detection techniques. Overall, this paper provides valuable insights and directions for researchers and practitioners working in the field of anomaly detection, offering a foundation for future studies and advancements in this important area of research.

REFERENCES

[1] J. Jiang et al., "Anomaly detection with graph convolutional networks for insider threat and fraud detection," in MILCOM 2019 - 2019 IEEE Military Communications Conference (MILCOM), 2019.

[2] R. Ma, G. Pang, L. Chen, and A. van den Hengel, "Deep graph-level anomaly detection by glocal knowledge distillation," in Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022.

[3] H. Kim, B. S. Lee, W.-Y. Shin, and S. Lim, "Graph anomaly detection with graph neural networks: Current status and challenges," IEEE Access, vol. 10, pp. 111820–111829, 2022.

[4] Q. Ding and J. Li, "AnoGLA: An efficient scheme to improve network anomaly detection," J. Inf. Secur. Appl., vol. 66, no. 103149, p. 103149, 2022.

[5] V. Patel, L. Pan, and S. Rajasegarar, "Graph deep learning based anomaly detection in ethereum blockchain network," in Network and System Security, Cham: Springer International Publishing, 2020, pp. 132–148.

[6] M. Mesgaran and A. B. Hamza, "Graph fairing convolutional networks for anomaly detection," Pattern Recognit., vol. 145, no. 109960, p. 109960, 2024.

[7] J. Tang, J. Li, Z. Gao, and J. Li, "Rethinking graph Neural Networks for anomaly detection," in International Conference on Machine Learning, 2022.

[8] F. Liu et al., "DAGAD: Data Augmentation for Graph Anomaly Detection," in IEEE International Conference on Data Mining (ICDM), 2022.

[9] X. Luo et al., "Comga: Community-aware attributed graph anomaly detection," in Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022, pp. 657–665.

[10] M. D. J. Peters, C. M. Godfrey, H. Khalil, P. Mcinerney, D. Parker, and C. B. Soares, "'Guidance for conducting systematic scoping reviews,''," Int. J. Evidence-Based Healthcare, vol. 13, no. 3, pp. 141–146, 2015.

[11] B. Chen et al., "GCCAD: Graph contrastive learning for anomaly detection," IEEE Transactions on Knowledge and Data Engineering, 2022.

[12] K. Ding, J. Li, R. Bhanushali, and H. Liu, "Deep anomaly detection on attributed networks," in Proceedings of the 2019 SIAM International Conference on Data Mining, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2019, pp. 594–602.

[13] X. Li et al., "A high accuracy and adaptive anomaly detection model with dual-domain graph convolutional network for insider threat detection," IEEE Trans. Inf. Forensics Secur., vol. 18, pp. 1638–1652, 2023.

[14] Y. Wu, H.-N. Dai, and H. Tang, "Graph neural networks for anomaly detection in industrial internet of things," IEEE Internet of Things J., vol. 9, no. 12, pp. 9214–9231, 2022.

[15] C. Cao, X. Zhang, S. Zhang, P. Wang, and Y. Zhang, "Adaptive graph convolutional networks for weakly supervised anomaly detection in videos," IEEE Signal Process. Lett., vol. 29, pp. 2497–2501, 2022.

[16] G. Zhang et al., "Dual-discriminative graph neural network for imbalanced graph-level anomaly detection," Advances in Neural Information Processing Systems, vol. 35, pp. 24144–24157, 2022.

[17] X. Huang et al., "DGraph: A large-scale financial dataset for graph Anomaly Detection," Advances in Neural Information Processing Systems, 2022.

[18] Z. Wang, N. Luo, and P. Zhou, "GuardHealth: Blockchain empowered secure data management and Graph Convolutional Network enabled anomaly detection in smart healthcare," J. Parallel Distrib. Comput., vol. 142, pp. 1–12, 2020.

[19] Z. Yuan, M. Shao, and Q. Yan, "Motif-level Anomaly Detection in Dynamic Graphs," IEEE Transactions on Information Forensics and Security, 2023.

[20] P. Kisanga, I. Woungang, I. Traore, and G. H. S. Carvalho, "Network anomaly detection using a graph neural network," in 2023 International Conference on Computing, Networking and Communications (ICNC), 2023.

[21] L. Cai et al., "Structural temporal graph neural networks for anomaly detection in dynamic graphs," in Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021.

[22] J. X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[23] T. Zhao, T. Jiang, N. Shah, and M. Jiang, "A synergistic approach for graph anomaly detection with pattern mining and feature learning," IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 6, pp. 2393–2405, 2022.

[24] J. Ren, F. Xia, I. Lee, A. N. Hoshyar, and C. C. Aggarwal, "Graph learning for anomaly analytics: Algorithms, applications, and challenges," ACM Trans. Intell. Syst. Technol., 2022.

[25] A. Markovitz, G. Sharir, I. Friedman, L. Zelnik-Manor, and S. Avidan, "Graph embedded pose clustering for anomaly detection," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[26] X. Lin, H. Wang, J. Guo, and G. Mei, "A deep learning approach using graph neural networks for anomaly detection in air quality data considering spatiotemporal correlations," IEEE Access, vol. 10, pp. 94074–94088, 2022.

[27] H. Chen, X. Mei, Z. Ma, X. Wu, and Y. Wei, "Spatial–temporal graph attention network for video anomaly detection," Image Vis. Comput., vol. 131, no. 104629, p. 104629, 2023.

[28] V. Patel, S. Rajasegarar, L. Pan, J. Liu, and L. Zhu, "EvAnGCN: Evolving graph deep neural network based anomaly detection in blockchain," in Advanced Data Mining and Applications, Cham: Springer Nature Switzerland, 2022, pp. 444–456.

[29] Y. Pei, T. Huang, W. van Ipenburg, and M. Pechenizkiy, "ResGCN: Attention-based deep residual modeling for anomaly detection on attributed networks," in 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), 2021.

[30] Z. You, X. Gan, L. Fu, and Z. Wang, "GATAE: Graph attention-based anomaly detection on attributed networks," in 2020 IEEE/CIC International Conference on Communications in China (ICCC), 2020.

[31] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," Data Min. Knowl. Discov., vol. 29, no. 3, pp. 626–688, 2015.

[32] C. Bilgin and B. Yener, "Dynamic network evolution: Models, clustering, anomaly detection," IEEE Networks, 2006

[33] A. Deng and B. Hooi, "Graph neural network-based anomaly detection in multivariate time series," Proc. Conf. AAAI Artif. Intell., vol. 35, no. 5, pp. 4027–4035, 2021.

[34] Y. Fan et al., "Metagraph aggregated heterogeneous graph neural network for illicit traded product identification in underground market," in 2020 IEEE International Conference on Data Mining (ICDM), 2020.

[35] H. Huang, R. Shi, W. Zhou, X. Wang, H. Jin, and X. Fu, "Temporal heterogeneous information network embedding," in Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021.

[36] J. Wang, Y. Zhang, Y. Wei, Y. Hu, X. Piao, and B. Yin, "Metro passenger flow prediction via dynamic hypergraph convolution networks," IEEE Trans. Intell. Transp. Syst., vol. 22, no. 12, pp. 7891–7903, 2021.

[37] X. Wang et al., "Heterogeneous graph attention network," in The World Wide Web Conference, 2019.

[38] C. Zhang et al., "Heterogeneous graph neural network," In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 793–803, 2019

[39] J. Zhao, X. Wang, C. Shi, B. Hu, G. Song, and Y. Ye, "Heterogeneous Graph Structure Learning for Graph Neural Networks," Proc. Conf. AAAI Artif. Intell., vol. 35, no. 5, pp. 4697–4705, 2021.

[40] Y. N. Zhu et al., "Heterogeneous mini-graph neural network and its application to fraud invitation detection," in 2020 IEEE International Conference on Data Mining (ICDM), 2020.

[41] J. Duan et al., "Graph anomaly detection via multi-scale contrastive learning networks with augmented view," in Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

[42] Y. Feng, J. Chen, Z. Liu, H. Lv, and J. Wang, "Full graph autoencoder for one-class group anomaly detection of IIoT system," IEEE Internet Things J., vol. 9, no. 21, pp. 21886–21898, 2022.

[43] J. Kim, K. Kim, G. Y. Jeon, and M. M. Sohn, "Temporal Patterns Discovery of Evolving Graphs for Graph Neural Network (GNN)-based Anomaly Detection in Heterogeneous Networks," J. Internet Serv. Inf. Secur, vol. 12, no. 1, pp. 72–82, 2022.

[44] Y. Gao, X. Wang, X. He, Z. Liu, H. Feng, and Y. Zhang, "Addressing heterophily in graph anomaly detection: A perspective of graph spectrum," in Proceedings of the ACM Web Conference 2023, 2023.

[45] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in 2nd International Conference on Learning Representations (ICLR), 2013.

[46] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," Advances in neural information processing systems, 2016.

[47] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," Advances in neural information processing systems, 2017.

[48] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks.", 2016

[49] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neurasl networks," IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 1, pp. 4–24, 2020.

Ascertaining Speech Emotion using Attention-based Convolutional Neural Network Framework

Ashima Arya[1], Vaishali Arya[2], Neha Kohli[3], Namrata Sukhija[4], Ashraf Osman Ibrahim[5]*, Salil Bharany[6]*, Faisal Binzagr[7], Farkhana Binti Muchtar[8], Mohamed Mamoun[9]

Department of Computer Science and Information Technology, KIET Group of Institutions, Delhi-NCR, Ghaziabad, India[1]
Department of Computer Science and Engineering, GD Goenka University, Sohna (Gurgaon), India[2, 3]
Department of Computer Science and Engineering, SRM University, Delhi-NCR, Sonipat 131029, India[4]
Creative Advanced Machine Intelligence Research Centre, Faculty of Computing and Informatics, Universiti Malaysia Sabah[5]
Department of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India[6]
Department of Computer Science, King Abdulaziz University, P.O. Box 344, Rabigh 21911, Saudi Arabia[7]
Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia[8]
Faculty of Computer Science and Information Technology, Alzaiem Alazhari University, Khartoum North 13311, Sudan[9]

***Abstract*—Conversation among people is a profuse form of interaction that also carries emotional information. Speech input has been the subject of numerous studies over the last ten years, and it is now crucial for human-computer connection, as well as for medical care, privacy, and stimulation. This research aims to evaluate if the suggested framework can aid in speech emotion recognition (SER) activities and determine if Convolutional Neural Network (CNN) systems are efficient for SER activities using transfer learning models on spectrogram. In this investigation, the authors present a brand-new attention-based CNN framework and evaluate its efficacy against several well-known CNN architectures from earlier research. The effectiveness of the suggested system is assessed using the SAVEE dataset, an open-access resource for emotive speech, compared to famous CNN models like VGG16, InceptionV3, ResNet50, InceptionResNetV2, and Xception. The authors used stacked 10-fold cross-validation on SAVEE for all of our trials. Amongst these CNN structures, the suggested model had the greatest accuracy (87.14%), followed by VGG16 (83.19%) and InceptionResNetV2 (82.22%). Compared to contemporary techniques, the test results and evaluation show our proposed approach to have steady and impressive results.**

***Keywords*—Convolutional neural network; emotions; speech; transfer learning models; spectrogram**

I. INTRODUCTION

Automatic recognition and identification of emotions from speech signals in speech emotion recognition (SER) using machine learning is challenging [1]. SER is a quick and usual method of communication and exchanging information among humans and computers and has many real-world applications in the domain of Human-computer interaction (HCI). Feelings expressed via speech must be accurately identified and appropriately handled to provide more natural and HCI. However, building an effective SER is a complex and arduous effort due to utterance levels and abstract emotions [2]. Moreover, determining a methodologically felicitous algorithm is crucial in realizing and achieving a performance superior to the established benchmarks. The preparation and entry of sound data, obtaining features, and identifying emotions are only a few of the basic steps covered by standard SER approaches. In the quickest-growing study area currently, emotion detection in speech signals, scientists have created techniques to identify sentiments in speech signals inherently [3]. As soon as they are suggested, the concept of SER will be widely applied in the domains of schooling and medical care [4].

Although multimodal techniques can more effectively accomplish algorithmic methods for emotion identification [5], audio has been a useful medium for this job owing to the variety of data given by a person's voice [6]. Choosing a reliable approach for obtaining prominent and distinguishing characteristics from spoken words to describe the feelings of an individual speaking based on their auditory elements has become a key problem for feature mining experts. For SER, low-level handmade characteristics, including vitality, zero-crossing, length, linear classifier factor, Mel-frequency MFCC, and non-linear attributes like tiger power activator, were extensively studied during the previous ten years. Most investigations now use as a Mel-scale filtering of the financial institution voice spectrogram source characteristic when employing ML approaches for SER. CNNs often utilize spectrograms, a 2-D model representing speech sounds, to gather notable and distinct characteristics for implementation in SER [7] alongside other computational purposes [3, 8, 9]. Most two-dimensional CNNs are created specifically for optical evaluations [10, 11], and investigators have been motivated to investigate two-dimensional CNNs in the discipline of SER by their effectiveness. The CNN model may obtain excellent, important data to identify feelings in messages using spectrograms as appropriate descriptions of verbal data.

The potential of ML techniques in SER includes the systematic retrieval of emotional qualities from the unprocessed utterance and comprehending the correlations among those features. It has proven to be more effective than traditional methods. For instance, the researchers in [12] presented a combination of models constructed using long short-term memory (LSTM) and CNNs to implement temporal participation. The length of the talk hadn't been planned or

resolved, regardless of the speaker's depiction. Since feelings can fluctuate throughout a prolonged conversation, missing crucial details if spoken data has been split could influence the outcome. Leveraging continuous SER, [13] created a CNN framework with two downsampling/upsampling architectures and variable stratum compression coefficients. A system's efficiency might be impacted by fluctuating variables, leading to an overfitting issue. Consequently, the degree of complexity and feature count must be reduced to address and mitigate the issues. Recent developments in AI are also demonstrated by SER modelling with the attention system [2], transfer learning [7], as well as deep neural systems [14, 15]. They mainly referred to exceptional models for speech characteristics. The core components of SER include characteristics extraction and choosing, notwithstanding the improvement of the CNN mentioned above models. Speech sentiment may be inferred from several speech cues.

Current research develops the model using a complete methodology in light of its significance and application. As a result, it lacks a second predictor to do the categorization. Additionally, extract the emotions in communication using the attention component as a CNN layer. Additionally, because our representation uses the fourth pooling stage of the VGG-16 approach, it needs fewer components. In particular, this pooling stage collects priceless intriguing speech data, facilitating quick emotion identification. The following are the primary benefits provided by our suggested approach:

- One of the best strategies for SER that our findings suggest is a unique CNN model that combines the VGG-16 with the attention unit.
- By combining the attention and convolution modules on VGG-16, the recommended approach may identify areas of utterance that are more prone to degrade at each level.
- Since the suggested CNN approach may be taught completely, an additional filter for development and evaluation is not necessary.
- The benchmark SAVEE datasets are used to assess our model.
- The proposed technique has been assessed both qualitatively and quantitatively. The assessment's findings show that our approach beats cutting-edge techniques.

The remainder of the research paper is structured in the following fashion: Several comparable investigations are included in Section II. The suggested model is presented in Section III. The results of the study are reviewed and examined in Section IV. Section V presents our conclusion.

II. Related Work

In the modern day, the study of computational signal processing is still in its infancy. Many academics have established a variety of approaches in this field for SER during the last ten years. The two primary parts of the SER work are often separated into choosing characteristics and grouping. It is difficult to find a prejudiced choice of characteristics and grouping approach that accurately detects the speaker's feelings in this area [16]. ML algorithms are being quickly employed for SER because of the rise in information and expense processing [17, 18, 19, 20], and numerous investigators are using these techniques for reliable depiction of features in various domains [21]. Huang et al. [8] introduced a CNN-inspired strategy for SER due to its outstanding success in detecting images. In a comparable vein, [22] employed CNN to acquire excellent prejudiced characteristics based on spoken wave spectrograms and identify individuals' emotions. The Gaussian mixture method has been implemented by certain investigators [23] to determine the feelings of the speaker using reliable information.

Today, the majority of scholars derive excellent differentiation characteristics from speech recordings using two-dimensional CNNs. To discover concealed data, SER researchers are now capturing spectrograms, graphing messages concerning duration, and sending the results to CNNs [7, 24]. Additionally, researchers may use transfer learning procedures for SER by sending audio spectrograms across already trained CNN networks such as VGG [25]. To identify the feelings of the individual speaking through the SER framework, the CNNs approach can deduce excellent prejudiced characteristics from communication signals using spectrograms [26]. Likewise, to how LSTM-RNNs are continually exploited in the SER structure, hidden time-related data in messages is primarily learned using these networks [27]. ML methodologies now significantly contribute to the growth of SER curiosity.

The latest research by [28] revealed a, throughout its entirety, LSTM-DNN driven framework for SER that automatically extracts expression using unprocessed information instead of creating features manually. The combined strategy of CNN-LSTM is described in [29] for obtaining the most prominent characteristics derived from unprocessed spoken word data employing CNN and provided to the LSTM system for collecting the orderly data reminiscent of [30]. Ma et al. [31] established a framework for the model to handle varying audio lengths for SER. In this technique, CNN represented the verbal spectrogram characteristics, while RNNs processed the varying length phrases. Zhang et al. [32] introduced a method for SER using the already trained Alex-Net system for characteristics encoding and the conventional support vector machine (SVM) for feelings categorization.

To increase the identification efficiency of spoken messages, several techniques in the discipline of SER use CNN simulations with various forms of data [33]. Corresponding to this, other investigators developed an independent predictor [34] for identification. They employed the previously trained algorithm for obtaining fundamental characteristics using speech spectrograms that improve costs associated with the system data processing. In the current investigation, the authors present a brand-new attention-based CNN model that integrates the attention component alongside VGG-16 and evaluates its efficacy against a number of well-known CNN approaches from previous investigations. The following section provides a thorough description of the suggested framework with illustrations.

III. PROPOSED MODEL

The attention component and the well-known already trained CNN framework (VGG-16) are the foundation for our suggested approach. Considering two distinct explanations, the authors recommend the VGG-16 architecture. Firstly, it uses its reduced kernel shape, making it ideal for SER having fewer layers than its alternative equivalent VGG-19 approach, to gather the characteristics at the lowest level. Furthermore, it offers improved feature mining capabilities for SER identification. Among the transfer learning strategies is the tweaking strategy that authors employ. Authors employ the already trained size of ImageNet [35] in conjunction alongside the VGG-16 framework to perform the fine-tuning procedure. Since there aren't enough speeches available for learning, avoiding the overfitting issue is possible. Fig. 1 displays the suggested model's comprehensive component layout.

A. Pre-processing

An abundance of accurately tagged information is of utmost importance because SER is an identification challenge. It is crucial to select a collection of data that members of the identical group already tag. In contrast, the collection's producers, since multiple individuals, may interpret the identical words as expressing different feelings. To replicate the desired consistency, authors experimented with simulated records wherein specific performers or individuals deliver aloud a series of lines. In this study, researchers employ the SAVEE [36] dataset, which comprises precisely classified and noise-free audio specimens. There are 480 English speeches in the SAVEE database. The seven distinct emotions—neutral, angry, sad, happy, fear, disgust, and surprise—are represented by 15 phrases. Except for neutral, which contains 120 speeches, every emotion has 60 instances. Among the data collection, 60 illustrations for every feeling had been captured. Additionally, 420 speeches were recorded for the experiment's assessment. 30% of the data collection had been employed for testing, while 70% had been leveraged for training.

1) Augmentation: For each CNN framework, the dataset dimensions are a key determining element. A lack of input hampers the ability of CNN algorithms to effectively project inputs to concrete labels. A significant issue that insufficient data might cause is high variation in the assessment data forecasts. Authors employ augmentation to generate numerous training examples using the sparse audio recordings in the SAVEE dataset to get around this issue. The method implemented in this time-shifting as in Eq. (1).

$$\delta_{t_{new}} = \delta[t_{old} \pm \mu] \quad (1)$$

where, μ is the quantity of instances shifted, and δt_{old} is the audio stream. Originally recorded sound is captured at 44100 Hz in the current composition, and shifting is accomplished by s instances moving to the opposite direction.

2) Feature extraction: Selecting powerful characteristics within the voice input is crucial to achieving outstanding functionality for the proposed framework. The Mel spectrogram is a perfect feature because authors leverage CNN techniques for learning the data we provide. Fourier Transforms on recordings are used to create spectrograms, done independently to each of the smaller time portions of the audio input. Consequently, authors are presented with a frequency vs. time chart where the hue of the spectrogram represents the harmonic's intensity. Individuals, on the other hand, interpret frequencies exponentially instead of linear. This issue is resolved by the Mel magnitude, which converts a tone's perceptual pitch to its actual pitch.

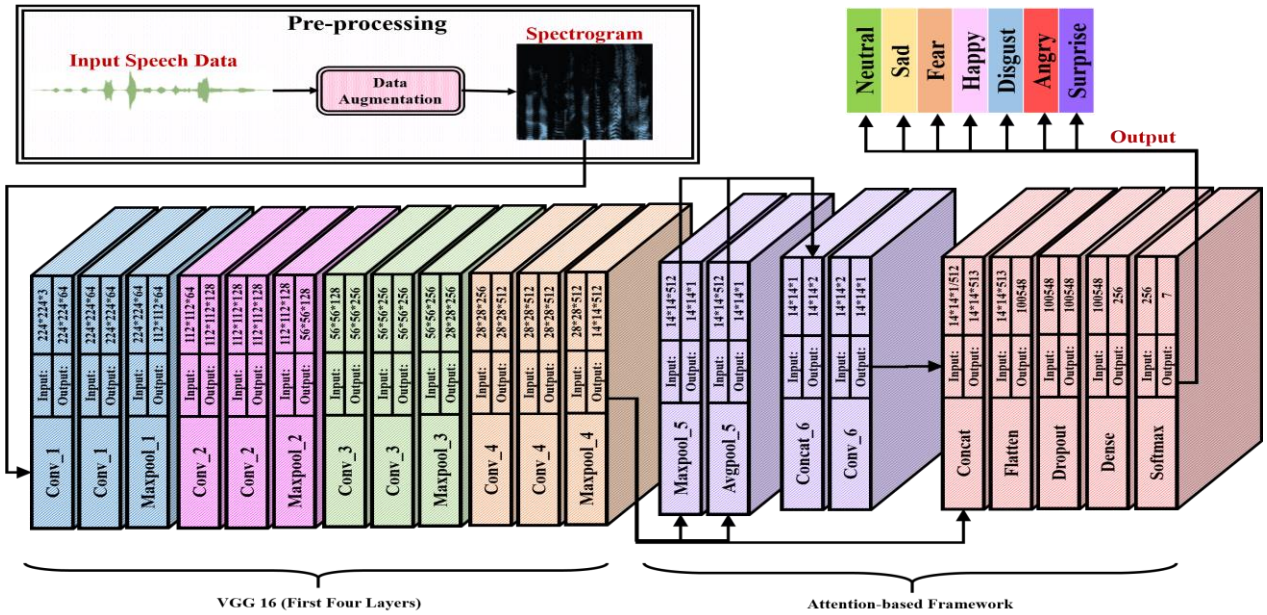


Fig. 1. The proposed architecture.

B. Attention-based Framework

Authors adopt the attention idea put forward by Woo et al. [37] for the proposed framework. The source tensor, which corresponds to the fourth pooling section of the VGG-16 framework utilized by the proposed technique, is subjected to maximum and average pooling operations. Then, employing the Sigmoid operation, these two resulting tensors are fused to conduct a convolution with a filtration value (ρ) of seven*seven. Fig. 1 displays the attention unit's layout. Eq. (2) defines the conjugated consequent tensor $(\partial_\mu(\beta))$.

$$\partial_\mu(\beta) = \theta(\rho[\beta_\mu^{avg}, \beta_\mu^{max}]) \quad (2)$$

where, the two-dimensional tensors obtained by average pooling and maximum pooling operations on the source tensor β are denoted by β_μ^{avg} and β_μ^{max}, respectively.

The fourth pooling section of the VGG-16 approach is employed in our approach. The scale-invariant component captures the intriguing hints of the picture. The midlevel area, or fourth pooling, better suitable for spectrogram, is where the intriguing hints are recovered. However, since spectrogram pictures are neither broader nor particular, the characteristics from other levels are inappropriate for spectrogram. As a result, the authors start by giving the attention component from the output of the fourth pooling layer. The production of the corresponding module is combined with the actual fourth pooling layer. Authors employ completely linked layers for expressing the concatenated characteristics obtained from the attention and convolution phase as a one-dimensional characteristic. In the proposed approach, the dense layer is set at 256, and the dropout is fixed at 0.5. The softmax layer groups the characteristics taken from the previous layers. The amount of groups in the softmax layer determines the measurement of the amount. The softmax layer produces the multinomial variation in likelihood ratings depending on the accomplished grouping. Eq. (3) defines the outcome of this distribution.

$$\varphi(x = z|y) = {e^{y_i}}/{\sum_n e^{y_n}} \quad (3)$$

where, y and z indicate the likelihoods that the softmax layer has been calculated and, correspondingly, one of the data categories employed by the suggested technique.

IV. Experiment and Results

A. State-of-the-art CNN Models

Five distinct CNN frameworks, including VGG16, InceptionV3, ResNet50, InceptionResNetV2, and Xception, have been employed in the present investigation. Among those most renowned and well-liked CNN designs is VGGnet. The distinctive VGGnet architecture consists of 138–144 million variables, approximately nineteen convolutional layers, Three*three convolutional filtering, five max–pooling stages, three completely linked layers, and a classification level as the final level [38]. By increasing its depth and breadth, Inceptionv3 is used to improve the processing capacity [39]. There are forty-eight layers in the design. The recommended framework is iterated with max-pooling to decrease the number of variables. ResNet is a standard feed-forward system with a residual link, in addition. The $(\gamma - 1)$ th results of the preceding level, also known as $(\gamma_t - 1)$, would be used to generate the residual layer outcome. The result of various procedures, including the convolution with various filtering widths and batch normalization accompanied by an activation operation on $(\gamma_t - 1)$, is referenced as $\sigma(\gamma_t - 1)$. Eq. (4) [38] can be employed to determine the residual section's ultimate result, γ_t.

$$\gamma_t = \sigma(\gamma_t - 1) + \gamma_t - 1$$

Each of the various fundamental residual blocks makes up the residual system. However, the tasks performed in the residual block fluctuate due to the various topologies of residual systems. A residual system with fifty levels is referred to as ResNet50. The InceptionResNetV2 is an amalgamation of 164-layer inception architectures featuring a residual link. To avoid any associated deterioration issues, the system uses multiple-sized CNNs that undergo training on various pictures [38]. "Extreme inception" is the abbreviation for the CNN structure known as Xception. The Xception design is a linear pile of residually connected separable by depth levels. Its 36 convolutional levels serve as the system's extraction of characteristics foundation [40].

B. K-Fold Cross-Validation

A popular method for determining the genuine forecasting errors of networks and fine-tuning the system's characteristics [41] to avoid generalization mistakes is cross-validation. Given the inconsistency in data collection, numerous models commonly encounter the overfitting problem. This outstanding approach is widely employed to address this issue [42]. The training information needs to be divided into K sections, each including an n/k specimen, wherein n is the initial sample amount, to begin the K-fold cross-validation operation. As a result, k-1 portions are employed in learning, whereas the residual portions are exploited in validation [43, 44]. The grid-search technique in our suggested method incorporates this significant strategy. The present research additionally employed the holdout approach, particularly a 3-split holdout, which is a way of partitioning the samples into several parts and engaging one part to learn the mathematical framework alongside additional parts for validating and testing the predictions. Additionally, the 10-fold cross-validation is used for several CNN networks; the study will go into more depth about the outcomes later [45-48].

C. Results

The performance of the proposed model and various CNN models is presented in Table I.

TABLE I. Performance of Various CNN Models

	Accuracy	***Precision***	***Recall***	***F1-score***
VGG16	83.19	90.68	90.97	90.82
InceptionV3	81.34	90.63	88.8	89.71
ResNet50	80.74	88.58	90.13	89.34
InceptionResNetV2	82.22	90.31	90.17	90.24
Xception	80.99	89.22	89.78	89.5
Proposed Model	87.14	92.85	93.42	93.13

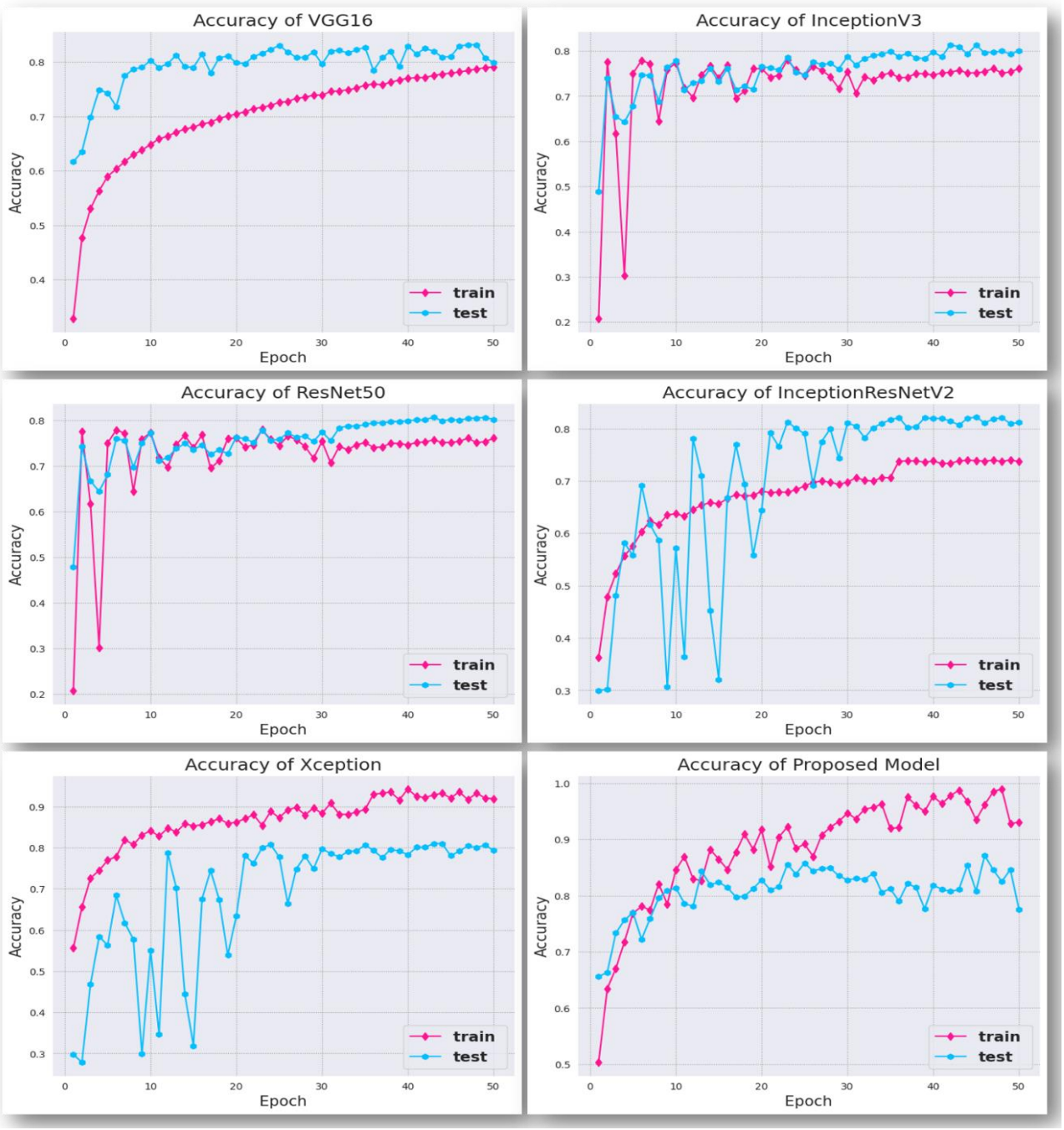


Fig. 2. Accuracy graph of various CNN models.

The proposed model achieved the highest performance with 87.14% accuracy, 92.85% precision, 93.42% recall and 93.13% f1-score. The second-best performance is demonstrated by VGG16 (83.19% accuracy, 90.68% precision, 90.97% recall and 90.24% f1-score) followed by InceptionResNetV2 (82.22% accuracy, 90.31% precision, 90.17% recall and 93.13% f1-score) and then InceptionV3 (81.34% accuracy, 90.63% precision, 88.8% recall and 89.71% f1-score).

ResNet50 observes the last performance among the underlying models with 80.74% accuracy, 88.58% precision, 90.13% recall and 89.34% f1-score. The accuracy graph, loss graph and confusion matrix for the various models are depicted in Fig. 2, Fig. 3 and Fig. 4.

The accuracy grows significantly in the initial few epochs, as seen in the Fig. 2, showing that the system is acquiring knowledge quickly. Following that, the trajectory becomes flatter, implying that there aren't sufficient epochs necessary for refining the simulation anymore. Overfitting occurs when the initial data precision improves, but the test accuracy deteriorates. It means that the framework has begun to remember the data. Inception V3 and ResNet 50 have been observed to have the minimum overfitting issue [49, 50].

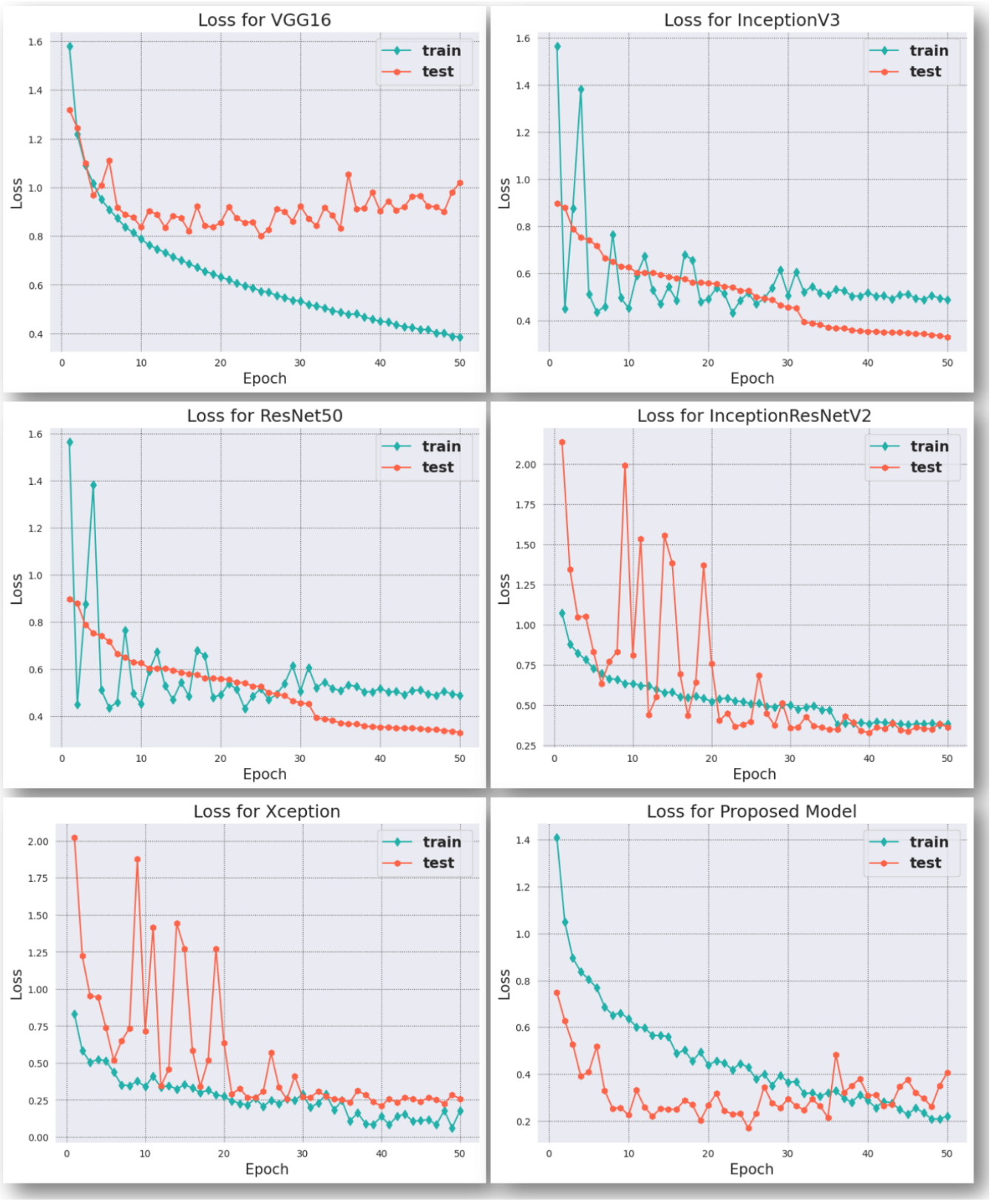


Fig. 3. Loss graph of various CNN models.

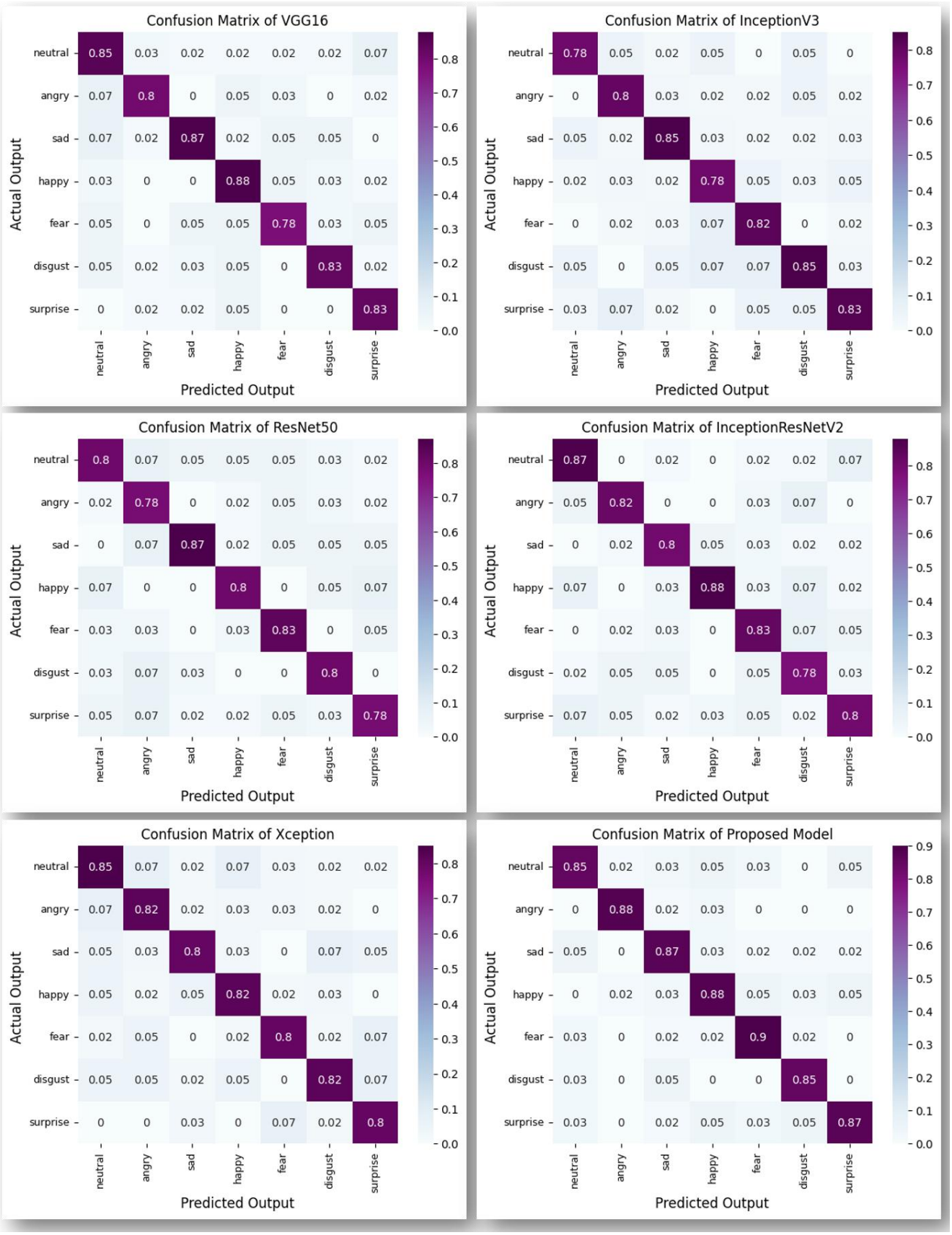


Fig. 4. Confusion matrix of different CNN models.

The loss on the learning set falls significantly during the initial few epochs, as seen in the Fig. 3. The loss in the test dataset is not declining at a comparable pace as in the preliminary set, but maintains nearly constant throughout numerous epochs. This indicates that the proposed framework generalizes effectively to new inputs.

The confusion matrix shows if the framework is "confused" in distinguishing among the various classes. It resembles a two-dimensional matrix, illustrated in the Fig. 4. The proposed model is observed to provide best results in comparison to the other prevailing models in the existing literature.

In future, authors aim to extend this research to incorporate the ensemble of various models to be fed to the attention module and compare it with existing research.

V. CONCLUSION

A key challenge to improving the model's accuracy compared to industry standards is creating a rigorous approach to gather relevant speech characteristics. Acquiring and evaluating voice elements for emotion recognition in spoken language may be challenging. The key challenges in designing a SER framework are extracting valuable characteristics and properly classifying those traits. However, developing contemporary ML techniques reduces the difficulty associated with these complicated activities. As a result, authors developed a novel SER model using attention-based CNN units, which acquire significant characteristics simultaneously with those required to group feelings. The SAVEE dataset, an open-access resource for emotional speech, is used to compare the performance of the proposed system against well-known CNN models such as VGG16, InceptionV3, ResNet50, InceptionResNetV2, and Xception. For all our experiments, the authors employed stacked 10-fold cross-validation on SAVEE. The recommended model exhibited the highest accuracy (87.14%) among these CNN architectures, followed by VGG16 (83.19%) and InceptionResNetV2 (82.22%). The test findings and assessment reveal that our suggested strategy has consistent and excellent outcomes compared to modern methods.

REFERENCES

[1] B. Liu, H. Qin, Y. Gong, W. Ge, M. Xia and L. Shi, "EERA-ASR: An energy-efficient reconfigurable architecture for automatic speech recognition with hybrid DNN and approximate computing," *IEEE Access*, vol. 6, pp. 52227-52237, 2018.

[2] Y. Zhang, J. Du, Z. Wang, J. Zhang and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Honolulu, HI, USA, 2018.

[3] Mustaqeem and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, 2019.

[4] M. Swain, A. Routray and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, pp. 93-120, 2018.

[5] J. Han, Z. Zhang, Z. Ren and B. Schuller, "EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 553-564, 2019.

[6] B. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90-99, 2018.

[7] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl and B. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proceedings of the 25th ACM international conference on Multimedia*, California, USA , 2017.

[8] J. Huang, B. Chen, B. Yao and W. He, "ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network," *IEEE access*, vol. 7, pp. 92871-92880, 2019.

[9] S. Kumar, P. Goswami and S. Batra, "Fuzzy Rank-Based Ensemble Model for Accurate Diagnosis of Osteoporosis in Knee Radiographs," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 4, 2023.

[10] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. Baik and V. de Albuquerque, "Cloud-assisted multiview video summarization using CNN and bidirectional LSTM," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 77-86, 2019.

[11] S. Khan, I. Haq, S. Rho, S. Baik and M. Lee, "Cover the violence: A novel Deep-Learning-Based approach towards violence-detection in movies," *Applied Sciences*, vol. 9, no. 22, p. 4963, 2019.

[12] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Piscataway, NJ, USA, 2016.

[13] S. Khorram, Z. Aldeneh, D. Dimitriadis, M. McInnis and E. Provost, "Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition," *arXiv preprint arXiv:1708.07050*, 2017.

[14] M. Lech, M. Stolar, C. Best and R. Bolia, "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding," *Frontiers in Computer Science*, vol. 2, p. 14, 2020.

[15] J. Zhao, X. Mao and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical signal processing and control*, vol. 47, pp. 312-323, 2019.

[16] S. Jiang, Z. Li, P. Zhou and M. Li, "Memento: An emotion-driven lifelogging system with wearables," *ACM Transactions on Sensor Networks (TOSN)*, vol. 15, no. 1, pp. 1-23, 2019.

[17] H. Wang, Q. Zhang, J. Wu, S. Pan and Y. Chen, "Time series feature learning with labeled and unlabeled data," *Pattern Recognition*, vol. 89, pp. 55-66, 2019.

[18] S. Batra, R. Khurana, M. Z. Khan, W. Boulila, A. Koubaa and P. Srivastava, "A Pragmatic Ensemble Strategy for Missing Values Imputation in Health Records," *Entropy*, vol. 24, no. 4, p. 533, 2022.

[19] S. Batra and S. Sachdeva, "Organizing standardized electronic healthcare records data for mining," *Health Policy and Technology*, vol. 5, no. 3, pp. 226-242, 2016.

[20] A. Pathak, S. Batra and V. Sharma, "An Assessment of the Missing Data Imputation Techniques for COVID-19 Data," in *Proceedings of 3rd International Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication: MARC 2021*, Singapore, 2022.

[21] R. Khalil, E. Jones, M. Babar, T. Jan, M. Zafar and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327-117345, 2019.

[22] A. Khamparia, D. Gupta, N. Nguyen, A. Khanna, B. Pandey and P. Tiwari, "Sound classification using convolutional neural network and tensor deep stacking network," *IEEE Access*, vol. 7, pp. 7717-7727, 2019.

[23] M. Navyasri, R. RajeswarRao, A. DaveeduRaju and M. Ramakrishnamurthy, "Robust features for emotion recognition from speech by using Gaussian mixture model classification," in *Information and Communication Technology for Intelligent Systems (ICTIS 2017)-Volume 2 2*, Cham, Switzerland, 2018.

[24] S. Batra, H. Sharma, W. Boulila, V. Arya, P. Srivastava, M. Z. Khan and M. Krichen, "An Intelligent Sensor Based Decision Support System for Diagnosing Pulmonary Ailment through Standardized Chest X-ray Scans," *Sensors*, vol. 22, no. 19, p. Sensors, 2022.

[25] A. Krizhevsky, I. Sutskever and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, p. 1097–1105, 2012.

[26] E. Ocquaye, Q. Mao, H. Song, G. Xu and Y. Xue, "Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition," *IEEE Access*, vol. 7, pp. 93847-93857, 2019.

[27] M. Zeng and N. Xiao, "Effective combination of DenseNet and BiLSTM for keyword spotting," *IEEE Access*, vol. 7, pp. 10767-10775, 2019.

[28] T. Sainath, O. Vinyals, A. Senior and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, South Brisbane, QLD, Australia, 2015.

[29] P. Tzirakis, G. Trigeorgis, M. Nicolaou, B. Schuller and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 8, pp. 1301-1309, 2017.

[30] X. Ma, H. Yang, Q. Chen, D. Huang and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, Amsterdam, The Netherlands, 2016.

[31] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," *Interspeech*, pp. 3683-3687, 2018.

[32] S. Zhang, S. Zhang, T. Huang and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576-1590, 2017.

[33] Q. Mao, M. Dong, Z. Huang and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE transactions on multimedia*, vol. 16, no. 8, pp. 2203-2213, 2014.

[34] P. Liu, K. Choo, L. Wang and F. Huang, "SVM or deep learning? A comparative study on remote sensing image classification," *Soft Computing*, vol. 21, pp. 7053-7065, 2017.

[35] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009.

[36] "Surrey Audio-Visual Expressed Emotion (SAVEE) Database," [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/. [Accessed 19 June 2023].

[37] S. Woo, J. Park, J. Lee and I. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, Munich, Germany, 2018.

[38] D. Theckedath and R. Sedamkar, "Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks," *SN Computer Science*, vol. 1, pp. 1-7, 2020.

[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[40] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[41] L. Yates, Z. Aandahl, S. Richards and B. Brook, "Cross validation for model selection: a review with examples from ecology," *Ecological Monographs*, vol. 93, no. 1, p. e1557, 2023.

[42] M. Kaariainen, "Semi-supervised model selection based on cross-validation," in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, Montreal, QC, Canada, 2006.

[43] D. Anguita, A. Ghio, S. Ridella and D. Sterpi, "K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines," in *DMIN*, Shenzhen, China, 2009.

[44] Ibrahim, A.O., Shamsuddin, S.M., Abraham, A. and Qasem, S.N., 2019. Adaptive memetic method of multi-objective genetic evolutionary algorithm for backpropagation neural network. Neural Computing and Applications, 31, pp.4945-4962.

[45] Bharany, S.; Sharma, S. "Intelligent Green Internet of Things: An Investigation." In Machine Learning, Blockchain, and Cyber Security in Smart Environments, Chapman and Hall: London, UK; CRC: Boca Raton, FL, USA, 2022; pp. 1–15.

[46] S. Alam et al., "Blockchain-Based Solutions Supporting Reliable Healthcare for Fog Computing and Internet of Medical Things (IoMT) Integration," Sustainability, vol. 14, no. 22. MDPI AG, p. 15312, Nov. 18, 2022. doi: 10.3390/su142215312.

[47] S. Bharany, S. Sharma, N. Alsharabi, E. Tag Eldin, and N. A. Ghamry, "Energy-efficient clustering protocol for underwater wireless sensor networks using optimized glowworm swarm optimization," Frontiers in Marine Science, vol. 10. Frontiers Media SA, Feb. 02, 2023. doi: 10.3389/fmars.2023.1117787.

[48] E. M. Onyema et al., "A Security Policy Protocol for Detection and Prevention of Internet Control Message Protocol Attacks in Software Defined Networks," Sustainability, vol. 14, no. 19. MDPI AG, p. 11950, Sep. 22, 2022. doi: 10.3390/su141911950.

[49] E. A. Adeniyi, P. B. Falola, M. S. Maashi, M. Aljebreen, and S. Bharany, "Secure Sensitive Data Sharing Using RSA and ElGamal Cryptographic Algorithms with Hash Functions," Information, vol. 13, no. 10. MDPI AG, p. 442, Sep. 20, 2022. doi: 10.3390/info13100442.

[50] A. Sundas, S. Badotra, S. Bharany, A. Almogren, E. M. Tag-ElDin, and A. U. Rehman, "HealthGuard: An Intelligent Healthcare System Security Framework Based on Machine Learning," Sustainability, vol. 14, no. 19. MDPI AG, p. 11934, Sep. 22, 2022. doi: 10.3390/su141911934.

Convolutional LSTM Network for Real-Time Impulsive Sound Detection and Classification in Urban Environments

Aigerim Altayeva[1], Nurzhan Omarov[2], Sarsenkul Tileubay[3],
Almash Zhaksylyk[4], Koptleu Bazhikov[5], Dastan Kambarov[6]
Al-Farabi Kazakh National University, Almaty, Kazakhstan[1, 2]
Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan[3]
Satbayev University, Almaty, Kazakhstan[4]
Yessenov University, Aktau, Kazakhstan[5]
Almaty University of Power Engineering and Telecommunications, Almaty, Kazakhstan[5]

***Abstract*—In recent years, the escalating challenges of noise pollution in urban environments have necessitated the development of more sophisticated sound detection and classification systems. This research introduces a novel approach employing a Convolutional Long Short-Term Memory (ConvLSTM) network tailored for real-time impulsive sound detection in metropolitan landscapes. Impulsive sounds, characterized by sudden onsets and short durations—such as honking, abrupt shouts, or breaking glass—are inherently sporadic but can significantly impact urban soundscapes and the well-being of city dwellers. Traditional sound detection mechanisms often falter in identifying these ephemeral noises amidst the cacophony of urban life. The ConvLSTM network proposed in this study amalgamates the spatial feature learning capabilities of Convolutional Neural Networks (CNN) with the temporal sequence retention attributes of LSTM, culminating in an architecture that excels in both sound detection and classification tasks. The model was trained and evaluated on a comprehensive dataset sourced from various urban settings and demonstrated commendable proficiency in discerning impulsive sounds with minimal false positives. Furthermore, the system's real-time processing capabilities ensure timely interventions, paving the way for smarter noise management in cities. This research not only propels the frontier of impulsive sound detection but also underscores the potential of ConvLSTM in addressing multifaceted urban challenges.**

Keywords*—Deep learning; CNN; LSTM; hybrid model; ANN; impulsive sound

I. Introduction

The burgeoning growth and urbanization of cities around the globe has, in recent decades, ushered in a plethora of environmental and societal challenges [1]. Among these challenges, noise pollution stands out as a particularly pervasive issue, affecting both the physical [2] and psychological health [3] of urban residents. While the continuous hum of traffic or the distant murmur of a crowded plaza might be classified as the usual sounds of urban life [4], impulsive sounds—those characterized by sudden onsets and fleeting durations—present a unique set of challenges. Whether it's the abrupt honk of a car, a sudden shout, or the shatter of glass, these noises can be more than just momentary disturbances; they can disrupt sleep, exacerbate stress, and even influence long-term health outcomes.

Historically, noise monitoring in urban spaces has relied predominantly on traditional sound detection methodologies [5]. However, these conventional systems often lack the precision and agility needed to discern between the myriad of auditory signals that coexist in a bustling urban environment [6]. Specifically, the ability to distinguish impulsive sounds from the ambient noise milieu and subsequently classify them in real-time has remained a significant gap in urban noise management systems [7]. This lacuna is further widened when considering the increasing heterogeneity of urban sounds as cities continue to evolve and densify.

Enter the era of deep learning and its transformative impact across various domains. Recent advancements in neural networks, especially Convolutional Neural Networks (CNNs) [8], have demonstrated remarkable success in image and sound processing tasks. CNNs, designed to automatically and adaptively learn spatial hierarchies from data, have fundamentally altered the landscape of sound analysis in controlled environments. However, the temporally fleeting nature of impulsive sounds in dynamic urban environments presents challenges that go beyond the scope of traditional CNNs. It necessitates the incorporation of temporal sequence learning, an attribute inherent to Long Short-Term Memory (LSTM) networks.

LSTM networks, a subtype of recurrent neural networks, excel at tasks that require the understanding of long-term dependencies, making them particularly suited for sequence prediction problems, like those seen in speech and time-series data [9]. The integration of CNN's spatial feature learning with LSTM's prowess in temporal sequence retention could potentially hold the key to a robust solution for impulsive sound detection and classification in urban locales. This potential amalgamation gave birth to the Convolutional LSTM (ConvLSTM) network—a hybrid model aiming to harness the strengths of both parent architectures.

This research pivots around the design, implementation, and evaluation of a ConvLSTM network tailored explicitly for

the detection and classification of impulsive sounds in real-time urban settings. Recognizing the profound implications of efficient noise management—ranging from urban planning and policy-making to the well-being and satisfaction of city residents—this study endeavors to bridge the existing technological gap. Through the marriage of convolutional and recurrent mechanisms, we embark on an exploration into a new frontier of urban sound management, positing a solution that promises both accuracy and timeliness in addressing the cacophony of modern urban life.

II. Related Works

The quest to discern and classify impulsive sounds within urban environments is embedded in a rich tapestry of research efforts, encompassing fields from acoustic engineering to artificial intelligence. This section delves into the pertinent literature, highlighting seminal works, and tracing the evolution of methodologies applied to this challenge.

A. Urban Sound Detection and Classification

One of the earliest works in urban sound classification was presented by [10], who utilized basic spectral features coupled with Support Vector Machines (SVM) to classify a limited set of urban sounds. Their model, though pioneering, had a limited scope in differentiating between closely related sounds. A more comprehensive approach was introduced by [11], which focused on extracting Mel-Frequency Cepstral Coefficients (MFCC) from urban soundscapes. Their work laid the foundation for many subsequent endeavors by demonstrating the potential of MFCC in capturing the nuances of urban noises.

B. Convolutional Neural Networks in Sound Analysis

The revolution brought about by deep learning in image processing soon trickled into the realm of acoustic analysis. Authors in the study [12] were among the first to employ CNNs for environmental sound classification. His model, though primarily geared towards stationary sounds, showcased the profound potential of CNNs in capturing intricate sound patterns. Further advancements by [13] extended the use of CNNs, leveraging transfer learning from pre-trained image-based networks to sound data, highlighting the shared hierarchical structures between the two domains.

C. LSTM and Sequence Modeling in Acoustics

Long Short-Term Memory (LSTM) networks, a subtype of recurrent neural networks (RNNs), have emerged as particularly influential in the realm of acoustic analysis. Their inherent capability to capture and model long-term dependencies within sequences makes them exceptionally suited for time-based sound data. Researchers in [14] effectively harnessed LSTMs for voice activity detection, shedding light on their potential in discerning complex temporal patterns. Furthermore, [15] built upon this by integrating LSTMs with attention mechanisms, aiming to identify anomalous sounds in industrial settings. These studies collectively underscore the pivotal role of LSTMs in advancing the frontier of sequence modeling within the acoustics domain.

D. ConvLSTM in Image and Video Processing

Before its foray into acoustic analysis, ConvLSTM made waves in the domain of video processing. Researchers in [16] introduced ConvLSTM as an extension to the traditional LSTM, integrating convolution operations into the recurrent updates. Their groundbreaking work in precipitation forecasting exhibited ConvLSTM's potential in spatiotemporal sequence forecasting. This novel architecture caught the attention of many, with [17] later applying it to video classification tasks, proving its versatility across multiple temporal data types.

E. Hybrid Models in Sound Detection

The intersection of diverse neural network architectures has given rise to hybrid models, which aim to leverage the unique strengths of each constituent network for enhanced performance in sound detection tasks. Recognizing the potential of such amalgamations, [18] introduced a Convolutional Recurrent Neural Network (CRNN) specifically tailored for detecting anomalous sounds within varied environments. By seamlessly integrating the spatial feature extraction capabilities of Convolutional Neural Networks (CNNs) [19] with the temporal sequence modeling prowess of Recurrent Neural Networks (RNNs) [20], their research set a new benchmark in the field. This innovative approach highlights the intrinsic benefits of harnessing both spatial and temporal dimensions in sound analysis. Hybrid models, as delineated by such pioneering works, elucidate the way forward, promising enhanced accuracy and adaptability in the complex domain of sound detection.

F. Real-Time Sound Processing

The necessity for real-time sound processing, particularly in urban contexts, is paramount. Early systems, as chronicled by [21], relied heavily on handcrafted features and simplistic classifiers. However, with the proliferation of deep learning, architectures evolved to address real-time demands. Authors in [22] presented a sound event detection system employing a stacked CNN-LSTM architecture capable of real-time sound localization and classification, illuminating the pathway for subsequent real-time models.

G. Challenges in Sound Detection Models

Navigating the complex domain of hybrid sound detection models introduces a myriad of challenges. Firstly, the integration of diverse architectures, such as CNNs and RNNs [23], presents computational burdens. The enhanced model complexity often results in increased training times, demanding more computational resources and potentially hindering real-time deployment. Additionally, the fusion of spatial and temporal data streams can lead to overfitting [24], especially when training on limited datasets [25], necessitating rigorous regularization techniques and data augmentation [26].

Another pertinent challenge is the adaptation to diverse acoustic environments [27]. Hybrid models, though versatile, may struggle with high variability in soundscapes, from fluctuating noise levels to the unique acoustic signatures of different urban settings [28]. Lastly, the interpretability of these hybrid models remains elusive. As the models grow in complexity, understanding their decision-making processes

becomes intricate, posing challenges for validation and further refinement. Addressing these challenges is crucial for the successful adoption and efficacy of hybrid models in real-world sound detection applications.

In summation, while the corpus of work surrounding urban sound detection is extensive, the specific challenge of real-time impulsive sound detection and classification in urban settings remained a largely unexplored niche. The incorporation of ConvLSTM in this context, as embarked upon in this research, represents a synthesis of past insights and current innovations, promising to further the field's understanding and capabilities.

III. Materials and Methods

This segment delineates the methodologies and resources employed throughout this investigation. It encompasses the dataset curated for discerning perilous urban acoustics, coupled with detailed insights into data assimilation, preparatory phases, and the construction of an intricate neural network model targeting the identification of hazardous urban noises. Fig. 1 provides a schematic representation of the devised framework, emphasizing real-time perilous urban acoustic detection. Subsequent subsections offer a comprehensive breakdown of the utilized resources and techniques, encapsulating the datasets engaged, the data assimilation paradigm, model conceptualization, challenges associated with impulsive acoustic detection, and a detailed exposition of the integrated CNN-LSTM architecture.

A. Data

At the study's inception, data acquisition was prioritized, recognizing the necessity of comprehensive information for robust research outcomes. An assortment of expansive datasets was engaged to scrutinize the sounds termed as "hazardous." The Environmental Sound Classification (ESC-50) dataset [29] emerged as the preferred choice for program evaluation. From its vast repertoire of 2,000 auditory samples, a curated subset of approximately 300 sounds was employed. The ESC-50's categorization encompassed:

- Faunal Acoustics (e.g., canines, felines, bovines, swine).
- Natural Phenomena (e.g., precipitation, oceanic waves, avian calls, electrical discharges).
- Human-Origin Sounds (e.g., neonatal distress, ambulatory noises, respiratory patterns).
- Domestic and Mundane Noises (e.g., door interactions, digital keystrokes, time alerts, vitreous fractures).
- Hazardous Acoustics (e.g., emergency vehicular alerts, rail transit, motor operations, timber-cutting equipment, aerial transport, pyrotechnics, detonations, canine alerts, ballistic discharges, and other precarious impulsive acoustics).

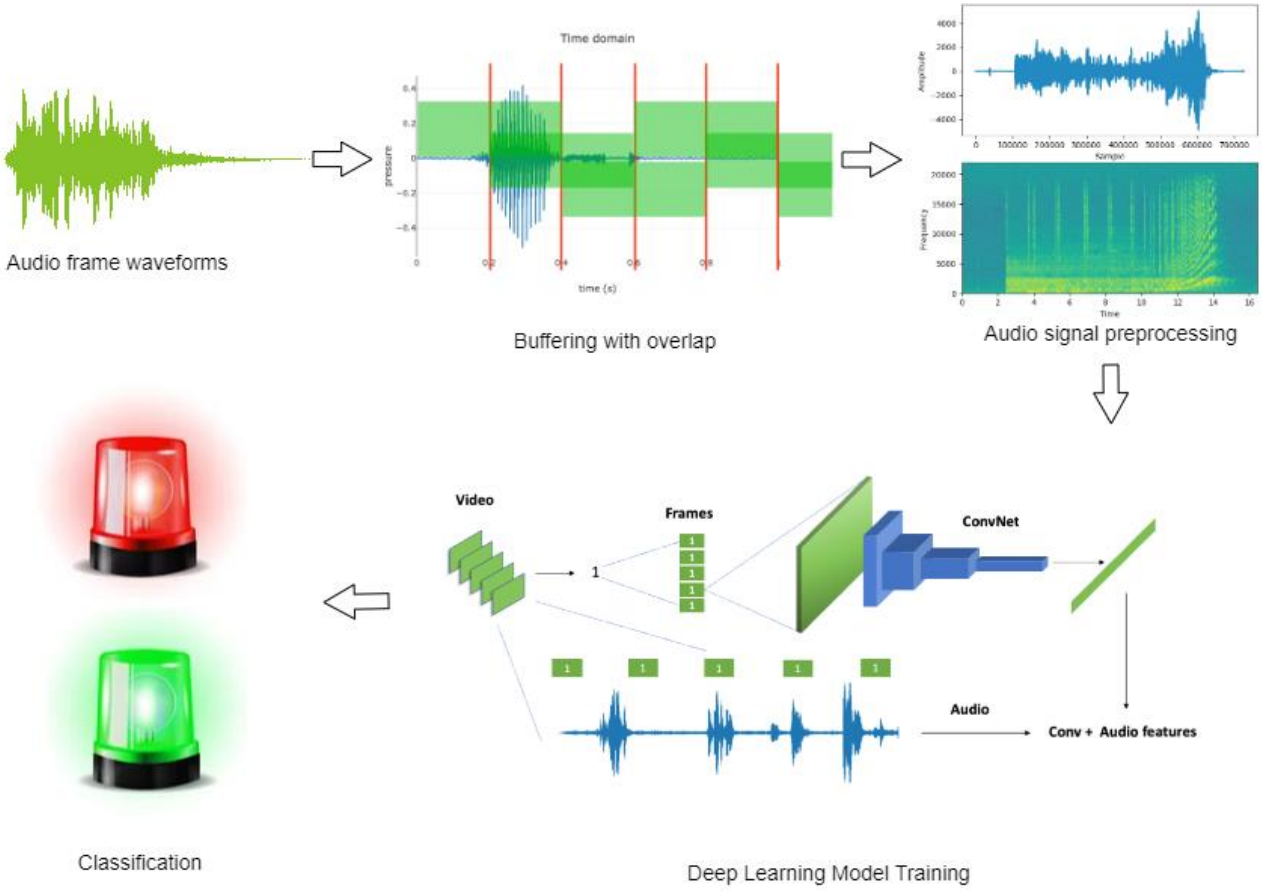


Fig. 1. Architecture of the proposed framework.

Despite the extensive nature of the ESC-50, this investigation was particularly centered on the hazardous acoustics subset, eschewing the remainder. A comparative analysis between the bespoke dataset's technical parameters and the ESC-50's original specifications can be referenced in Table I.

TABLE I. TECHNICAL PARAMETERS OF THE DEVELOPED DATASET

Characteristics	Accuracy
Overall size	661 MB
Size after preprocessing	45 MB
Number of files	2000
Number of files after preprocessing	301
Extension of files	.ogg

Within the observed region, incidents characterized by gunshots, vocal distress, and fragmented glass were identified as anomalous or "atypical." Consequently, the efficacy of the proposed framework was scrutinized, targeting its applicability in automated monitoring systems.

In pursuit of this objective, the research synthesized a dataset amalgamating diverse audio samples recorded across multifaceted environments within railway stations. This curated dataset encompassed 8,000 distinct perilous urban sound manifestations distributed across eight categorizations. The dataset's intention lies in facilitating the training and validation of both machine learning and advanced deep learning architectures in discerning and classifying hazardous urban acoustics.

Predominantly, the dataset resonated with ambient auditory elements, notably picks, gunshots, and glass rupture cues. To encapsulate the nuances of diverse operational environments, ambient acoustics were assimilated from both indoor and outdoor milieus.

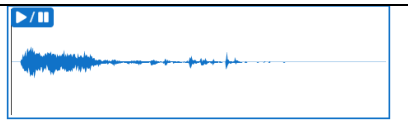
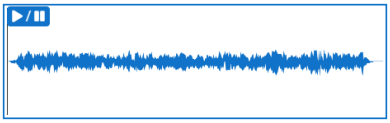
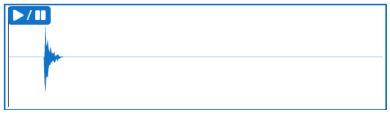
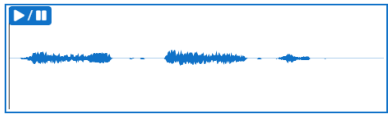
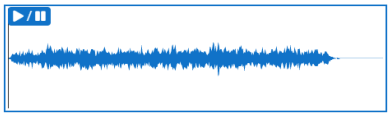
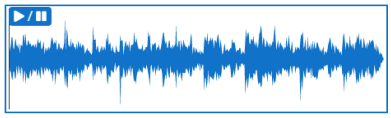
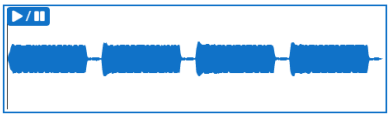
For analytical rigor, the acoustic cues were partitioned into segments lasting one second—reflecting the typical duration of the identified events of significance. Each of these segments was further dissected into 200 MS frames, exhibiting a 50% overlap. To elucidate, each one-second segment was articulated into nine distinct frames.

Table II offers a meticulous breakdown of the dataset, elucidating the composition of signals, frames, and segmented intervals. This tabulated exposition illuminates the heterogeneity of perilous urban acoustics, with an emphasis on their respective spectrograms. The table furnishes insights into the spectrographic analysis of varied impulsive acoustics, encompassing phenomena like vehicular glass rupture, canine alerts, emergency vehicular signals, infantile distress, security alarms, and various fire warning systems. This tabulation underscores the salience of the curated dataset and the pioneering CNN-LSTM deep learning paradigm.

B. Model Overview

Subsequent to initial preparations, the focus shifted towards logic programming. Central to this phase was the objective of formulating methodologies for comprehensive audio detection. The intricacies of discerning potentially alarming acoustic events can be bifurcated into two specific sub-endeavors:

TABLE II. SAMPLES OF IMPULSIVE SOUNDS IN THE DEVELOPED DATASET

Sound	Time (sec)	Spectrograms
Automobile glass shattering	3.84	
Dog barking	22.15	
Police siren	24.19	
Ambulance siren	15.41	
Constant wail from police siren	56.87	
Single gunshot	3.84	
Explosion	7.78	
Baby crying	6.66	
Burglar alarm	11.13	
Fire alarm beeping	1.41	
Fire alarm bell	1.59	
Smoke alarm	0.99	
Fire alarm yelp	2.3	

- Firstly, within the continuous audio data stream, there is a need to detect and isolate discrete pulse signals, ensuring their distinction from ambient auditory noise.
- Secondly, once extracted, the signal must then be classified, ascertaining its alignment with one of the multiple predefined acoustic events.

C. Detection of Impulsive Sound Events

The quantification of power for a series of consecutive, non-overlapping audio signal blocks serves as a cornerstone for various methodologies [9]. The computational approach to ascertain the power of the kth signal block, comprising N samples, is articulated by Eq. (1):

$$e(k) = \frac{1}{N}\sum_{n=0}^{N-1} x^2(n+kN) \qquad k = 0,1,... \tag{1}$$

Consider, for instance, an auditory manifestation of a gunshot, registered at approximately 4.6 seconds. For blocks consisting of N = 4,000 entries, corresponding to each block's duration, the power value span equates to roughly 90 milliseconds. The identification of blocks autonomously, especially in the context of transient pulse noises, can be executed via several distinct strategies, contingent upon the chosen approach:

- Grounded in the standard deviation of data that's been calibrated in terms of power metrics;
- Through the application of the median value from a median filter operating on the power units;
- By setting adaptive thresholds pertinent to the power units.

A deeper analysis reveals that this methodology predominantly leans on the standard deviation of power units' normalized values. Further examinations have deduced that normalized power block values situated within the interval [0, 1] stand as a pivotal element in this analytical schema.

$$e_{norm}(j) = \frac{e_{win}(j) - \min_j\left(e_{win}(j)\right)}{\max_j\left(e_{win}(j) - \min_j(e_{win}(j))\right)} \tag{2}$$

Following the initial processes, the focus transitioned to evaluating the standard deviation, commonly referred to as variance, for a specified set of data points:

$$\mathrm{var}(k) = \frac{1}{L-1}\sum_{j=0}^{L-2}\left[e_{norm}(j,k) = \bar{e}_{norm}(k)\right]^2 \tag{3}$$

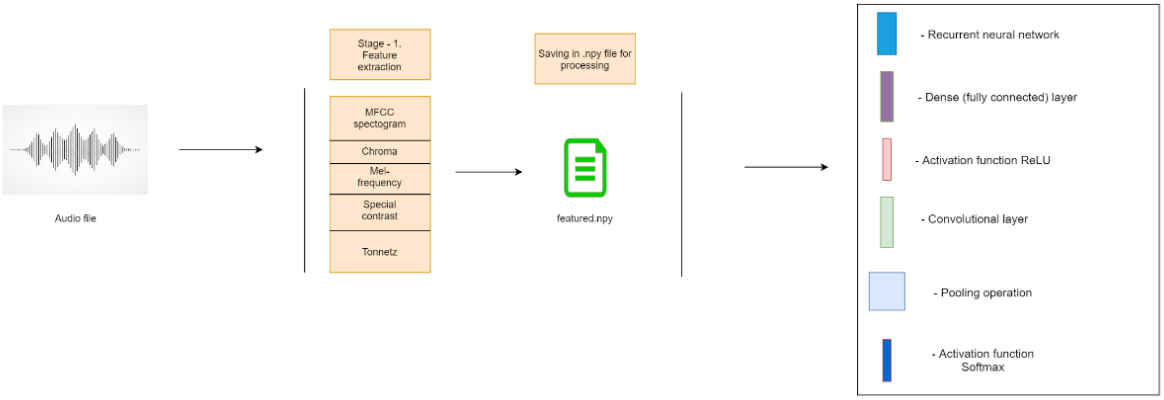
In scenarios characterized by the presence of ambient noise, block powers generally exhibit a uniform distribution within the interval [0,1] (as illustrated on the left). Upon recalibrating the power value for an audio segment to fit within this defined range, any significant deviation above the established power levels of background units triggers the automatic detection of a pulse signal. Gradual alterations in signals can be discerned by observing the average value of normalized power metrics. Notably, this methodology displays resilience in the face of fluctuations in ambient noise intensity.

D. Proposed Model

In the present research, a synergistic architecture has been postulated, integrating Convolutional Neural Networks (CNN) with Recurrent Neural Networks (RNN). In this structure, the RNN does not function as a recursive layer within the CNN. Instead, it operates independently, employing a Rectified Linear Unit (ReLU) activation for information processing. The RNN dimension is set at 128. A detailed representation of this integrated architecture can be viewed in Fig. 2.

E. Feature Extraction

In this investigation, the process of feature extraction from auditory signals spanned approximately 90 minutes, given that the dataset under scrutiny amounted to 6.6 GB. This specific size was selected intentionally, with the research aiming to assess methodologies on a comparatively modest dataset. In subsequent phases, the established techniques will be applied to data in a singular pass. Following the comprehensive analysis of the auditory files, a resultant set of 8,674 sounds, equating to a cumulative duration of 5,439 seconds or 90.65 minutes, was obtained. A closer examination of the feature extraction component can be understood by referring to the coding segment, and the entire process is graphically represented in Fig. 3.



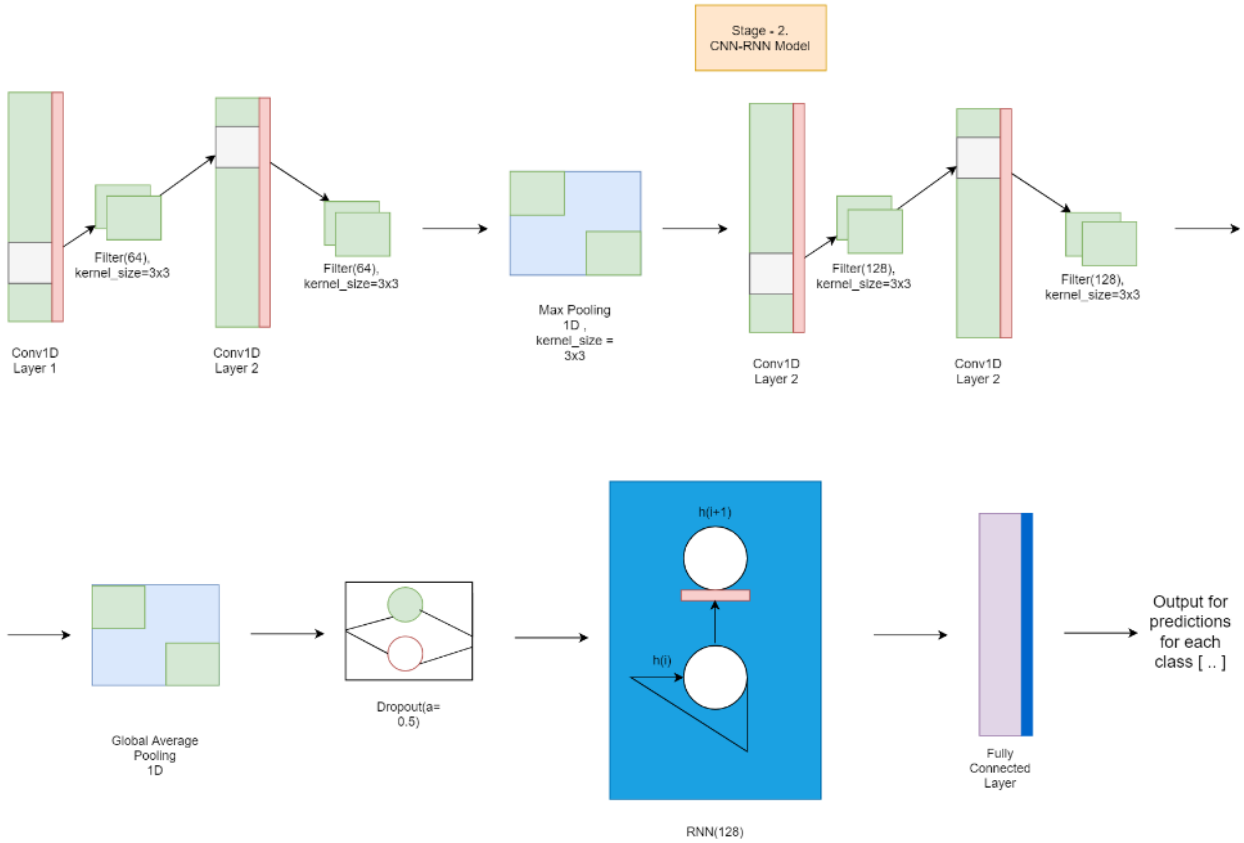


Fig. 2. Architecture of the proposed model.

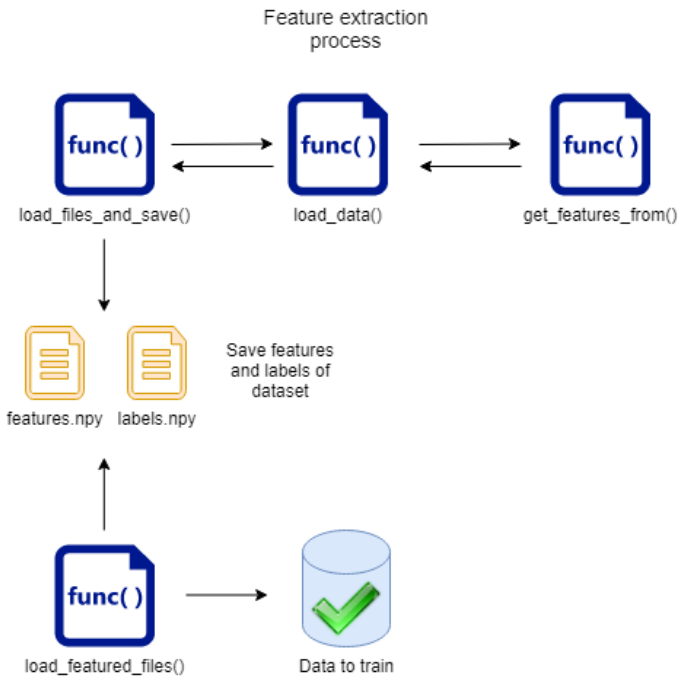


Fig. 3. The proposed framework.

Within the scope of this research, four distinct functions were delineated. Among them, three were explicitly designated for the extraction of features and subsequent data preservation. To elaborate:

1) The load_files_and_save() function invoked load_data(). This latter function systematically iterated over the dataset to acquire individual sound samples.

2) Following this, get_features_from() was summoned for each sound sample to extract its pertinent features.

3) Post-extraction, these attributes, along with their corresponding labels, were committed to persistent storage in two separate .npy formatted files.

Subsequent to the storage phase, data retrieval was facilitated by the load_featured_files() function. This prepared the dataset for the training phase, utilizing the integrated RNN-CNN model previously delineated in Fig. 2. This model's architecture encompassed two convolutional layers: one derived from a global maximum pooling mechanism and the other from a global average pooling paradigm.

IV. Experimental Results

This segment elucidates the empirical outcomes derived from employing the synergized CNN-LSTM model for the identification of hazardous urban auditory events. Initially, the

metrics tailored for appraising the efficacy of the aforementioned deep learning model are delineated. This is succeeded by a presentation of the results from both training and testing phases, encompassing model accuracy, associated losses, the distinct confusion matrices, and AUC-ROC curve impulsive sound classification.

A. Evaluation Metrics

The efficacy of any machine learning or deep learning model, especially in contexts like hazardous urban sound detection, necessitates a rigorous and comprehensive evaluation strategy. This section elucidates the primary metrics employed to assess the proposed CNN-LSTM model's performance:

Accuracy: This is the most fundamental metric, representing the ratio of correctly predicted instances to the total instances in the dataset [30]. It provides a broad understanding of the model's effectiveness, encapsulating its capacity to correctly identify both dangerous and non-dangerous sounds.

$$accuracy = \frac{TP + TN}{P + N} \tag{4}$$

Precision: Precision ascertains the proportion of true positive predictions in the pool of all positive predictions [31]. In the realm of urban sound detection, high precision denotes that the sounds flagged as 'dangerous' by the model are indeed perilous with minimal false alarms.

$$preision = \frac{TP}{TP + FP} \tag{5}$$

Recall (or Sensitivity): This metric quantifies the model's ability to correctly identify all potential hazards [32]. In essence, recall measures the fraction of actual dangerous sounds that were rightly detected by the model.

$$recall = \frac{TP}{TP + FN} \tag{6}$$

F-score: Harmonizing precision and recall, the F-score is the harmonic mean of these two metrics [33]. It assists in providing a balanced measure, especially when the class distribution is skewed. An optimal model would strive for a high F-score, indicating both robust precision and recall.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{6}$$

Utilizing these metrics provides a holistic understanding of the model's capabilities, ensuring it is adept at identifying genuine threats while minimizing false alarms.

B. Results

This section offers a comprehensive assessment of the experimental results emanating from the dangerous urban sound detection exercises. The results of the CNN-LSTM model's endeavor at discerning impulsive sounds are visually illustrated in Fig. 4 and Fig. 5.

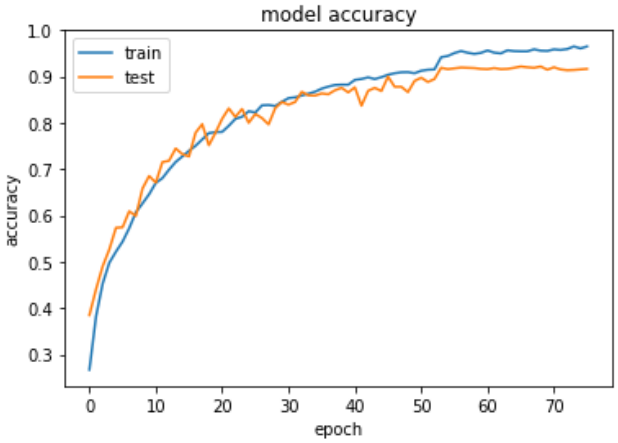


Fig. 4. Model training accuracy.

Fig. 4 specifically delineates the performance metrics during both training and testing phases for the inaugural dataset furnished by the research team. Notably, the CNN-LSTM model exhibited an impressive proficiency, registering an accuracy rate of 95% during its training phase. This was achieved over an approximate span of 80 epochs. Diving deeper into the model's architecture, it was observed to have approximately 87,822 parameters.

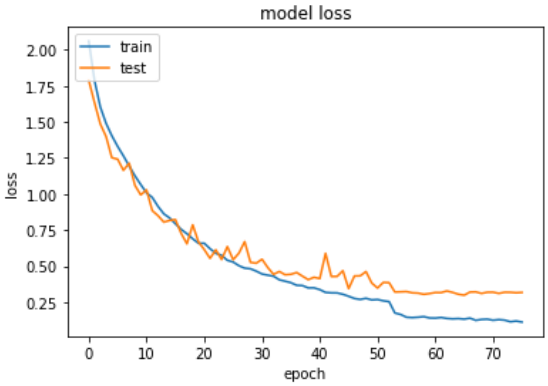


Fig. 5. Model training loss.

Furthermore, the duration of the training phase offers insight into the model's computational efficiency. The entire training process was completed in roughly 267 seconds, translating to slightly more than four minutes. This timeline underscores not only the model's accuracy but also its expedient processing capabilities.

In tandem, precision, recall, and F-score – though not explicitly detailed in the current dataset visualizations – remain paramount for comprehensive model assessment. These metrics, when considered in conjunction, ensure a holistic appreciation of the model's true capability, especially in real-world urban soundscapes.

In this section, we turn our focus to the findings from the second dataset, sourced from an open repository, and their implications as demonstrated in Fig. 6.

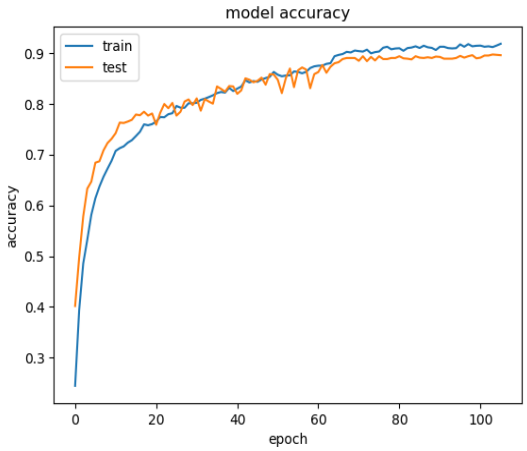


Fig. 6. Model test accuracy.

The CNN-LSTM model's performance on the second dataset is showcased in Fig. 7, providing invaluable insights into its adaptability and precision. Over a span of approximately 110 epochs, the model achieved an accuracy of 92%. This underscores its consistent performance, even when confronted with potentially disparate sound data from varied sources.

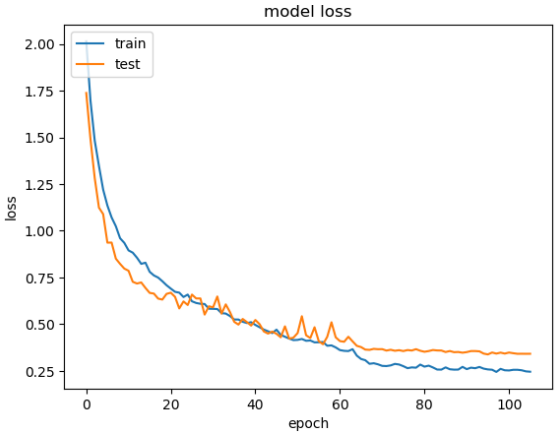


Fig. 7. Model test loss.

In the post-training phase, the model's performance was quantified using a confusion matrix, providing clarity on its precision across various classifications—specifically delineating true positives, true negatives, false positives, and false negatives, in the context of diverse urban acoustics. Fig. 8 presents a graphical representation of this matrix, elucidating the categorization efficacy for impulsive sounds. The implemented CNN model facilitated the differentiation of eight distinct hazardous urban auditory signals.

Fig. 9 depicts the AUC-ROC curve pertinent to the detection of perilous auditory events. The curve provides insights into the model's sensitivity to variations within the training dataset. The outcomes suggest that the integrated CNN-LSTM framework adeptly discerns hazardous acoustic events with commendable precision. Observations from the graph indicate a consistent performance, signifying the model's robust training tailored specifically for the identification of hazardous acoustical scenarios.

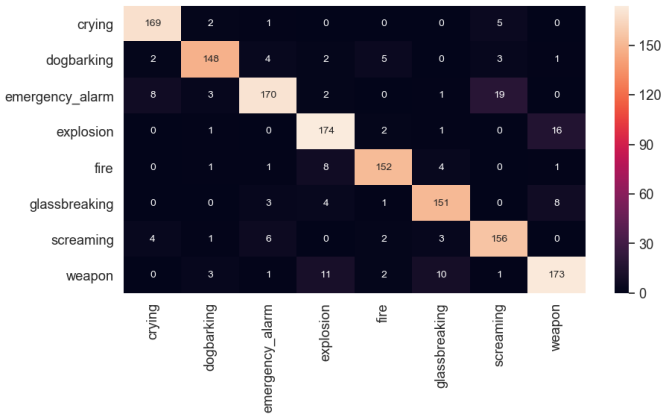


Fig. 8. Confusion matrix.

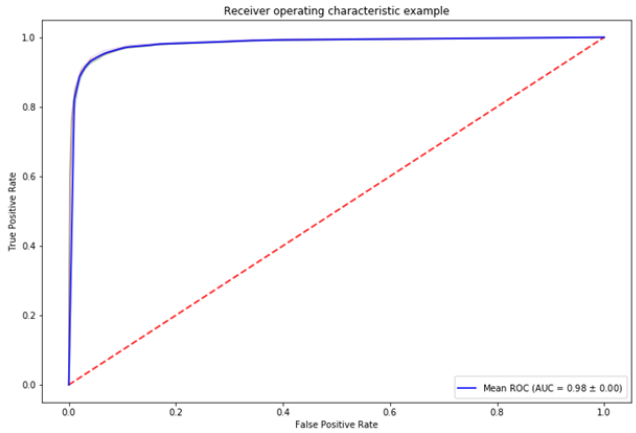


Fig. 9. ROC-AUC curve.

Consequently, the introduced deep neural network exhibits superior efficacy in consistently detecting hazardous urban sounds across all evaluation metrics. The success of the proposed methodology may be attributed to the utilization of the advanced RNN-CNN for weight and bias adjustments, coupled with an optimized training duration. The findings indicate that the presented deep neural network can be readily adapted to cater to both concise and extensive auditory inputs in contemporary applications.

V. Discussion

The task of detecting and classifying dangerous urban sounds using deep learning architectures has garnered considerable attention given the importance of public safety and efficient urban management. This study presented a novel Convolutional LSTM (CNN-LSTM) network specifically tailored for real-time impulsive sound detection and classification in urban settings. The outcome of this research offers significant insights and implications, which are discussed in this section.

A. Comparative Performance of the Proposed Model

The CNN-LSTM architecture, as revealed by the results, showcases a notable advancement over previously proposed models for urban sound detection. In comparison to standard CNN architectures, the introduction of RNN allows the model to effectively process temporal sequences in the auditory data,

proving its prowess in capturing the temporal dynamics inherent in sound samples [34]. Not only does this substantiate the architectural choices made in this research but it also suggests potential avenues for further refining and expanding the hybrid deep learning models in auditory signal processing.

B. Efficacy in Hazardous Sound Detection

A pivotal achievement of this study is the high classification accuracy for sounds that have immediate safety implications, such as gunshots, alarms, and screams [35]. Precision in detecting these sounds is crucial for real-time monitoring systems that aim to promptly respond to emergencies. The proposed model's ability to discern these sounds from a cacophony of urban noises, with significant accuracy, positions it as a strong candidate for deployment in urban surveillance systems.

C. Model Generalizability and Robustness

Another salient point worth discussing is the model's performance across diverse datasets [36]. Its consistent results, even with different datasets including open-source ones, indicate robustness and generalizability. The implications here are two-fold: First, the model appears to be resilient to overfitting [37], a frequent pitfall in deep learning paradigms. Second, its generalizability suggests that with minor modifications [38], the proposed architecture could potentially be employed in varied urban environments, extending beyond the specific settings of this study.

D. Computational Efficiency and Real-time Implementation

The study indicates that the model's training lasted a mere few minutes, emphasizing the computational efficiency of the CNN-LSTM architecture. This is crucial for scaling up the approach and integrating it into real-time surveillance systems, where rapid model training and updating are of essence. Given the emergent nature of urban sounds and the ever-evolving urban landscape, the ability to quickly train and retrain models can be a game-changer.

E. Challenges and Limitations

While the findings are promising, it is essential to acknowledge certain challenges. Ambient noises, characteristic of dynamic urban settings [39], can sometimes interfere with the accurate detection of impulsive sounds. Furthermore, while the model has been tested on selected datasets, its performance in other global urban contexts – each with its unique soundscapes – remains to be evaluated.

F. Future Research Directions

Several prospective avenues emerge from this study:

- Data Augmentation: Experimenting with more extensive and diverse datasets, inclusive of global urban soundscapes, could further test and improve the model's robustness.
- Model Refinements: While the CNN-LSTM architecture demonstrates efficacy, the integration of attention mechanisms might enhance its ability to focus on critical sound segments, thereby potentially improving accuracy.
- Transfer Learning: Given the computational efficiency of the proposed model, it would be intriguing to investigate the benefits of transfer learning, applying knowledge from pre-trained models to expedite the training process even further.
- Integration with Visual Surveillance: A holistic urban surveillance system could combine auditory cues from the CNN-LSTM model with visual data from CCTV cameras, enhancing the accuracy and response time of emergency systems.

In conclusion, the CNN-LSTM model's performance in detecting dangerous urban sounds signals a promising step forward in urban surveillance and safety systems. Its computational efficiency, robustness, and high accuracy across datasets underpin its potential for real-world applications. Nevertheless, like all research, it sets the stage for further inquiries, refinements, and innovations in this domain.

VI. Conclusion

The paramount importance of ensuring urban safety cannot be overstated, and the deployment of advanced technological measures is crucial in these endeavors. This research aimed to bridge the extant gaps in urban sound detection by proposing a novel Convolutional LSTM (CNN-LSTM) architecture tailored for real-time impulsive sound detection and classification. The results, as delineated in the study, highlight the efficacy of this hybrid model in discerning and classifying hazardous urban sounds amidst the complex soundscape of urban environments.

The model's comparative performance, evidenced by its high classification accuracy, demonstrates its potential utility for urban surveillance systems. Especially noteworthy is its ability to accurately detect sounds of immediate safety concern, such as alarms, screams, and gunshots. Moreover, the robustness and generalizability of the model, as indicated by its consistent performance across diverse datasets, fortify its position as a leading contender for wide-scale implementation in urban settings globally.

However, while the findings undoubtedly underscore the potential of the CNN-LSTM architecture, they also pave the way for future research. There remains a vast expanse of uncharted territory in this domain, especially concerning model refinements, transfer learning, and the integration of auditory and visual surveillance cues. Such advancements could further refine detection accuracy and foster comprehensive urban safety measures.

In sum, this study serves as a testament to the untapped potential of deep learning paradigms in enhancing urban security. The proposed CNN-LSTM model, with its impressive results, sets a foundational precedent for further innovations and refinements. As urban centers continue to grow and evolve, it is our sincere hope that the fruits of this research contribute to safer, more secure, and harmonious urban living experiences for all.

Acknowledgment

This work was supported by the research project — Development of a system for detecting and alerting dangerous

events based on the audio analysis and machine learning funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan. Grant No. IRN AP19175674.

REFERENCES

[1] J. Bajzik, J. Prinosil, R. Jarina and J. Mekyska, "Independent channel residual convolutional network for gunshot detection," International Journal of Advanced Computer Science and Applications, vol. 13, no.4, pp. 950-958, 2022.

[2] K. M. Nahar, F. Al-Omari, N. Alhindawi and M. Banikhalaf, "Sounds recognition in the battlefield using convolutional neural network," International Journal of Computing and Digital Systems, vol. 11, no.1, pp. 189-198, 2022.

[3] I. Estévez, F. Oliveira, P. Braga-Fernandes, M. Oliveira, L. Rebouta et al., "Urban objects classification using Mueller matrix polarimetry and machine learning," Optics Express, vol. 30, no.16, pp. 28385-28400, 2022.

[4] Z. Peng, S. Gao, Z. Li, B. Xiao, Y. Qian, "Vehicle safety improvement through deep learning and mobile sensing" IEEE Network, vol. 32, no.4, pp. 28-33, 2018.

[5] Y. Wei, L. Jin, S. Wang, Y. Xu and T. Ding, "Hypoxia detection for confined-space workers: photoplethysmography and machine-learning techniques," SN Computer Science, vol.3, no.4, pp.1-11, 2022.

[6] Y. Arslan, H. Canbolat, "Sound based alarming based video surveillance system design," Multimedia Tools and Applications, vol. 81, no.6, pp. 7969-7991, 2022.

[7] K. Pawar, V. Attar, "Deep learning approaches for video-based anomalous activity detection," World Wide Web, vol. 22, no.2, pp.571-601, 2019.

[8] S. U. Amin, M. S. Hossain, G. Muhammad, M. Alhussein, M. A. Rahman, "Cognitive smart healthcare for pathology detection and monitoring," IEEE Access, vol. 7, no.1, pp. 10745-10753, 2019.

[9] Omarov, B., & Altayeva, A. (2018, January). Towards intelligent IoT smart city platform based on OneM2M guideline: smart grid case study. In 2018 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 701-704). IEEE.

[10] C. Heipke, F. Rottensteiner, "Deep learning for geometric and semantic tasks in photogrammetry and remote sensing," Geo-spatial Information Science, vol. 23, no.1, pp. 10-19, 2020.

[11] Omarov, B., Altayeva, A., & Cho, Y. I. (2017). Smart building climate control considering indoor and outdoor parameters. In Computer Information Systems and Industrial Management: 16th IFIP TC8 International Conference, CISIM 2017, Bialystok, Poland, June 16-18, 2017, Proceedings 16 (pp. 412-422). Springer International Publishing.

[12] A. Rajbanshi, D. Das, V. Udutalapally, R. Mahapatra, "DLeak: an IoT-based gas leak detection framework for smart factory," SN Computer Science, vol. 3. no.4, pp. 1-12, 2022.

[13] Y. Arslan and H. Canbolat, "Sound based alarming based video surveillance system design," Multimedia Tools and Applications, vol. 81, no. 6, pp. 7969-7991, 2022.

[14] R. Sun, Q. Cheng, F. Xie, W. Zhang, T. Lin et. al., "Combining machine learning and dynamic time wrapping for vehicle driving event detection using smartphones," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no.1, pp.194-207, 2019.

[15] G. Chen, F. Wang, S. Qu, K. Chen, J. Yu et. al., "Pseudo-image and sparse points: vehicle detection with 2D LiDAR revisited by deep learning-based methods," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no.12, pp. 7699-7711, 2020.

[16] Omarov, B., Omarov, B., Shekerbekova, S., Gusmanova, F., Oshanova, N., Sarbasova, A., ... & Sultan, D. (2019). Applying face recognition in video surveillance security systems. In Software Technology: Methods and Tools: 51st International Conference, TOOLS 2019, Innopolis, Russia, October 15–17, 2019, Proceedings 51 (pp. 271-280). Springer International Publishing.

[17] V. Osipov, N. Zhukova, A. Subbotin, P. Glebovskiy, E. Evnevich, "Intelligent escalator passenger safety management," Scientific Reports, vol. 12, no.1, pp.1-16, 2022.

[18] I. H. Peng, P. C. Lee, C. K. Tien, J. S. Tong, "Development of a cycling safety services system and its deep learning bicycle crash model," Journal of Communications and Networks, vol. 24, no. 2, pp. 246-263, 2022.

[19] Sultan, D., Omarov, B., Kozhamkulova, Z., Kazbekova, G., Alimzhanova, L., Dautbayeva, A., ... & Abdrakhmanov, R. (2023). A Review of Machine Learning Techniques in Cyberbullying Detection. Computers, Materials & Continua, 74(3).

[20] L. M. Bine, A. Boukerche, L. B. Ruiz, A. A. Loureiro, "Leveraging urban computing with the internet of drones," IEEE Internet of Things Magazine, vol. 5, no.1, pp. 160-165, 2022.

[21] S. Khan, L. Alarabi and S. Basalamah, "Toward smart lockdown: a novel approach for COVID-19 hotspots prediction using a deep hybrid neural network," Computers, vol. 9, no. 4, pp. 1-16, 2020.

[22] M. Dua, D. Makhija, P. Manasa and P. Mishra, "A CNN–RNN–LSTM based amalgamation for Alzheimer's disease detection," Journal of Medical and Biological Engineering, vol. 40, no. 5, pp. 688-706, 2020.

[23] H. Gill, O. Khalaf, Y. Alotaibi, S. Alghamdi and F. Alassery, "Multi-model CNN-LSTM-LSTM based fruit recognition and classification," Intelligent Automation & Soft Computing, vol. 33, no. 1, pp. 637-650, 2022.

[24] K. Chandriah and R. Naraganahalli, "RNN/LSTM with modified Adam optimizer in deep learning approach for automobile spare parts demand forecasting," Multimedia Tools and Applications, vol. 80, no. 17, pp. 26145-26159, 2021.

[25] Tursynova, A., & Omarov, B. (2021, November). 3D U-Net for brain stroke lesion segmentation on ISLES 2018 dataset. In 2021 16th International Conference on Electronics Computer and Computation (ICECCO) (pp. 1-4). IEEE.

[26] Y. Xue, P. Shi, F. Jia, H. Huang, "3D reconstruction and automatic leakage defect quantification of metro tunnel based on SfM-Deep learning method," Underground Space, vol. 7, no.3, pp. 311-323, 2022.

[27] L. Zhang, L. Yan, Y. Fang, X. Fang, X. Huang, "A machine learning-based defensive alerting system against reckless driving in vehicular networks," IEEE Transactions on Vehicular Technology, vol. 68, no.12, pp.12227-12238, 2019.

[28] A. M. Youssef, B. Pradhan, A. Dikshit, M. M. Al-Katheri, S. S. Matar et. al., "Landslide susceptibility mapping using CNN-1D and 2D deep learning algorithms: comparison of their performance at Asir Region, KSA," Bulletin of Engineering Geology and the Environment, vol. 81, no.4, pp. 1-22, 2022.

[29] S. Asadianfam, M. Shamsi, A. Rasouli Kenari, "Hadoop Deep Neural Network for offending drivers," Journal of Ambient Intelligence and Humanized Computing, vol. 13, no.1, pp. 659-671, 2022.

[30] L. M. Koerner, M. A. Chadwick, E. J. Tebbs, "Mapping invasive strawberry guava (Psidium cattleianum) in tropical forests of Mauritius with Sentinel-2 and machine learning," International Journal of Remote Sensing, vol. 43, no.3, pp. 841-872, 2022.

[31] D. K. Dewangan, S. P. Sahu, "Deep learning-based speed bump detection model for intelligent vehicle system using raspberry Pi," IEEE Sensors Journal, vol. 21, no.3, pp. 3570-3578, 2020.

[32] Z. Fang, B. Yin, Z. Du and X. Huang, "Fast environmental sound classification based on resource adaptive convolutional neural network," Scientific Reports, vol. 12, no. 1, pp. 1-18, 2022.

[33] V. Gugnani, R. K. Singh, "Analysis of deep learning approaches for air pollution prediction," Multimedia Tools and Applications, vol. 81, no.4, pp. 6031-6049, 2022.

[34] X. Yang, L. Shu, Y. Liu, G. P. Hancke, M. A. Ferrag et. al., "Physical security and safety of IoT equipment: a survey of recent advances and opportunities," IEEE Transactions on Industrial Informatics, vol. 18, no.7, pp. 4319-4330, 2022.

[35] H. Kyle, P. Agarwal, J. Zhuang, "Monitoring misinformation on Twitter during crisis events: a machine learning approach," Risk Analysis, vol. 42, no.8, pp. 1728-1748, 2022.

[36] M. Esmail Karar, O. Reyad, A. Abdel-Aty, S. Owyed, M. F. Hassan, "Intelligent iot-aided early sound detection of red palm weevils," Computers, Materials & Continua, vol. 69, no.3, pp. 4095–4111, 2021.

[37] T. Thomas Leonid and R. Jayaparvathy, "Classification of elephant sounds using parallel convolutional neural network," Intelligent Automation & Soft Computing, vol. 32, no.3, pp. 1415–1426, 2022.

[38] Z. Ma, G. Mei, , F. Piccialli, "Machine learning for landslides prevention: a survey," Neural Computing and Applications, vol. 33, no.17, pp. 10881-10907, 2021.

[39] X. Zhao, L. Zhou, Y. Tong, Y. Qi and J. Shi, "Robust sound source localization using convolutional neural network based on microphone array," Intelligent Automation & Soft Computing, vol. 30, no. 1, pp. 361–371, 2021.

Breast Cancer Detection System using Deep Learning Based on Fusion Features and Statistical Operations

Suleyman A. AlShowarah
Faculty of Information Technology, Mutah University, Karak, Jordan

***Abstract*—Breast cancer is considered as the second cause of death for women. The earlier is diagnosed, the easier the patients can be recovered. The need for studies to detect this kind of cancer easily and accurately came from the growing rate of infected patents by breast cancer exponentially. This study is conducted to investigate the use of deep-learning model for breast cancer detecting using the technique VGG-19 and ultrasound images. Two layers of VGG19 structure were used: (i.e. fc6 and fc7. Based on these two layers (fc6 and fc7), new datasets were created, which are named as statistical operations. These datasets will be employed as input for the following Machine Learning classifiers: K-Nearest Neighbors, Random Forest, Naïve Bayes and Decision Tree. Data augmentation was considered to increase the dataset size for better learning of CNN. Random Forest achieved high accuracy (88.63), precision (0.88), recall (0.88) and F-Measure (0.88). The results of the classification accuracy in the three scenarios are slightly similar; this proves that the breast cancer can be detected even if the size of data in the training dataset was minimal.**

Keywords*—Breast cancer detection; breast cancer classification; deep learning; vgg-19; breast tumor*

I. INTRODUCTION

Cancer is considered as an uncontrolled growth of cells in human body. Breast cancer is one type of cancers, which considered as the second cause of death for women. It is known for patients and doctors, the earlier the cancer is diagnosed and detected, the easier the patients can be recovered. Because of the growing rate of infected by breast cancer exponentially [1]; there is need for studies to detect this kind of cancer easily and accurately. It is considered as a motivation for such study in that the diagnosed people in this kind of cancers is growing day by day. For example, in the US [2], most women are diagnosed with breast cancer compared to any other type of cancer, except for skin cancer. This cancer affects one in three of new female annually. In a year 2023, the estimated diagnosed women in US with invasive breast cancer to be 297,790, and with non-invasive breast cancer will be 55,720 [2]. Although these statistical numbers reflect what is obtainable in most advanced economies, and it was illustrated by studies that about 58% of deaths occur because of this cancer in less advanced countries. The high death rate from breast cancer is because of the lack of early detection, as more than 33% are for population aged 30-49 and 81% for 30-59 years old are diagnosed for this cancer [3], [4], [5].

In order for early detection and saving lives of breast cancer, a mammography was developed by scientists with some limitations of its functions. Despite of that, some of studies showed a reduction in death rate of about 40% after a mammogram screening [6, 7]. About 15 of the 1,000 women seen with mammography are recommended for a biopsy, and about 13 women of these biopsies show a false positive results (not present) [8]. A major limitation of mammographic screening was highlighted by C.K. et al. [9]: breast cancers of prognostic significance are not diagnosed.

Several of strategies were implemented to enhance the performance of screening mammography: including double checking and screening at annual intervals [10], apply two views for each breast [11], and make a comparison with previous mammograms [12]. The serious features can be detected by radiologists for each scan such as architectural distortions, micro-calcifications, and asymmetries as cancer biomarkers or cancer risk. Detecting these serious features manually leads to additional costs and will let the radiologists pay more efforts for mammography [13]. One of systems emerged in the 1990s named as Computer-aided detection (CAD), this is to detect and then classify breast cancers in mammograms automatically. But still the performance of such these traditional systems has not improve screening process significantly, this is mainly due to their lack specificity [14,15]. Specificity relates to how abnormalities can be discriminated by algorithms when screening, which differs from how it is diagnosis; it employs causal inference as to the origin of the abnormality. However, detecting anomalies in screening mammographs is important in the diagnosis.

Recently, the researchers in [16] reported that novel algorithms based on CNN can be used to improve the performance of screening mammography and also to increase the efficiency of mammography professionals. In this matter, some of researchers developed different CNNs-based algorithms for automated mammographic analysis purposes [17].

This study aims to detect and classify the images of breast cancer using deep-learning, which can be employed as system used to help doctors and radiologists in their diagnosis automatically. To achieve this aim, it will be conducted based on CNN using VGG-19. The pre-trained technique VGG-19 is used to achieve high accuracy by finding distinctive details features of image [18] [19]. The two layers of the VGG19 structure were used (i.e. layer 6; which called fc6, and layer 7; which called fc7), and each contains 4096 features. Also, more feature vectors were created from (fc6 and fc7), which named

as statistical operations. Statistical operations are used to generate more datasets using Average (Avg), Minimum (Min), Maximum (Max), and fusion between fc6 and fc7. All aforementioned datasets will be used as input for the processes of classification using different algorithms (i.e. K-Nearest Neighbors (KNN), Random Forest (RF), Naïve Bayes (NB) and Decision Tree (DT)).

The results illustrated that Random Forest algorithm achieved high accuracy (88.63), precision (0.88), recall (0.88) and F-Measure (0.88) for fc7 of second scenario. The results were slightly similar; this approves that these features can provide a better accuracy when used in detection studies. Also, the results of the classification accuracy in the three scenarios are slightly similar; this proves that the breast cancer can be detected even if the number of images in the training dataset were minimal.

The motivation of conducting this study is represented by:

1) The literature need for researches of detecting breast cancer using CNN with new model like VGG-19 based on the two layers: fc6 and fc7.

2) To the best of my knowledge, I could not find any research in the literature performed based on statistical operations for detecting breast cancer using VGG-19.

3) Based on features that were extracted by VGG-19, it can be a contribution for this study by providing the literature with a differentiation between the results of different aforementioned classifiers and with different three scenarios.

This research is designed into five sections. The overview of related studies in literature is introduced in Section II. Section III discusses the methodology for the proposed model and the experiment design. Then, the experimental results and discussion are discussed in Section IV. Finally, Section V presents the conclusion.

II. Literature Review

Several of studies showed several of automated, and computer vision approaches to classify breast cancer- based images [20, 21]. Some of them have focused on segmentation process, and then features were extracted from images [22]. While in some other studies followed the pre-processing steps for better feature extraction, this is to improve the contrast in the images and then to detect the infected part of image [23]. For example, the most important and the first of the pre-processing steps in the mammogram analysis are applying the segmentation for the infected region, which allows focusing on region of interest in the images. The researchers in [24] applied the technique called texture filter in the segmentation of the breast region.

Lastly, in this matter, there was a study conducted by de Vos, et al. [25] who implemented DL for extraction features for region of interest from cancer images. In their study, they used three techniques of convolutional neural network (ConvNet) to detect and to extract features from a 3D image, which are: the presence of the anatomical structure of interest in the following: 1) axial, 2) coronal, 3) sagittal slices. The method of their localization was compared to the manual method using the distances between the centroids and the walls with an automatically and manually defined reference frame. Many other researches have adapted a pure deep learning based on its layers for extraction features [26–30], and also using one of most interesting methods such a high pass isotropic filter [31].

Image cropping aims to enhance image quality by removing distracting content/also adding aesthetics, which are mainly categorized-based on that. Different methods are available to achieve such this task. Often, these methods can apply techniques like: machine learning, deep learning, segmentation, saliency-based, and sparse coding. For example, the study conducted by Mishra et al. [32] used ML radiology using classification pipeline. They segmented the region of interest, and then extracted the useful features. Their study was performed on the dataset: (BUSI), and the results showed improving in classification accuracy. While the study conducted by Byra [33] used DL for the classifying the cancer parts from images. The transfer learning (TL) was used and then deep representation scaling (DRS) layers were added between the blocks of pre-trained CNNs to enhance the provided information. In order to analyze these parts classification, the enhancement was only for the parameters of the deep representation scaling layers during training, this is to enhance the pre-trained CNNs, which was much better compared to other techniques.

Some of researchers in [34] developed algorithm: Dilated Semantic Segmentation Network (Di-CNN) and then they used it to detect and classify the breast cancer. The pre-trained DenseNet201 deep model was used in their work and then trained using TL that was used for feature extraction. In addition, they applied a 24-layer CNN and fusion features in their work. The results of the fusion process have improved the classification accuracy in the detection process.

Ahmed et al. [35] used patch selection to classify breast tissue based images using TL. The features were extracted using CNN to discriminate patches, which are an input for an Efficient-Net architecture that is considered as an architecture of CNN and employed for scaling technique that scales all dimensions of: width, depth, and resolution based on a compound coefficient; these input were trained on the dataset: ImageNet. The classifier support vector machine was used for classifying features that were extracted from the Efficient-Net. The results showed that the suggested model achieved better results compared to the standard methods.

The use of DL has outperformed most recent methods. A good example of this, it is the study in [35], which built based on the geometric properties of the edge features to extract the abnormal patches structures in the expected regions. For example, the researchers in [36] conducted a study and the features were extracted from the image using CNN (DenseNet); which are then provided in fully connected layers (FC) for classifying the benign and cancerous cells of breast cancer image. However, the researchers in [37] presented deep learning methods for detecting and classifying models. Deep learning can be used perfectly for computer vision problems, especially image optimization and interpretation. This has led to a wave of pioneering applications of medical imaging, and

available databases of image have presented the growth of DL algorithms aimed to detect cancer images.

Other researchers in [38] conducted a study to extract the most important and useful information using DL including convolution layers for breast cancer detection. They showed that the features extracted by DL models are better in term of accuracy than traditional and manual methods. DL methods showed that it has ability to detect pathological forms of cancer that were previously thought to be difficult to diagnose using conventional and manual methods. The researchers in [39] presented DL-based techniques for detection breast cancer images. In their work, a dataset was published containing canine mammary tumor (CMT). Also, (VGG-16) was used to investigate the performance of hybrid frameworks using different algorithms on CMT and other datasets of breast cancer.

The works that have been done in the literature on DL using CNN among others, have provided an insight to the researchers on how an automatic representation method without supervised descriptors in extracting, i.e. independent of any human intervention that could influence these representations [40].

In fact, Deep CNNs usually have too large number of parameters, so that it is not reasonably trained without a very large dataset. Moreover, medical datasets are usually not large enough to adequately train a deep CNN model from scratch. Thus, transfer learning in deep learning was explored to be used in the medical imaging to solve such problem. So, the transfer learning transfers knowledge between large source and small target domains [41], which can be done by using pre-training a CNN model with the source of images, then re-training parts of the model with the target images. In [42], the researchers used CNN AlexNet to detect the images of breast cancer from dataset named: BreakHis [43]. Their results showed that the classification accuracy is 79.85%.

While the researchers Han et al. in [44] have proposed a framework for breast cancer multi-classification using class structure-based deep CNN model on the dataset named: BreakHis. The results showed 93.2% of classification accuracy.

The researchers Nuh et al. in [45] have conducted a study to discriminate between samples infected and non-infected breast cancers images based on CNN using different spatial patches. The results showed for window sizes: 5x5 and 7x7 are 86.91% and 86.17% respectively.

The researchers in [46] have presented a CNN classifier for the visual analysis of area of cancer in images of malignant breast cancer. The results showed a higher performance for their proposed classifier compared to random forest classifier: 84.23% classification accuracy. It was approved by results of Hafemann et al. [47] that used a CNN; It showed better results compared to traditional approach that always needs huge efforts and effective expert in the field of knowledge [48].

III. Methodologies

This section presents the dataset design, and the experimental setup model.

A. Database Design

The Dataset of Breast Ultrasound Images (Dataset BUSI) is used in this research and can be obtained online [49]. The dataset consists of 780 images with size 500×500 pixels; including the segmentation masks that refers to 600 patients. The dataset consists of three classes: normal, malignant, and benign. The whole dataset was divided into training and test dataset. However, this is not enough as dataset to train data using the model of deep learning; therefore, a data augmentation step is achieved to increase the dataset size for better learning of CNN. These implemented steps are achieved multiple times until the size of dataset of each class has reached 5872.

B. Experimental Setup Model

The experiment of this study was designed based on three scenarios: 1) 50% for training and 50% for testing, 2) 70% for training and 30% for testing, and 3) 80% for training and 20% for testing. So, for each scenario - the experimental setup for the proposed model is displayed in Fig. 1 and consists of set of steps, as follows:

Step 1: The MATLAB is used for automatically extracting feature form images based on Pre-trained VGG-19. The outputs are two datasets for fc6 and fc7. Each of fc6 and fc7 contains 4096 features. These datasets will be used in the step 3.

Step 2: Creating a new dataset from step 1 by performing the statistical operations (i.e. Avg, Min, Max, and fusion of fc6 and fc7). These datasets will be used also in step 3.

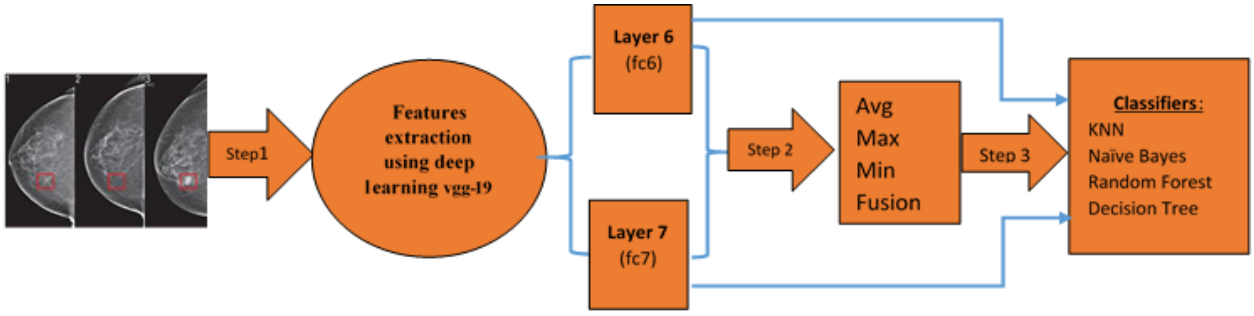


Fig. 1. Proposed experimental setup model.

The explanation for the statistical operations is in the following: 1) Max: It is the largest value of fc6 and fc7. 2) Min: It is minimum value of fc6 and fc7. 3) Avg: It is the average for (fc6 and fc7). 4) Fusion between fc6 and fc7: It is used to combine the first group of fc6: (4096) next to the second group of fc7: (4096), and thus that will create dataset, which contains 8192 features.

Step 3: The aforementioned classifiers will be applied on the datasets that obtained from step 1 and step 2 to provide the results represented by Accuracy, Recall, F-measure, Precision, and duration time.

IV. Experimental Results and Discussions

The study is designed for three scenarios. In each scenario, the results of the evaluation of performance for the breast cancer images is represented by: Accuracy (Acc), Recall, F-measure, Precision (Pre) and duration training time for each classifier. The evaluation of performance is applied on the following four classifiers; KNN, NB, RF, and DT. The results of each scenario are illustrated in the following subsections.

A. First Scenario

This scenario was designed based on the percentage of 50% for training and 50% for testing. Its aim is to investigate the influence of 50% of the data size in the training dataset on the classification accuracy. In this scenario, three results are obtained. First, results for original fc6 and original fc7 datasets. Second, results for the statistical operations. Finally, results for fusion feature between (original fc6 and original fc7) dataset.

1) Results for original fc6 and original fc7 datasets separately: Table I and Table II show the results of the Fully Connected: fc6 feature vector dataset and fc7 feature vector dataset which were obtained from CNN outputs based on using different classifiers.

The results showed that Random Forest outperforms other algorithms in classifying breast cancer if it is malignant or benign for fc6 and fc7 which are (88.24) and (88.35) of classification accuracy respectively. However, the training time required to conduct the experiment shows that DT required more time (i.e. 16.03) compared to others, but KNN required little time (i.e. 0s). The reason for that, because there is no training model; the comparison occurs directly between the test row with other training rows (examples), and this explains the slow in time for testing, especially if there is large size of data (examples) for the training [52-53]. This results match with the results in [50] in term of that RF outperforms other classifiers used in their study. Their study compared Random Forest with Support Vector Machine, DT, Multilayer Perceptron, and KNN.

To the best of my knowledge, there was no the same study achieved to detect breast cancer based on the same proposed model in using deep learning with these four classifiers together (i.e. KNN, NB, RF, and DT), and also using statistical operations (i.e. Avg, Max, Min, and Fusion of fc6 and fc7), or using the three scenarios.

2) Results for the statitsical operations: The results of three datasets that created for statistical operations (i.e. Avg, Max, and Min) are presented in this section.

Tables III to V show results of the three statistical operations, whereas Random Forest algorithm outperforms other algorithms in classifying breast cancer if it is malignant or benign for Avg, Max, and Min, which are (88.28), (87.87), and (88.07) respectively. Despite of the Random Forest have showed an acceptable classification accuracy that outperformed other classifiers, it showed also an acceptable training time. While the training time required to conduct the experiment, the classifier Decision Tree required more time (i.e. 14.38s) compared to others, but KNN required little time (i.e. 0s), this is for the same reason mentioned in Section A of First Scenario. This results match with the results in [50] as mentioned in Section A of First scenario in the field of conducting study on RF, but not in using the statistical operation or three scenarios.

3) Results for fusion feature between (orginal fc6 and orginal fc7) dataset: This dataset is created by fusion of fc6 (4096 feature) and fc7 (4096 feature). The total feature will be 8192. The results in Table VI showed that Random Forest algorithm (88.38) outperformed other algorithms in classifying breast cancer if it is benign or malignant. The second-high accuracy is for KNN (86.78). While the training time for the classifiers, DT required large time (28.79s), but KNNs required less time (0s) when compared to other classifiers.

The summary of first scenario is that the results for all datasets used in first scenario are slightly similar to each other; especially for RF and KNN. This means all the features used in the study can have the same influence on the classification accuracy.

B. Second Scenario

This scenario was designed based on the percentage of 70% for training and 30% for testing. Its aim is to investigate the influence of 70% data size in the training dataset on the classification accuracy. In this scenario, three results are obtained. First, results for original fc6 and original fc7 datasets. Second, results for the statistical operations. Finally, results for fusion feature between (original fc6 and original fc7) dataset.

1) Results for original fc6 and original fc7 datasets separately: Table I and Table II show the results of the Fully Connected: fc6 feature vector dataset and fc7 feature vector dataset which were obtained from CNN outputs based on using different classifiers.

The results showed that Random Forest outperformed other algorithms in classifying breast cancer if it is benign or malignant for fc6 and fc7, which are (88.24) and (88.63) of classification accuracy respectively. While the training time for the classifier Decision Tree required more time (i.e. 31.09s) for fc7 compared to others. But KNN required little time (i.e. 0s), this is for the same reason mentioned in Section A of First Scenario. This results match with the results in [50] as mentioned in Section A of First scenario in the field of

conducting study on RF, but not in using the statistical operation or three scenarios.

2) Results for the statitsical operations: The results of three datasets that created for statistical operations (i.e. Avg, Max, and Min) are presented in this section.

Tables from III to V show results of the three datasets (avg, max, and min). The classifier Random Forest outperformed other algorithms in classifying breast cancer for Avg, Max, and Min, which are (88.51), (87.95), and (88.36) respectively. Despite of the RF have showed an acceptable classification accuracy that outperformed all other classifiers, and the training time was also an acceptable compared with others. While the training time for the classifier Decision Tree required more time (i.e. 81.13s) compared to others. But KNN required little time (i.e. 0s), as this explained earlier in Section A of First Scenario. This results match with the results in [50] as mentioned in Section A of First scenario in the field of conducting study on RF, but not in using the statistical operation or three scenarios.

TABLE I. DETECTION RESULTS OF FC6 FEATURE VECTOR

Algorithms	Acc	Pre	Recall	F-measure	Time (s)	Acc	Pre	Recall	F-measure	Time (s)	Acc	Pre	Recall	F-measure	Time (s)
	Fc6 (First Scenario)					*Fc6 (Second Scenario)*					*Fc6 (Third Scenario)*				
KNN	86.03	0.86	0.86	0.86	0.01	86.39	0.86	0.86	0.86	0	85.71	0.86	0.85	0.85	0.01
NB	78.54	0.82	0.78	0.79	1.29	77.98	0.82	0.78	0.78	1.76	78.71	0.82	0.78	0.79	3.27
RF	88.24	0.88	0.88	0.87	3.11	88.24	0.88	0.88	0.87	4.38	88.10	0.88	0.88	0.87	9.38
DT	81.19	0.81	0.81	0.812	13.71	81.36	0.81	0.81	0.81	20.18	82.07	0.82	0.82	0.82	57.63

TABLE II. DETECTION RESULTS OF FC7 FEATURE VECTOR

Algorithms	Acc	Pre	Recall	F-measure	Time (s)	Acc	Pre	Recall	F-measure	Time (s)	Acc	Pre	Recall	F-measure	Time (s)
	Fc7 (First Scenario)					*Fc7 (Second Scenario)*					*Fc7 (Third Scenario)*				
KNN	86.75	0.87	0.86	0.86	0	85.86	0.86	0.85	0.86	0	86.54	0.86	0.86	0.86	0.01
NB	82.93	0.84	0.82	0.83	1.19	83.52	0.84	0.83	0.83	1.98	83.39	0.84	0.83	0.83	2.2
RF	88.35	0.88	0.88	0.87	2.97	88.63	0.88	0.88	0.88	4.52	88.35	0.88	0.88	0.87	5.58
DT	81.77	0.81	0.81	0.81	16.03	82.65	0.82	0.82	0.82	31.09	81.60	0.81	0.81	0.81	31.52

TABLE III. DETECTION RESULTS OF AVERAGE (AVG) FEATURE VECTOR

Algorithms	Acc	Pre	Recall	F-measure	Time (s)	Acc	Pre	Recall	F-measure	Time (s)	Acc	Pre	Recall	F-measure	Time (s)
	Avg (First Scenario)					*Avg (Second Scenario)*					*Avg (Third Scenario)*				
KNN	86.06	0.86	0.86	0.86	0	86.64	0.87	0.86	0.86	0	85.86	0.86	0.85	0.86	0
NB	79.63	0.83	0.79	0.80	1.14	79.22	0.82	0.79	0.80	1.98	80.54	0.83	0.80	0.81	3.17
RF	88.28	0.88	0.88	0.87	3.01	88.51	0.88	0.88	0.88	8.77	88.35	0.88	0.88	0.87	10.4
DT	81.02	0.81	0.81	0.811	15	82.99	0.82	0.83	0.82	81.13	82.65	0.82	0.82	0.82	43.23

TABLE IV. DETECTION RESULTS OF MAXIMIUM (MAX) FEATURE VECTOR

Algorithms	Acc	Pre	Recall	F-measure	Time (s)	Acc	Pre	Recall	F-measure	Time (s)	Acc	Pre	Recall	F-measure	Time (s)
	Max (First Scenario)					*Max (Second Scenario)*					*Max (Third Scenario)*				
KNN	85.35	0.85	0.85	0.85	0	85.69	0.86	0.85	0.85	0	85.48	0.85	0.85	0.85	0
NB	82.45	0.84	0.82	0.83	1.24	82.74	0.84	0.82	0.83	2.8	83.07	0.84	0.83	0.83	2.99
RF	87.87	0.87	0.87	0.87	3.1	87.95	0.87	0.88	0.87	8.14	88.20	0.88	0.88	0.87	8.53
DT	81.36	0.81	0.81	0.81	14.38	81.58	0.81	0.81	0.81	31.7	82.12	0.82	0.82	0.82	35.71

TABLE V. DETECTION RESULTS OF MINIMUM (MIN) FEATURE VECTOR

Algorithms	Acc	Pre	Recall	F-measure	Time (s)	Acc	Pre	Recall	F-measure	Time (s)	Acc	Pre	Recall	F-measure	Time (s)
	Min (First Scenario)					*Min (Second Scenario)*					*Min (Third Scenario)*				
KNN	86.23	0.86	0.86	0.86	0	86.32	0.86	0.86	0.86	0	86.67	0.86	0.86	0.86	0.02
NB	76.19	0.81	0.76	0.77	1.23	75.45	0.81	0.75	0.76	2.82	76.64	0.81	0.76	0.77	5.02
RF	88.07	0.88	0.88	0.87	3.13	88.36	0.88	0.88	0.87	7.76	87.88	0.87	0.87	0.87	8.61
DT	83.10	0.83	0.83	0.83	14.45	82.31	0.82	0.82	0.82	38.24	81.03	0.81	0.81	0.81	51.01

TABLE VI. DETECTION RESULTS OF FUSION FEATURE VECTOR

Algorithms	Acc	Pre	Recall	F-measure	Time (s)	Acc	Pre	Recall	F-measure	Time (s)	Acc	Pre	Recall	F-measure	Time (s)
	Fusion (First Scenario)					*Fusion (Second Scenario)*					*Fusion (Third Scenario)*				
KNN	86.78	0.87	0.86	0.86	0	86.00	0.86	0.86	0.86	0	86.63	0.86	0.86	0.86	0
NB	81.77	0.83	0.81	0.82	2.42	82.18	0.84	0.82	0.82	3.84	82.75	0.84	0.82	0.83	6.51
RF	88.38	0.88	0.88	0.87	4.31	88.61	0.88	0.88	0.88	10.7	88.250	0.88	0.88	0.87	12.37
DT	81.23	0.81	0.81	0.81	28.79	83.09	0.82	0.83	0.83	73.95	82.48	0.82	0.82	0.82	91.39

TABLE VII. COMPARISON BETWEEN PROPOSED MODEL VS. MOST RELATED WORKS

Authors' References	Description of method/ technique	Dataset Size	Performance metric/s (value/s)
[54]	It consists of two stages: 2D detection and 3D aggregation. In 2D, the magnitude and orientation field map generated using Gabor filters. Then, features extracted using Faster-RCNN to determine the spatial pattern of AD. In 3D, the fusion strategy for regions is used for 2D into 3D. They used Soft classifier.	265 DBT image. They were collected from 68 patients	Mean True positive fraction (MTPF) of 50 +_ 0.04 ACC=90%
[55]	They used VGG-16 with progressive fine-tuning to evaluate its performance on AD detection. Then, the results were compared with a custom CNN architecture that were trained from scratch.	280 image	AUC=0.89
[56]	Breast masses detected system was developed based on texture description, spectral clustering, and support victor machine (SVM). ROIs were segmented using spectral clustering relaying on texture. Then, the optimal features were submited to SVM. They used SVM classifier.	-	Accuracy 90%
[51]	Different of CNN models are used for feature extraction: VGG16, InceptionV3, and ResNet50. Data were enlarged up to 400X to increase the accuracy. They used Softmax classifier.	7,909 images (Number of patients 82: 24 benign patients and 58 malignant patients).	98.26%.
The proposed model	The following are used in this study: - CNN (VGG-19 technique): features extracted from two layers: (fc6+fc7). - Fusion for two layers: (fc6+fc7) - Apply classifiers on: Fc6, fc7, statistical operations (avg, max, min, and fusion). - Three scenarios in the training dataset. - Used the following classifiers: K-Nearest Neighbors, Random Forest, Naïve Bayes, and Decision Tree.	5872	Accuracy (88.63%)

3) Results for fusion feature between (original fc6 and original fc7) dataset: This dataset is created by fusion of fc6 (4096 feature) and fc7 (4096 feature). The total feature will be 8192. The results in Table VI showed that there was not big difference between them, but Random Forest algorithm achieved higher accuracy compared to other algorithms in classifying breast cancer if it is benign or malignant for (88.61) of accuracy. The second-high accuracy is for KNN (86). While for the training time for classifiers; Decision Tree required more time (73.95s), but KNNs required little time (0s) compared to other.

The summary of second scenario is that results for all datasets used in the second scenario are slightly similar to each other; especially for RF and KNN. This means that all the features used in the study can have the same influence on the classification accuracy

C. Third Scenario

This scenario was built based on the percentage of 80% for training and 20% for test data set. Its aim is to investigate the influence of 80% data size in the training dataset on the classification accuracy. In this scenario, three results are obtained. First, results for original fc6 and original fc7 datasets.

Second, results for the statistical operations. Finally, results for fusion feature between (original fc6 and original fc7) dataset.

1) Results for orginal fc6 and orginal fc7 datasets separatly: Table I and Table II show the results of the Fully Connected: fc6 feature vector dataset and fc7 feature vector dataset which were obtained from CNN outputs based on using different classifiers.

The results showed that Random Forest outperforms other algorithms in classifying breast cancer if it is benign or malignant for fc6 and fc7, which are (88.10) and (88.35) of classification accuracy. While the training time for the classifier Decision Tree required more time (i.e. 57.63s) for fc7 compared to others, but KNN required little time (i.e. 0s. This results match with the results in [50], as discussed in Section A of First and Second Scenarios.

2) Results for the statitsical operations: The results of three datasets that created for statistical operations (i.e. Avg, Max, and Min) are presented in this section.

The results of statistical operations are presented in Tables from III to V. The results show the Random Forest algorithm outperformed other algorithms in classifying breast cancer for Avg, Max, and Min, which are (88.35), (88.20), and (87.88) respectively. In addition to these results of having acceptable accuracy for the Random Forest, the confusion matrix; F-measure, recall, and precision are scored high values among all other classifiers.

The training time for RF was also in an acceptable compared with others. While the training time for the classifier Decision Tree required more time (i.e. 51.01s) compared to others, but KNN required little time (i.e. 0.02s. This results match with the results in [50] in the field of classification accuracy results for RF, as they did not conduct their study in using the statistical operation or three scenarios.

3) Results for fusion feature between (original fc6 and original fc7) dataset: This dataset is created by fusion of fc6 (4096 feature) and fc7 (4096 feature). The total feature will be 8192. The results in Table VI showed that there were not big difference between the values of accuracy. Therefore, Random Forest algorithm outperformed other algorithms in classifying breast cancer if it is malignant or benign in fusion dataset that achieved (88.25) of accuracy. The second-high accuracy is for KNN (86.63). While the training time for the classifiers Decision Tree required large time (91.39s), and KNNs required less time (0s) compared to other classifiers.

The summary results for the three scenarios show that the classifier Forest showed better classification accuracy compared to other classifiers. In general, the required time to achieve was high for the case; fusion dataset in the third scenario, this is because the size of data is huge in the training, which was 80% and required a lot of time. In term of investigating the influence of the three scenarios on the classification accuracy, the results have approved that the detection for breast cancer can be achieved with almost similar classification accuracy even if the training dataset was minimal.

Table VII shows the comparison between our proposed model with most similar studies conducted for breast cancer detection. The proposed model can be considered as one of the interesting study, for number of reasons, mentioned below.

Some others of previous studies were performed on small dataset compared to proposal model.

Some others of previous studies were performed on large data size in training dataset (examples) compared to proposal model which contained a few images (examples). Three scenarios were considered in the proposal model. It was approved by our proposal model that using few number of images in the training dataset usually leads to low classification accuracy compared to use large number of images.

Some of previous data enlarge the data up to 400 times categories to increase the data size such as in [51]. This may influence on the training dataset that then would effect on the accuracy. In our proposed model, there were no enlarge in the training datasets.

To the best of my knowledge, we have not come across to any research performed based on statistical operations nor using three scenarios to detect breast cancer for the classification accuracy purposes, which considered a new method in this field.

V. Conclusion

The aim of this study is to investigate the use of deep learning model for breast cancer detecting using VGG-19 that used for extracting features from ultrasound images. Two layers of VGG19 structure were used: (i.e. fc6 and fc7); each of layer contains 4096 features. Also, more feature vectors were created from (fc6 and fc7), which called statistical operations. Different statistical operations are used to generate more datasets such: average, minimum, maximum and fusion of both fc6 and fc7. These datasets will be employed as input for the following ML classifiers: KNN, Random Forest, Naïve Bayes and Decision Tree. Data augmentation was considered to increase the dataset size for better learning of CNN.

Based on the results; Random Forest achieved high accuracy (88.63), precision (0.88), recall (0.88) and F-Measure (0.88) for fc7 of second scenario. The results were slightly similar; this approves that these features can provide a better accuracy when used in detection studies. Also, the results of the classification accuracy in the three scenarios are slightly similar, this approves that the breast cancer can be detected even if the size of data in the training dataset was minimal.

In the future work, it is recommending to conduct more investigation to improve the classification accuracy results and reduce training time using different algorithms.

References

[1] WebMD, (2022). “Cancer Health Center (CHC): Understanding Cancer—Diagnosis and Treatment”. WebMD. Medically Reviewed by Carmelita Swiner, MD on January 31, 2022, Last access: Juley. 13, 2023. [Online]. Available: https://www.webmd.com/cancer/understanding-cancer-treatment.

[2] Cancer.Net Editorial Board, (2023). "Breast Cancer: Statistics". American Society of Clinical Oncology (ASCO). Accessed: https://www.cancer.net/cancer-types/breast-cancer/statistics.

[3] J. Ferlay, H. R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin, (2008). "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008" Int. J. Cancer, vol. 127, no. 12, pp. 2893–2917, Dec. 2010, doi: 10.1002/ijc.25516.

[4] D. Adeloye, O. Y. Sowunmi, W. Jacobs, R. A. David, A. A. Adeosun, A. O. Amuta, S. Misra, M. Gadanya, A. Auta, M. O. Harhay, and K. Y. Chan, (2018). "Estimating the incidence of breast cancer in africa: A systematic review and meta-analysis" J. Global Health, vol. 8, no. 1, Jun. 2018, Art. no. 010419, doi: 10.7189/jogh.08.010419.

[5] R. Ben-Ari, A. Akselrod-Ballin, L. Karlinsky, and S. Hashoul, (2017). "Domain specific convolutional neural nets for detection of architectural distortion in mammograms". 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). pp. 552–652. DOI:10.1109/ISBI.2017.7950581.

[6] C. Nickson, K.E. Mason, D.R. English, A.M. Kavanagh, (2012). "Mammographic screening and breast cancer mortality: a case-control study and meta-analysis". Cancer Epidemiol. Biomark. Prev. 21 (9) (Sep. 2012) 1479–1488, https://doi.org/ 10.1158/1055-9965.epi-12-0468.

[7] M. Broeders, et al., (2012). "The impact of mammographic screening on breast cancer mortality in europe: a review of observational studies" J. Med. Screen 19 (1) (Sep. 2012) 14–25, https://doi.org/10.1258/jms.2012.012078.

[8] W.A. Berg, R.E.H. PhD, D.B. Kopans, P. Robert, A. Smith, (2020). "Frequently asked questions about mammography and the USPSTF recommendations: a guide for practitioners". Soc. Breast Imag. (2020) 1–17.

[9] C.K. Kuhl.et. al., (2015). "The Changing World of Breast Cancer: A Radiologist's Perspective". Invest. Radiol. 50 (9) (Sep. 2015) 615–628, DOI: https://doi.org/10.1097/rli.0000000000000166.

[10] S. Destounis, A. Santacroce, (2018). "Age to begin and intervals for breast cancer screening: balancing benefits and harms". AJR Am J Roentgenol. 210 (2) (2018) 279–284. DOI: https://doi.org/10.2214/AJR.17.18730.

[11] Robert K. Horsley MD a, Juliana M. Kling MD, MPH b, Suneela Vegunta MD b, Roxanne Lorans MD c, H'hamed Tem kit PhD d, Bhavika K. Patel MD, (2019). "Baseline Mammography: What Is It and Why Is It Important? A Cross-Sectional Survey of Women Undergoing Screening Mammography". Volume 16, Issue 2, February 2019, Pages 164-169.

[12] A.A. Roelofs, et al., (2007). "Importance of comparison of current and prior mammograms in breast cancer screening" Radiology 242 (1) (2007) 70–77. DOI: https://doi.org/ 10.1148/radiol.2421050684.

[13] A. Rimmer, 2017. "Radiologist shortage leaves patient care at risk, warns royal college" BMJ Br. Med. J. (Clin. Res. Ed.) 359 (2017). DOI: https://doi.org/10.1136/bmj.j4683.

[14] J.J. Fenton, et al., (2007). "Influence of computer-aided detection on performance of screening mammography" N. Engl. J. Med. 356 (14) (Apr. 2007) 1399–1409. DOI: https://doi.org/10.1056/nejmoa066099.

[15] C.D. Lehman, R.D. Wellman, D.S.M. Buist, K. Kerlikowske, A.N.A. Tosteson, D. L. Miglioretti, (2015). "Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Internal Med. 175 (11) (Nov. 2015) 1828. DOI: https://doi.org/10.1001/jamainternmed.2015.5231.

[16] A.D. Trister, D.S.M. Buist, C.I. Lee, (2017). "Will machine learning tip the balance in breast cancer screening?". JAMA Oncol. 3 (11) (Nov. 2017) 1463. DOI: https://doi.org/ 10.1001/jamaoncol.2017.0473.

[17] American College of Radiology, Data science Institute, (2020). " Empowering machine learning in radiology". FDA Clear. AI algorith. (2020), 1–1.

[18] M. P. Sampat, G. J. Whitman, M. K. Markey, and A. C. Bovik, (2005). ''Evidence based detection of spiculated masses and architectural distortions'. Proc. Med. Imag. Image Process., San Diego, CA, USA, Apr. 2005, pp. 26–37.

[19] Computer Creations, (2023). "Why is the VGG Network Commonly Used?". NNART, LLC 2023. Accessed: https://nnart.org/why-is-vgg-commonly-used/.

[20] Masud, M.; Rashed, A.E.E.; Hossain, M.S., (2022). "Convolutional neural network-based models for diagnosis of breast cancer". Neural Computing and Applications. volume 34, pages11383–11394 (2022). [Google Scholar] [CrossRef].

[21] Jiménez-Gaona, Y.; Rodríguez-Álvarez, M.J.; Lakshminarayanan, V., (2020). "Deep-Learning-Based Computer-Aided Systems for Breast Cancer Imaging: A Critical Review". Appl. Sci. 2020, 10, 8298. [Google Scholar] [CrossRef].

[22] Zeebaree, D.Q., (2020). "A Review on Region of Interest Segmentation Based on Clustering Techniques for Breast Cancer Ultrasound Images". J. Appl. Sci. Technol. Trends 2020, 1, 78–91. [Google Scholar].

[23] Huang, K.; Zhang, Y.; Cheng, H.; Xing, P., (2021). "Shape-adaptive convolutional operator for breast ultrasound image segmentation". In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; IEEE: New York, NY, USA, 2021; pp. 1–6. [Google Scholar].

[24] P. Kus and I. Karagoz, (2010). "Accurate segmentation of the breast region with texture filter in mammograms for CAD applications", in Proc. 15th Nat. Biomed. Eng. Meeting, Apr. 2010, pp. 1–8.

[25] J. M. Wolterink, P. de Jong, T. Leiner, W. Viergever, and I. Iágum, (2017). ''Convnet based localization of anatomical structures in 3-D medical images'', IEEE Trans. Med. Imag., vol. 36, no. 7, pp. 1470–1481, Jul. 2017.

[26] M. A. Al-masni, M. A. Al-antari, J. M. Park, G. Gi, T. Y. Kim, P. Rivera, E. Valarezo, S.-M. Han, and T.-S. Kim, (2017). ''Detection and classification of the breast abnormalities in digital mammograms via regional convolutional neural network'', in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Seogwipo, Japan, Jul. 2017, pp. 1230–1233, doi: 10.1109/EMBC.2017.8037053.

[27] P. Kaur, G. Singh, and P. Kaur, (2019). ''Intellectual detection and validation of automated mammogram breast cancer images by multi-class SVM using deep learning classification''. Informat. Med. Unlocked, vol. 16, Oct. 2019, Art. no. 100151, doi: 10.1016/j.imu.2019.01.001.

[28] D. Singh and M. Singh, (2016). "Investigation on ROI selection for mammography using texture models and machine learning classifier". Int. J. Control Theory Appl., vol. 9, p. 45, Dec. 2016.

[29] M. Izadpanahkakhk, S. Razavi, M. Taghipour-Gorjikolaie, S. Zahiri, and A. Uncini, 2018. ''Deep region of interest and feature extraction models for palmprint verification using convolutional neural networks transfer learning''. Appl. Sci., vol. 8, no. 7, p. 1210, Jul. 2018.

[30] Y. Hou, H. Zhang, S. Zhou, and H. Zou, (2017). "Efficient ConvNet feature extraction with multiple RoI pooling for landmark-based visual localization of autonomous vehicles". Mobile Inf. Syst., vol. 2017, pp. 1–14, Apr. 2017, doi: 10.1155/2017/8104386.

[31] R. Lakshmanan, S. T. P., V. Thomas, S. M. Jacob, and T. Pratab, (2014). "A preprocessing method for reducing search area for architectural distortion in mammographic images", in Proc. 4th Int. Conf. Adv. Comput. Commun., Aug. 2014, pp. 101–104.

[32] Mishra, A.K.; Roy, P.; Bandyopadhyay, S.; Das, S.K., (2021). "Breast ultrasound tumour classification: A Machine Learning—Radiomics based approach". Expert Syst. 2021, 38, e12713. [Google Scholar] [CrossRef].

[33] Byra, M., 2021. "Breast mass classification with transfer learning based on scaling of deep representations". Biomed. Signal Process. Control 2021, 69, 102828. [Google Scholar] [CrossRef].

[34] Irfan, R.; Almazroi, A.A.; Rauf, H.T.; Damaševičius, R.; Nasr, E.; Abdelgawad, A., (2021). "Dilated Semantic Segmentation for Breast Ultrasonic Lesion Detection Using Parallel Feature Fusion". Diagnostics 2021, 11, 1212. [Google Scholar] [CrossRef].

[35] Ahmad, N., Asghar, S., & Gillani, S.A., (2022). "Transfer learning-assisted multi-resolution breast cancer histopathological images classification". Vis. Comput., 38, 2751-2770.

[36] Kousalya, K., & Saranya, T., (2021). "Improved the detection and classification of breast cancer using hyper parameter tuning". Volume

81, Part 2, 2023, Pages 547-552. DOI: https://doi.org/10.1016/j.matpr.2021.03.707.

[37] Duggento, A., Conti, A., Mauriello, A., Guerrisi, M., & Toschi, N., (2020). "Deep computational pathology in breast cancer". Seminars in cancer biology, 72, 226– 237. doi: 10.1016/j.semcancer.2020.08.006.

[38] Duc My Vo, Ngoc-Quang Nguyen, Sang-Woong Lee, (2019). "Classification of breast cancer histology images using incremental boosting convolution networks". Information Sciences. Volume 482, May 2019, Pages 123-138. https://doi.org/10.1016/j.ins.2018.12.089.

[39] Kumar, A., Singh, S.K., Saxena, S., Lakshmanan, K., Sangaiah, A.K., Chauhan, H., Shrivastava, S., & Singh, R.K., (2020). "Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer" Information Sciences, Volume 508, pages 405-421.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, (2012). "Imagenet classification with deep convolutional neural networks". In Advances in neural information processing systems, 1097–1105 (2012).

[41] Z. Weiming, W. Henry, Y. Fung, C. Zhenghao, Z. Seid Miad, L. Zhicheng, and C. Yuk Ying, (2017). "Using Transfer Learning with Convolutional Neural Networks to Diagnose Breast Cancer from Histopathological Images". ICONIP 2017, Part IV, LNCS 10637, pp. 669–676, 2017. https://doi.org/10.1007/978-3-319-70093-9_71.

[42] F.A. Spanhol, L.S. Oliveira, C. Petitjean, C. and L. Heutte, (2016). "Breast cancer histopathological image classification using convolutional neural networks". In International Joint Conference on Neural Networks (2016), p. 2560-2567.

[43] F.A. Spanhol, A. Fabio., L.S Oliveura., C. Ptitjean., 2016. "A dataset for breast cancer histopathological image classification". IEEE Transactions on Biomedical Engineering, (2016), vol. 63, no 7, p. 1455-1462.

[44] Z. Han, B. WEI, Y. ZHENG, "Breast cancer multi-classification from histopathological images with structured deep learning model". Scientific reports, 2017, vol. 7, no 1, p. 4172.

[45] N. Hatipoglu and B. Gokhan, (2014). "Classification of histopathological images using convolutional neural network". Image Processing Theory, Tools and Applications (IPTA), 2014 4th International Conference on. IEEE, 2014. p. 1-6.

[46] R. Cruz , A. Basavanhally, A. Gonzalez et al., (2014). "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks". Medical Imaging 2014: Digital Pathology. International Society for Optics and Photonics, 2014. p. 904103.

[47] L. G. Hafemann, L. S. Oliveira, and P. Cavalin, (2014). "Forest species recognition using deep convolutional neural networks". International Conference on Pattern Recognition, 2014, pp. 1103–1107.

[48] T. H. Vu, H. S. Mousavi, V. Monga, U. A. Rao, and G. Rao, (2015). "Dfdl: Discriminative feature-oriented dictionary learning for histopathological image classification," in Proceedings of the IEEE 12th International Symposium on Biomedical Imaging (ISBI). IEEE, Apr. 2015, pp. 990– 994.

[49] Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A., (2020). "Dataset of breast ultrasound images. Data in Brief". 2020 Feb;28:104863. DOI: 10.1016/j.dib.2019.104863.

[50] Manas Minnoor and Veeky Baths,. (2023). "Diagnosis of Breast Cancer Using Random Forests". Procedia Computer Science. International Conference on Machine Learning and Data Engineering. Volume 218, pages 429-437. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2023.01.025. Accessed: https://www.sciencedirect.com/science/article/pii/S187705092300025X.

[51] Xiaofan Cheng & Liang Tan & Fangpeng Ming, 2021. "Feature Fusion Based on Convolutional Neural Network for Breast Cancer Auxiliary Diagnosis," Mathematical Problems in Engineering, Hindawi, vol. 2021, pages 1-10.

[52] Hassanat, A. (2018). "Furthest-pair-based binary search tree for speeding big data classification using k-nearest neighbors". Big Data, 6(3): 225-235.

[53] S. Al-Showarah et. al. 2020. "The Effect of Age and Screen Sizes on the Usability of Smartphones Based on Handwriting of English Words on the Touchscreen", Mu'tah Lil-Buhuth wad-Dirasat, Natural and Applied Sciences series, Vol. 35, No. 1, 2020. ISSN: 1022-6812.

[54] Y. Li, Z. Xie, Z. He, X. Ma, Y. Guo, W. Chen, and Y. Lu, ''Architectural distortion detection approach guided by mammary gland spatial pattern in digital breast tomosynthesis,'' Proc. SPIE, vol. 11314, Mar. 2020, Art. no. 1131417.

[55] A. C. Costa, H. C. R. Oliveira, L. R. Borges, and M. A. C. Vieira, ''Transfer learning in deep convolutional neural networks for detection of architectural distortion in digital mammography,'' Proc. SPIE, vol. 11513, May 2020, Art. no. 115130N.

[56] Ketabi, H., Ekhlasi, A. & Ahmadi, H. A computer-aided approach for automatic detection of breast masses in digital mammogram via spectral clustering and support vector machine. Phys Eng Sci Med 44, 277–290 (2021). https://doi.org/10.1007/s13246-021-00977-5.

Detecting Threats from Live Videos using Deep Learning Algorithms

Rawan Aamir Mushabab AlShehri[1], Abdul Khader Jilani Saudagar[2]
Information Systems Department, College of Computer and Information Sciences,
Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

***Abstract*—Threat detection is an important area of research, particularly in security and surveillance applications. The research is focused on developing a threat detection system using DL techniques. The system aims to detect potential threats in real-time video streams, enabling early identification and timely response to potential security risks. The study uses two state-of-the-art DL models, MobileNet and YOLOv5, to train the object detection system. The TensorFlow object detection API is employed for training and evaluating the models. The results of the study indicate that MobileNet outperforms YOLOv5 in terms of detection accuracy, speed, and overall performance. The justification for selecting MobileNet over YOLOv5 is based on several factors. First, MobileNet has a lightweight architecture, making it suitable for real-time applications where processing speed is critical. Second, it is efficient in terms of memory usage, enabling it to operate effectively on low-resource devices. Third, MobileNet provides high accuracy in detecting objects of different sizes and shapes. The study evaluated the performance of the threat detection system using various evaluation metrics, including mean average recall (mAR), mean average precision (mAP) and Intersection over union (IoU). The results show that the system achieved high accuracy in detecting threats, with an overall mAP (mean average precision) of 0.9125, mAR (mean average recall) of 0.9565 and Intersection over union (IoU) of 0.9045. In this study, researchers present a highly efficient and successful method for identifying threats through the utilization of deep learning methods. The research demonstrates the superiority of MobileNet over YOLOv5 in terms of performance, and the results obtained validate the effectiveness of the proposed system in detecting potential threats in real-time video streams.**

Keywords*— *Deep learning; machine learning; object detection; threat detection

I. INTRODUCTION

Nowadays, technologies are experiencing unprecedented growth and advancement, particularly in the field of Data Sciences (DS). DS encompasses a wide range of disciplines, including Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL). AI involves the external creation of intelligence for various systems, enabling them to make decisions based on insights derived from available data. ML empowers machines to process data and generate knowledge and intelligence by integrating applied statistics and optimization theory. DL is a subfield of ML that specifically concentrates on constructing, training, and deploying extensive and intricate neural networks [1]. By leveraging data in parallel, these neural networks carry out their operations with the aim of accomplishing their assigned tasks. To develop an intelligent system for object detection, it is necessary to identify and categorize moving objects. Object detection entails the ability of computers and software systems to identify the presence and location of various objects within an image, such as vehicles, animals, humans, and more. Deep convolutional neural networks have gained significant prominence in tasks like image classification, object classification, localization, and object detection [2]. DL technologies have been increasingly applied in image classification, target tracking, object detection, image segmentation. These technologies have helped in different social life events and cases. The focus in this research is to use these technologies to raise safety and security. Considering the escalating global crime and terrorism rates, the demand for automated video surveillance has surged. The combination of surveillance and detection has become critically significant. While human detection and tracking are desirable, the unpredictable nature of human movement presents significant challenges in effectively tracking and categorizing suspicious activities. The focus of this research is to identify and detect potentially threatening objects captured by Close Circuit Television Cameras (CCTV) [3]. This idea comes out of the great need for detecting threatening objects to help evaluate situations or prevent any further crimes. Finally, the idea behind this research is extracted and built based on the previous studies in the field of DL techniques. Also, among the search between the previous studies, there was no use of the MobileNet model in the field of recognizing and detecting objects in videos especially threatening objects.

A. Object Detecting Technique

Object detection is a computer vision technique that empowers software systems to identify, locate, and track objects within images or videos. One of its notable features is the ability to classify objects (such as people, tables, chairs, etc.) and precisely determine their coordinates within the image. This is achieved by drawing a bounding box around the object, although the accuracy of the bounding box may vary. The effectiveness of the detection algorithm is measured by its capacity to accurately locate objects within the image. An example of object detection is facing detection.

Object detection algorithms can either be pre-trained or trained from scratch, with pre-trained weights from existing models often utilized and fine-tuned to suit specific requirements or use cases. Object detection is a crucial area of research in computer vision and finds widespread applications in various fields, including surveillance, robotics, autonomous vehicles, and image and video search engines. The objective of object detection is to enable machines to comprehend and

interpret the visual world in a similar manner to humans. Deep learning techniques, such as convolutional neural networks (CNNs), are commonly employed in object detection algorithms to analyze images or videos and identify objects within them. These algorithms are trained on large datasets comprising labeled images, where the object's location and class are annotated. The training process involves adjusting the weights and biases of the neural network to minimize the error between predicted and actual object locations and classes. Once trained, object detection algorithms can be applied to detect objects in new images or videos. The algorithm typically outputs a set of bounding boxes corresponding to the detected objects, accompanied by their class labels and confidence scores. Post-processing techniques can be further employed to refine the bounding boxes and enhance the accuracy of the object detection process.

In recent years, there has been significant progress in object detection, driven by advances in DL and the availability of large datasets. State-of-the-art object detection algorithms can achieve high accuracy and real-time performance on a wide range of object types and scenarios. In addition to the traditional object detection methods, there are also several advanced techniques that have emerged in recent years. Below are a few examples:

1) One-shot object detection: Traditional object detection algorithms require a large amount of labeled data during the training phase. One-shot object detection, on the other hand, is a technique that can detect objects with only one or a few examples of each object class during training.

2) Instance segmentation: Instance segmentation is an extension of object detection that not only detects the objects in an image but also segments each object instance from the background. This technique is useful in scenarios where precise object boundariesare necessary, such as in medical imaging or autonomous driving.

3) 3D object detection: While traditional object detection works on 2D images, 3D object detection can detect objects in 3D space. This is important for applications such as robotics and autonomous vehicles, where the detection of objects in 3D is necessary for navigation and obstacle avoidance.

4) Few-shot object detection: Similar to one-shot object detection, few-shot object detection is a technique that can detect objects with only a few examples of each object class during training. However, few-shot object detection is more challenging than one-shot object detection as it requires the algorithm to generalize to unseen object classes.

B. Object Detecting from Image and Video

Object detection is an important and prominent area of research that combines deep learning (DL), computer vision, and image processing. Its primary objective is to identify specific semantic objects, such as people or animals, in digital images and videos. Within the field of object detection, there are well-established areas of study, such as pedestrian detection and face detection. However, object detection has broader applications in various DL fields, including image restoration and video surveillance [4]. To differentiate between different objects, object detection employs a multi-label classifier, although it does not determine the specific identity of each object. This is where Image Localization comes into play, as it precisely determines the object's location within the image by providing a bounding box around it. Image Localization technology has practical applications in image retrieval, with facial detection being a widely used example. Additionally, it can assist in pedestrian detection at traffic lights to enhance traffic flow and aid visually impaired individuals. It also facilitates Sign Language Detection (SLD) to support communication with the deaf and mute community. The process of object detection involves providing an image or video frame, using algorithm-based models to search for targeted objects, and assigning specific categories to each identified target [5]. Object detection algorithms commonly employ machine learning (ML) and DL techniques to achieve meaningful results. Furthermore, the ultimate goal of object detection models is to emulate the rapid comprehension and recognition of objects by the human brain when observing images or videos. Training DL algorithms and models to match the intelligence of the human brain using computer technology is a challenging yet crucial task, particularly in the context of detecting potentially threatening objects.

C. Problem Statement

The research focuses on the detection and recognition of objects using ML and DL techniques, which is a prominent area of study for DL enthusiasts. The aim is to enable machines to learn autonomously by simulating the functioning of neurons in the human brain. While previous works have explored object detection for various social issues, there is a lack of research specifically addressing the detection of threatening objects in videos and evaluating the MobileNet model for this purpose. This research aims to enhance community safety and security by improving the detection of threatening objects, as incidents and accidents often result from inadequate security measures or delayed responses to immediate threats. The research seeks to answer two main questions: is there an efficient way to detect threats from live videos using DL algorithms, and how accurate and efficient is the MobileNet model in detecting objects in live videos? The objectives of the study are to detect threats in live videos to prevent the criminal use of threatening objects and to assess the efficiency and accuracy of the MobileNet model in detecting threatening objects. The research utilizes DL techniques to identify and detect threats in live video clips or surveillance videos. The idea for this research originated from the urgent need for video analysis on the internet, and it builds upon a review of previous studies in DL and its techniques. The researchers identified a gap in utilizing the MobileNet model for identifying and detecting potential shapes and threats in videos. The MobileNet model will be employed in this research to detect threatening objects, and the results will be analyzed in terms of accuracy and quality.

II. LITERATURE REVIEW

The DL fields are full and rich by the research in this domain. The related work to this contribution was different in the model that they used, or they study the same model to different dataset, or they wanted to detect images instead of

video. There are very limited kinds of literature that have adopted detecting threats by using DL or ML algorithms. This section presents most of the notable research worked on DL, ML, and object detection.

A. Object Detection Algorithms

Object Detection is a very trending domain among researchers. It was thoroughly studied from the beginning of 2010 until the current time [6]. That shows how much this domain is important and rich. Fig. 1 shows the number of publications with the keywords of "Object Detection" during the past 10 years.

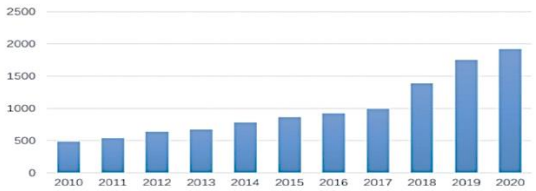


Fig. 1. The growing numbers of publications in object detection algorithms.

Object detection plays a crucial role in computer vision applications, enabling the automatic identification and localization of objects within images or videos. The researcher Shaukat Hayat. el at [7], studied a DL framework for object recognition purposes. The proposed model was further tuned, and the recognition performance was improved. They use nine different object classes taken from wide varied image dataset caltech101 and deploy five layers CNN model. They compared the proposed model's performance with different classical bag-of-words (BOW) approaches trained on a novel. This algorithm achieved an accuracy level of 90.12% is much better than different classical BOW approaches. While Kanimozhi S. el at [8], tried to detect and track the object in the sports field to make the computer learn Deeply, which is none other than the application of DL. The proposed method increases the accuracy level in identifying real-time household objects. Aniruddha Srinivas Joshi, el at [9], adopted the idea of detecting face masks from video footage. A highly effective face detection model is applied for obtaining facial images and cues. A distinct facial classifier is built using DL to determine the presence of a face mask in the facial images detected. The proposed method has shown its good effectiveness in identifying facial masks by achieving high precision, recall, and accuracy. As for using ML algorithms to detect threats included in the social media post, Shatha Alajlan. el at [10] has published research in that domain. She utilized the CNN model on the TensorFlow platform to classify Instagram content (images and Arabic comments) for threat detection. The results of this research showed that the accuracy of the developed model is 96% for image classification and 99% for comment classification.

1) Using object detection algorithms: Muhammad Shakeel. el at [11] have developed a passenger security screening system that employs DL techniques to detect potential security threats by rotating the image of a person's body. However, this method has limitations in identifying the specific type of threat and precisely locating its position within the body. Shaoqing Ren. el at [12] have conducted a study on object classification and detection performance, achieving a remarkable 5-7 frame rate per second and 73.2% mean average precision (mAP). Their approach involves categorizing objects and identifying their precise location, allowing for the development of an efficient security screening algorithm that accurately detects threats at specific locations using an enhanced faster R-CNN model. A comprehensive analysis of cargo X-ray image analysis automation was carried out by Jaccard, Nicolas et al. [13]. The review emphasized the importance of employing image pre-processing techniques, including image quality enhancement, manipulation, material discrimination, and segmentation. These techniques play a crucial role in improving the accuracy of automated image understanding algorithms and rectifying errors that may arise during image acquisition.

Moreover, Jaccard's paper proposes an automated threat detection method to further improve the accuracy in the analysis of cargo X-ray images. Nicolas Jaccard el at. [13] have developed a CNN model specifically designed for detecting threats in X-ray images. Their model was trained using an augmented dataset that included real threat images, resulting in a high detection rate of 90% and a low false alarm rate of only 0.8%. The effectiveness of image manipulation and quality improvement techniques in enhancing the proposed solution is highlighted in their study. In their study, Akcay, Samet et al. [14] investigated the potential of convolutional neural networks (CNNs) for object classification in X-ray baggage images. Their research focused on utilizing CNNs to improve the accuracy of object classification in this domain.

On the other hand, Riffo, Vladimir and Flores Sebastian [15], proposed an innovative automated approach for object detection in X-ray images, specifically for baggage screening purposes. Their solution involved the utilization of an adapted implicit shape model (ASIM), an enhanced version of the implicit shape model introduced in the research conducted by Leibe, Bastian et al. [16]. The ASIM approach employed SIFT descriptors to describe objects using multiple X-ray images from different perspectives. The visual vocabulary of object parts was then used to characterize the object, and targets were detected by searching for similar visual words and spatial distributions. Although the object detector incorporated pose estimation and Q-learning computer vision techniques, it may not be ideal for region-based threat detection.

Furthermore, Nurhopipah, Ade et al. [17] conducted a study that delved into various aspects of motion detection, face detection, data training, and face identification. Their research aimed to explore the complexities associated with these areas in the context of threat detection. Their utilization of the Accumulative Differences Images (ADI) approach for motion segmentation proved successful, with motion detection reaching a high success rate of 92.655%. The Haar cascade classifier was employed for face detection, with a success rate of 76%, and face identification reached 60%. Meanwhile, Busarin Eamthanakul. el at [18] implemented the background subtraction method and median filter to compute data and analyze traffic conditions. Their system was able to detect the

number of objects or cars on the road, providing useful data for traffic management.

Nipunjita Bordoloi. el at [19] developed a security system that effectively tracks object movement and detects anomaly motion in real-time. Background subtraction was utilized to track objects, and the system achieved success in detecting suspicious activity. Alavudeen Basha. el at [20] also delved into suspicious activity detection, using a CNN-DBNN algorithm to detect human activity. The technique of foundation subtraction was used for human detection, with larger bounding boxes to enclose individuals. A Discriminative Deep Belief Network (DDBN) was implemented for activity classification, with an impressive accuracy rate of 90%. The research in computer vision systems continues to advance, providing new and innovative ways to detect and analyze data.

2) Using mobilenet model: MobileNet models offer an efficient solution for on-device intelligence across various recognition tasks. Developed specifically for TensorFlow, MobileNets are a family of computer vision models designed with a mobile-first approach. These models prioritize accuracy while considering the limited resources available for on-device or embedded applications. MobileNets fall under the category of lightweight deep convolutional neural networks, significantly smaller in size and faster in performance compared to many other popular models. Their small footprint, low latency, and low power consumption make them well-suited to meet the resource constraints of diverse use cases. MobileNet models can be leveraged for classification, detection, embeddings, and segmentation tasks, providing a versatile framework for on-device intelligent applications.

3) The Gap on the Literatures: This research aims to complete what other researchers have done by using DL algorithms and the MobileNet model to detect objects in images and videos. All the studied and reviewed literature were specifying different domain of study or different type of purposes. However, this research focus on Using the MobileNet model to detect the threatening objects on videos. According to previous studies, various researchers have reported high AP values for this model's ability to detect different classes such as cars, persons, and chairs. Some research findings suggest that the AP reaches as high as 99.76%, while others claim to achieve a slightly lower value of 97.76% [21]. Researchers want to approve if the same percentage would be detected with the same purpose they aim to study. This action has not been previously investigated by researchers in the field of object detection by using DL algorithms.

III. Methodology

This research aims to investigate the effectiveness of the MobileNet model in detecting threatening objects. Object detections a crucial task in computer vision, and it has numerous applications in various domains, such as security, surveillance, and autonomous driving. The research is focused on detecting threatening objects in public places, and the results could have significant implications for improving public safety as its none of the research objectives. The research approach adopted is a qualitative exploratory approach, which is a suitable method for gaining an in-depth understanding of a phenomenon.

A. Research Design

A research design serves as a structured framework or strategy for collecting, measuring, and analyzing data with the purpose of addressing specific research inquiries [22]. Fig. 2 shows the research scenario that the researcher follows to answer the research question and fulfill its objectives.

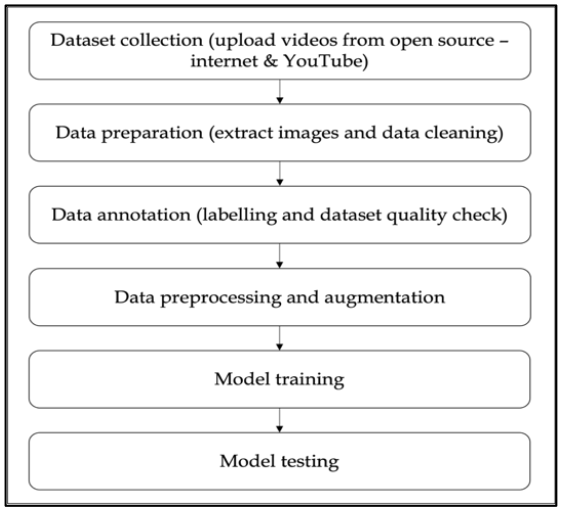


Fig. 2. Research scenario.

B. Data Collection

The dataset used in this study was curated for the purposes of the study. It was generated and collected by the researcher from various sources. It consists of five classes of objects that are considered potentially dangerous or threatening, including: fire, guns, knives, arrows, and swords. These objects were chosen due to their prevalence in public safety incidents and their potential to cause harm. To obtain the data, various sources were utilized, including YouTube videos that contained CCTV footage, educational videos for learning fighting skills, and demo videos. The videos were processed by extracting frames at a rate of 1 frame per second (1 fps) to capture the necessary images for the dataset. The use of CCTV footage is particularly useful as it allows for the collection of authentic data from public places where security cameras are commonly used. To annotate the dataset, the RoboFlow tool was utilized, which is a popular image annotation tool used for object detection tasks. Annotations provide additional information about the images, indicating the location and class of the object within the image. These annotations are essential for training ML models to accurately detect objects. To increase the diversity of the dataset, augmentations such as rotation and contrast difference were applied to the images. These augmentations help to create variations of the images, which can improve the performance of ML models by exposing them to a wider range of data. The purpose behind curing such dataset is essential for training ML models to accurately detect these objects within the five mentioned classes and improve public safety.

C. Data Preparation

Object detection from images is a fundamental task in DL that requires a significant amount of labeled data for training. The MobileNet SSD model is a popular DL model for object detection, but it requires a large, labeled dataset for effective training. In this study, a sufficient amount of data was obtained by collecting threatening videos and CCTV footage from online resources such as YouTube. The videos collected for this study mainly contained real-life activities, such as ATM robberies and police operations, to ensure that the dataset reflects real-world scenarios. Additionally, training and rehearsal-based videos were included to increase the number of images for each object in the dataset. This was an important step because having a larger dataset with a balanced distribution of classes can improve the performance of the model. To create the labeled dataset, frames were extracted from the videos at a rate of 30 frames per second (30 FPS) using the RoboFlow tool. Each frame was manually labeled for the presence of five different objects, including guns, knives, swords, arrows, and fire. Manual labeling is a crucial step in creating a high-quality labeled dataset because it ensures that the labels are accurate and consistent. The use of real-life video footage and manual labeling ensures that the dataset is of high quality and accurately reflects the threatening objects. This can improve the accuracy and reliability of the model when detecting threatening objects in real-world scenarios.

1) Data cleaning: After extracting all the images from videos uninformative and vague frames are discarded for data cleaning and maintaining data quality. Only frames with clear object visibility are remained after cleaning that are labelled manually for five object classes. Data classes are sampled in such a way that create a balanced number of images per object in training, validation, and testing.

D. Data Annotation

In this phase, the researcher used the RoboFlow annotation tool that enabled drawing bounding boxes around the objects of interest. RoboFlow enables the uploading of videos and extraction of images with varying FPS rates. Each image is labeled for its particular class name and bounding box. The RoboFlow tool allows saving the bounding box values in different formats, as required by the model. In this case, the labeling is saved as a CSV file to comply with the MobileNet SSD file format. The minimum and maximum values for the bounding box x and y sides are saved to draw the bounding box rectangle. The RoboFlow interface for image labeling is depicted in the Fig. 3. When labeling the images using the RoboFlow tool, a bounding box is drawn around the object of interest in the image. The bounding box is represented by a rectangular box with four values: the x-coordinate and y-coordinate of the top-left corner of the box, and the width and height of the box. To save the bounding box values in the CSV file, the RoboFlow tool records the minimum and maximum values for the x and y coordinates of the top-left corner of the box, as well as the width and height of the box. These values are saved in separate columns in the CSV file, along with the class name of the object in the image. For example, there is an image containing a gun, and the bounding box around the gun has a top-left corner coordinate of (100, 150), a width of 50 pixels, and a height of 100 pixels. The RoboFlow tool would save the following values in the CSV file for this object as follow:

- Class Name: Gun
- Minimum x-coordinate: 100
- Maximum x-coordinate: 150
- Minimum y-coordinate: 150
- Maximum y-coordinate: 250

To start the annotation step, the researcher opened each image using the chosen annotation tool. For every image, the researcher carefully drew bounding boxes around the instances of the objects that aimed to be detected. For each bounding box, the researcher assigned the correct class label from the five mentioned classes which are: fire, gun, knives, arrows, or swords. This step was done manually to ensure accurate placements and labels, as these annotations would serve as the ground truth for the models training phase. With the annotated dataset in hand, the next step was to integrate it with the respective training frameworks. The researcher utilized the annotations they had created to train the object detection models.

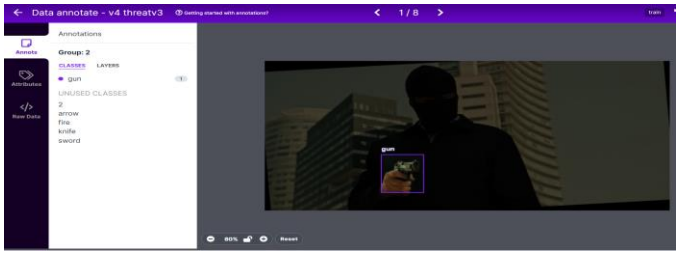


Fig. 3. RoboFlow interface.

1) Dataset quality check: Ensuring data quality is a critical step before building models for object detection. Poor data quality can lead to inaccurate and unreliable models. Below the used methods taken to ensure data quality in object detection:

- Annotation Quality Control
- Data Cleaning and Preprocessing
- Balanced Class Distribution
- Data Augmentation
- Validation and Anomaly Detection
- Consistent Image Quality
- Real-World Scenario Simulation
- Review Annotations for Ambiguity
- Class Label Consistency
- Cross-Validation
- Continuous Monitoring
- Use External Data Sparingly

- Expert Review
- Feedback Loop

Once reached the confident in the quality of the annotations, we proceeded to export the annotated dataset from RoboFlow. The platform typically provided options to export the data in formats that were compatible with various deep learning frameworks. Given this scenario, the researchers ensured that the exported format aligned with the chosen model architecture, whether it was YOLO or MobileNet SSD.

E. Data preparation and Augmentation

Data preprocessing is applied in order to make it suitable for effective model training. Images are resized to one scale 416x416 for MobileNet SSD model. Data preprocessing enhances data quality and decreases training time. Images are scaled, cropped, and resized to make them in same format before model training. Data augmentation is also applied to increase veracity of data so model can learn better with more data as addressed before. Images are rotated, blurred, and adjusted contrast and orientation for data augmentation purposes.

F. Data Limitations and Issues

DL models require a large amount of data to train effectively, and the more data feed into the model, the better its performance would be. In this study, the researchers have around 2000 images for each object after augmentations. However, the number of images per sample is relatively low, and the diversity of real scenarios is limited, which are some of the limitations of this dataset. To improve the performance of the model, more data can be collected from real-time scenarios. Furthermore, due to privacy concerns and issues, many establishments and markets are hesitant to share their camera recordings. Therefore, more collaboration and support are needed to collect a diverse range of datasets to improve the model's performance and deliverables.

G. Model Building

This research focus on two DL models. The YOLOv5 (You Only Look Once version five) model, which has undergone thorough scrutiny by researchers, stands out as the initial model that has gained recognition for its effectiveness in object detection. This model showcases highly promising accuracy outcomes based on two key metrics: mAP (mean average precision) and FPS (frames per second). In a study conducted by S. Murthy et al. [23], the application of YOLOv5 was investigated, and it demonstrated superior speed and a 95% accuracy rate compared to other object detection algorithms examined in the comparative analysis. Additionally, it achieved an average precision ranging between 67 and 70, along with a frames per second rate ranging between 65 and 124. A thorough comparison of the YOLO model versions in the below sections. In reference to the Debojit Biswas et al. work that has been done [24] MobileNet SSD model achieved 92.97% average detection accuracy in the experiment. Sanjay Kumar et al. confirm in his work that the SSD on MobileNet has the highest mAP among the models targeted for real-time processing [25]. That was promising to start the investigation upon this case.

1) YOLOv1: First object detection network that combines the problem of identifying class labels and determining bounding boxes for a set amount of classes, making it a one-stage detector (rather than two-stage detectors which first detects the regions of interest, and then classify that region as a specific class based on given input during training). This is possible by fully connecting the two important steps of bounding box prediction and classification of labels to an end-to-end differentiable network [26].

2) YOLOv2: From it's iteration of version one of YOLO, works have been done too dramatically improve the performance of the accuracy through the addition of BatchNorm, improved resolutions, and the use of anchor boxes [27].

3) YOLOv3: Improvements made from the previous model included the use of more connections in its backbone network layers as well as adding a new network that aids in the model's ability to identify smaller objects better (with the use of feature pyramid network (FPN) that allows the model to learn objects of different sizes simultaneously). Added an objectness score for the model's bounding box predictions, which helps determine the bounding box to take for all the bounding boxes overlapping a specific ground truth object within an image [26][27].

4) YOLOv4: Additional improvements were introduced into the YOLO series in YOLOv4 through the introduction of:

- Feature Aggregation which combines the features extracted from previous layers
- Bag of Freebies - several methodologies and functions added to improve its performance without affecting the model's inference during production. The main additions are related to data augmentation such as rotation, flip, crop, hue, saturation, mosaic, MixUp, Blur, etc.
- Self-Adversarial Training which allows the model to find the region of the image that its network relies most on and subsequently editing the image to remove this reliance to enable generalisation of the model.
- CIoU loss as the loss function which not only observes the overlap of bounding boxes with the ground truth (which is already done for IoU), but also how close the box was to the ground truth box in terms of the pixel distances within the image, which is an additional part of the loss function that is trained so that it enables the network to pull the predicted bounding box closer to the ground truth box.
- Using Mish activation as the activation function instead of ReLU which improves the performance of the model due to its ability to push the features created by the model towards its optimal [27][28].

5) YOLOv5: YOLOv5 represents the most recent iteration of the YOLO (You Only Look Once) series of object detection models, originally introduced in 2016. Developed by Ultralytics, a reputable computer vision research company,

YOLOv5 offers notable advancements in both accuracy and speed when compared to its predecessors. It attains state-of-the-art performance across multiple benchmark datasets while preserving real-time inference speeds on modern GPUs. This algorithm employs a single convolutional neural network (CNN) to predict object classes and bounding boxes by dividing the input image into a grid and making predictions based on each grid cell. This approach enables faster inference times and improved accuracy compared to region-based CNNs used in other object detection models. YOLOv5 has gained significant traction within the computer vision community and finds applications in various domains, including autonomous vehicles, robotics, security, and more [26]. YOLOv5 contains three layers as an object detection model: Backbone as the feature extractor, the Neck which combines and mixes different features extracted from the Backbone, and the Head which takes the outputs from the Neck and predicts bounding boxes and classification. The Backbone that mentioned in Fig. 4 was CSPDarknet. CSPDarknet is a neural network that contains a set of convolutional layers which are useful for consolidating images and extracting useful features which can be learnt from the model. As shown in Fig. 4, the Backbone consists of a set of BottleNeckCSP (Cross Stage Partial) blocks and a Spatial Pyramid Pooling (SPP) block. The bottleneck part of BottleNeckCSP helps reduce the number of feature maps, which in turn reduces the model size and computation. The cross stage partial part of BottleNeckCSP also aids in reducing the model size and computation required by reducing the amount of gradient information during optimization within the network while maintaining good accuracy for the model. It does this by splitting the base layer into two parts (one as the base layer, the other is partitioned into multiple blocks) and merging them back together again through a cross-stage hierarchy strategy. The SPP block performs pooling of the features from the previous CSP block to generate fixed-length outputs. This avoids the need to do any cropping, warping or preprocessing at the start of the input to the neck and is done through pooling which is an information aggregation function [30].

After the Backbone, the feature mixing and combining model used was Path Aggregation Network (PANet). PANet is a network architecture with a bottoms-up approach, where there is a feature hierarchy which aggregates and passes the information of multiple convolutional layers at different stages, enhancing the signals between lower layers and upper layers. Linking different feature levels together to allow the model to accurately detect both larger and small objects when performing object detection.

As shown in Fig. 4, there are several concatenation blocks which combine the lower and higher-level features together to be fed into the final head layers. The final layer is a set of 1x1 convolutional layers that takes in the input of the Neck (PANet) to pass into the regression that detects the bounding boxes and classifies them, and these are then used for training and inference/prediction [28]. During training, YOLOv5 will see the images inputted from the training dataset, use the Backbone (CSPDarknet) to extract out relevant features, thereafter, utilizing the PANet to concatenate lower and higher-level features together, and these are finally passed to output the bounding boxes and classes for different region of the image as predictions. This will be trained using the training dataset and can be used for inference after enough training is done for the model [29].

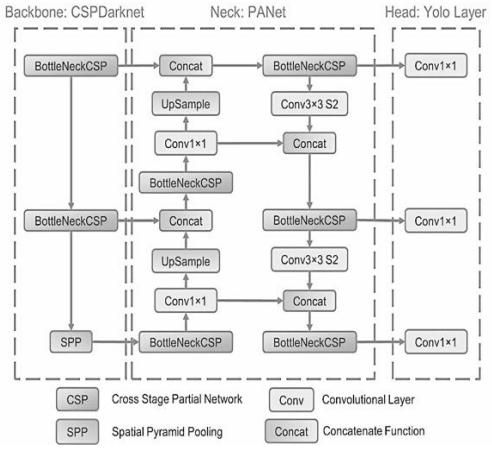


Fig. 4. YOLOv5 model architecture.

6) MobileNet Single Shot MultiBox Detector (SSD): The MobileNet Single Shot MultiBox Detector (SSD) is an object detection algorithm that combines the Single Shot Detector (SSD) framework with the MobileNet architecture. This algorithm was developed by Wei Liu et al. in 2016 [31]. MobileNet is specifically designed for mobile and embedded devices, offering a lightweight convolutional neural network architecture that utilizes depth-wise separable convolutions. These convolutions help reduce computational requirements while maintaining high accuracy. On the other hand, the SSD framework is a popular approach for object detection that utilizes a single convolutional neural network to predict object classes and bounding boxes [31]. MobileNet SSD enables real-time object detection on devices with limited computational resources, such as mobile devices and robotics. It is particularly beneficial for applications that require real-time object detection, including autonomous vehicles and surveillance. Despite its efficiency, MobileNet SSD achieves high accuracy on benchmark datasets like PASCAL VOC and COCO, all while maintaining fast inference times. As a result, it has gained significant adoption within the computer vision community and finds applications in various domains such as security, surveillance, and robotics [15][24]. MobileNet SSD comprises two key layers: a backbone model used to extract relevant features (in this case, VGG-16 was employed as the feature extractor), and the detector head, which outputs crucial information for object detection.

The VGG-16 model, proposed by researchers at the University of Oxford in 2014, serves as the backbone for many computer vision tasks. It is a convolutional neural network architecture comprising 16 layers, including 13 convolutional layers, 5 max pooling layers, and 3 fully connected layers. The

primary purpose of the convolutional layers is to extract meaningful features from the input image, enabling the model to capture relevant patterns and structures. On the other hand, the pooling layers play a crucial role in reducing the spatial dimensionality of the extracted features. By down sampling the feature maps, these pooling layers enhance computational efficiency during subsequent processing stages. Together, the combination of convolutional and pooling layers in the VGG-16 architecture enables effective feature extraction and representation for a wide range of computer vision applications. Finally, the fully connected layers learn to classify the extracted features into their corresponding categories. The detector head consists of several convolutional blocks that link to the detection block, as well as a post-processing step called Non-Maximum Suppression (NMS) as shown in Fig. 5. The purpose of the convolutional blocks that are linked at different levels to the detection head is to extract features at multiple levels, which enables the model to detect both small and large objects by extracting features for them. Subsequently, the NMS block will take the bounding boxes that are outputted from the model and pick out the bounding box that is closest to the ground truth bounding box using Intersection-over-Union (IoU) as the metric. NMS is only used during training. As an overview for the methodology of MobileNet SSD, the model will be fed the images from the training dataset, which goes through a feature extraction process in VGG-16 as the backbone. These feature extractors are then further convoluted, and features at different levels of convolutions are passed to the detection head for bounding box and classification prediction. For training, there is an additional process of NMS which choose the most prominent bounding box for each ground truth box. The pretrained SSD MobileNet v1 FPN with dimension of 640x640 were used for detecting the objects. it is an object detection model based on a single-shot detection (SSD) architecture with a feature pyramid network (FPN) and uses the MobileNet V1 neural network as a base feature extractor. This model is designed to detect objects in images of size 640x640 pixels. The SSD architecture is a popular object detection approach that predicts object categories and bounding boxes in a single forward pass through the neural network. The SSD MobileNet v1 FPN 640x640 model consists of a base network, feature pyramid network, and detection network.

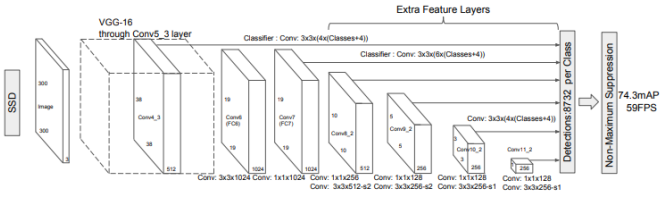


Fig. 5. MobileNet Single Shot Multi box Detector (SSD).

- Base Network: The base network of the model is the MobileNet V1 neural network. It is a lightweight deep neural network architecture that uses depth wise separable convolutions to reduce the number of parameters and improve computational efficiency. The MobileNet V1 architecture consists of a sequence of depth wise separable convolutional layers followed by standard convolutional layers, which are used to extract feature maps from the input image.
- Feature Pyramid Network: The feature pyramid network (FPN) is used to combine feature maps from different levels of the MobileNet V1 base network. The FPN is a top-down architecture that aggregates high-resolution feature maps from the lower levels of the base network with lower-resolution feature maps from the higher levels of the network. This creates a pyramid of feature maps with rich semantic information at multiple scales, which is useful for detecting objects of varying sizes in the input image.
- Detection Network: The detection network is used to predict the bounding boxes and object categories in the input image. The detection network consists of a set of convolutional layers that process the feature maps generated by the FPN. These layers are used to predict the locations and class scores of the objects in the input image. The SSD MobileNet V1 FPN 640x640 model uses a set of default anchor boxes at different scales and aspect ratios to generate object proposals. These proposals are then refined by the detection network to improve the accuracy of the final object detection results.

The SSD MobileNet V1 FPN 640x640 is a powerful object detection model that combines the strengths of the MobileNet v1 architecture, the feature pyramid network, and the single shot detection approach to achieve high accuracy and computational efficiency.

a) Parameters of the MobileNet SSD Model

The parameters for this model are as follows:

- Backbone architecture: This model uses a MobileNet V1 architecture as the backbone. MobileNet is a lightweight convolutional neural network architecture designed for mobile devices, which makes it suitable for real-time object detection on low-power devices.
- Feature Pyramid Network (FPN): This model uses a Feature Pyramid Network (FPN) to generate a multi-scale feature map. FPN is a technique used to extract features from images at different scales, which helps improve the accuracy of object detection.
- Input size: The input size for this model is 640x640 pixels. This means that the model can detect objects in images up to 640x640 pixels in size.
- Batch size: The batch size is the number of images that are processed simultaneously. The batch size of 8 was used for training the model.
- Learning rate: The learning rate is a hyperparameter that controls how much the model adjusts its parameters during training.
- Number of classes: The number of classes is the number of object categories that the model can detect. In this case it was five classes.

- Anchor boxes: Anchor boxes are a set of predefined bounding boxes of different sizes and aspect ratios that the model uses to detect objects.
- NMS threshold: The Non-Maximum Suppression (NMS) threshold is a parameter that controls how much overlapping bounding boxes are merged into a single detection. The default NMS threshold for this model is 0.6, but it can be changed depending on the specific use case.

b) Model Architecture

The details of model architecture with parameter setting for threat detection is given as:

- Number of classes: This specifies the number of classes or object categories that the model will detect. In this case, the model is trained to detect five classes.
- Image resizer: This defines how the input image is resized to fit the input size of the model. Here, a fixed shape resizer is used with a height and width of 640 pixels.
- Feature extractor: This defines the feature extraction backbone of the model. In this case, the SSD MobileNet V1 FPN Keras architecture is used. The depth multiplier is set to 1.0, and the minimum depth of the network is set to 16. The conv hyperparams section specifies the hyperparameters used for the convolutional layers of the feature extractor, including the L2 regularization weight, random normal weight initializer, and batch normalization parameters.
- Override base feature extractor hyperparams: This indicates that the hyperparameters specified in this pipeline config file will be used to override the default hyperparameters of the base feature extractor.
- FPN: This specifies the Feature Pyramid Network used for multi-scale feature extraction. The minimum and maximum levels of the feature pyramid are set to 3 and 7, respectively.
- Box Coder: In object detection, the box coder is used to encode and decode the predicted boxes, which is necessary because the predicted boxes are in a relative format and need to be converted back to the absolute coordinates of the image. The box coder section specifies the method used to encode and decode boxes. In this particular case, the box coder is using the faster R-CNN box coder method, which encodes boxes using their center coordinates, width, and height. The y scale and x scale values specify the scaling factors for the center coordinates, while the height scale and width scale values specify the scaling factors for the height and width. These scaling factors are used to normalize the box coordinates to a similar range.
- Matcher: The matcher section specifies the method used to match predicted boxes to ground truth boxes. In this case, the argmax matcher method is used, which matches predicted boxes to ground truth boxes based on their maximum intersection-over-union (IoU) overlap. The matched threshold value specifies the minimum IoU overlap required for a predicted box to be considered a match, while the unmatched threshold value specifies the maximum IoU overlap allowed for a predicted box to be considered unmatched.
- Similarity Calculator: It specifies the method used to calculate the similarity between predicted boxes and ground truth boxes. In this case, the IoU similarity method is used, which calculates the IoU overlap between two boxes.
- Box Predictor: In the SSD MobileNet V1 FPN 640x640 model, the box predictor is responsible for predicting the bounding boxes for the detected objects. The weight shared convolutional box predictor is used as the box predictor, which shares weights between the class prediction and box prediction layers. This helps to reduce the number of parameters in the model. The depth parameter specifies the number of filters in each convolutional layer of the box predictor. In this model, it is set to 256.
- Number of layers before Predictor: this parameter specifies the number of convolutional layers before the predictor layers. In this model, four convolutional layers are used before the predictor.
- Kernel Size: It specifies the size of the convolutional kernel used in the predictor layers. In this model, a kernel size of 3 is used.
- Class prediction Bias init: It initializes the bias for the class prediction layer. In this model, it is initialized to -4.599999904632568.
- Convolutional Hyperparameters: It specifies the hyperparameters for the convolutional layers in the box predictor. It includes the regularizer, initializer, activation function, and batch normalization parameters.
- L2 Regularizer: It applies L2 regularization to the convolutional layers to prevent overfitting. The weight value provided is 3.9999998989515007e-05.
- Random Normal Initializer: This parameter initializes the weights of the convolutional layers using a normal distribution with a mean of 0 and a standard deviation of 0.009999999776482582.
- Activation: The activation parameter specifies the activation function used in the convolutional layers. In this model, the RELU_6 activation function is used.
- Batch Normalization: It applies batch normalization to the convolutional layers to improve the training process. It includes the decay rate, scale, and epsilon values. In this model, the decay rate is set to 0.996999979019165, the scale is set to true, and the epsilon is set to 0.0010000000474974513.
- Anchor boxes: In object detection, anchor boxes are pre-defined bounding boxes of various sizes and aspect ratios that are used to identify objects in an image.

- Anchor Generator: It specifies how these anchor boxes should be generated. In the SSD MobileNet V1 FPN 640x640 model, the anchor generator uses the multiscale anchor generator which generates anchors at multiple scales and aspect ratios. The values provided in the multiscale anchor generator section is the following:

o Min level and max level: These specify the minimum and maximum levels of feature maps in the FPN.

In this case, the feature maps are generated at levels 3 to 7.

- Anchor Scale: This specifies the base size of the anchor boxes. The size of the anchor boxes is proportional to the square root of the area of the feature map.
- Aspect Ratios: These specify the aspect ratios of the anchor boxes. In this case, three aspect ratios are used: 1.0, 2.0, and 0.5.
- Scales per Octave: This specifies the number of scales to be used per octave. In this case, two scales are used per octave.
- Score Threshold: minimum confidence score for detections to be considered. Value is 9.99999993922529e-09.
- IoU Threshold: intersection over union (IoU) threshold used for non-maximum suppression. Value is 0.6000000238418579.
- Max Detections per Class: maximum number of detections to keep per class after non-maximum suppression. Value is 100.
- Maximum total Detections: maximum number of detections to keep over all classes after non-maximum suppression. Value is 100.
- Use Static Shapes: whether to use static shapes for the output tensor shapes. Value is false.
- Score Converter: method for converting scores. Value is SIGMOID.
- Normalize loss by number of Matches: whether to normalize the total loss by the number of matched ground truth boxes. Value is true.
- Localization Loss: weighted_smooth_l1: the localization loss function. No values provided, uses default parameters.
- Freeze Batch norm: whether to freeze the batch normalization parameters during training. Value is false.
- Batch Size: The number of images that are fed into the network at once during training. In this case, the batch size is set to 8.
- Data Augmentation Options: A list of data augmentation options to apply to the input images during training. In this case, two types of data augmentation are used: random horizontal flips and random crops.
- Sync Replicas: A Boolean variable that controls whether to use synchronous gradient updates during training. When set to true, the gradients are computed and averaged across all replicas before the weights are updated. This can lead to better convergence but requires more memory and communication.
- Optimizer: Specifies the optimizer used during training. In this case, the momentum optimizer is used with a cosine learning rate schedule.
- Learning Rate: The learning rate schedule used during training. The learning rate is decreased according to a cosine schedule that decreases the learning rate from a base value of 0.04 to a final value of 0 over 25,000 steps. The learning rate is also gradually increased from a warmup value of 0.0133 over 2,000 steps.
- Momentum Optimizer Value: The momentum value used by the optimizer. In this case, the momentum is set to 0.9.
- Use Moving Average: A Boolean variable that controls whether to use a moving average of the model weights during training. When set to false, the raw weights are used. When set to true, the moving average of the weights is used instead, which can improve the robustness of the model.

7) Python: Model implementation, training and evaluation is done in python programing using several libraries as listed below:

- PyTorch: PyTorch is a python library, a deep learning framework for building and training neural networks, widely used for research and production in machine learning programming language and the Torch library. Torch is an open-source ML library used for creating deep neural networks and is written in the Lua scripting language. It's one of the preferred platforms for deep learning research. Outcome of PyTorch is a model file that can be loaded in mobile device and can be used for prediction. The researchers used PyTorch for MobileNet and yolov5 implementation, training, and evaluation.
- TensorFlow: An open-source machine learning framework for developing and deploying machine learning models, including deep learning models. Both yolov5 and MobileNet SSD models can be implemented in PyTorch and TensorFlow these are just two standard libraries for implementing neural networks. Researchers tried TensorFlow for implementation, but PyTorch was more user friendly, so PyTorch was adopted.
- OpenCV: An open-source computer vision library offering tools for image and video processing, including object detection and analysis processing. OpenCV supports a wide variety of programming languages like Python, C++, Java, etc. It can process images and

videos to identify objects, faces, or even the handwriting of a human. Researchers have used OpenCV for reading images and applying preprocessing steps like scaling and normalization of images.

- RoboFlow: A platform for managing, annotating, and preprocessing data for computer vision projects, assisting in training machine learning models.
- Matplotlib**:** A Python plotting library used for creating static, interactive, and animated visualizations in data analysis and model output visualization. Matplotlib is also used for evaluation and analysis of results like building confusion metrics after prediction is done through this library. Outcome of matplot lib is plots and charts generated as images that can be used for visualization of training loss and accuracy with each epoch.
- Seaborn: A statistical data visualization library built on top of Matplotlib, designed to generate informative and attractive statistical graphics. Seaborn is also used for plotting various analysis charts of training loss and accuracy values.

H. Model Evaluation

1) Model evaluation metrics: Mean Average Precision (MAP) is the universal standard metric used to compare performance between object detection models created by different authors [32]. This metric is specifically derived from Average Precision (AP). Since classification is performed during object detection for different bounding boxes along with the provision of the ground truths, the fundamentals of the confusion matrix apply. The matrix enables computations to be made with accuracy, precision, and recall. It consists of the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values [32]. In the context of object detection:

- True Positive: Detection made correctly by the model.
- True Negative: Background region correctly detected by the model (where there are no objects)
- False Positive: Wrongly detected regions made by the model.
- False Negative: Regions where the ground truths are missed by model.

Intersection-over-Union (IoU) is the next metric used to determine whether bounding boxes are TP, TN, FP or FN. IoU is defined as the area of overlap between the bounding box predictions and the ground truth, divided by the area of union between them [33]. An IoU of 1 means the bounding boxes predicted match exactly the ground truth boxes, whereas an IoU of 0 depicts no overlap between the two bounding boxes [34]. Fig. 6 shows the IoU equation.

A threshold is a hyperparameter predetermined to decide between TP, TN, FP or FN. For example, with a threshold of 0.5 for a ground truth bounding box, if the predicted bounding box has an IoU greater than 0.5 for that particular ground truth, it is considered a TP. Whereas an IoU lower than 0.5 means the predicted bounding box is a FP. Through the IoU, we are able to determine the confusion matrix (TP, TN, FP, FN) for every bounding box, and subsequently calculate the precision and recall.

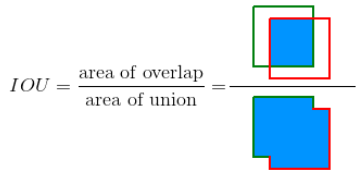


Fig. 6. Intersection over union equation.

- Precision: of the positive classes that are correctly detected, how many are actually positive? This follows the following Eq. (1) [33]:

$$\text{Precision} = TP / TP+FP \quad (1)$$

- Recall: of all positive classes, how much can we predict the class correctly? It is preferred that this measure is as high as possible. It follows the following Eq. (2) [33]:

$$\text{Recall} = TP / TP + FN \quad (2)$$

We also, calculate the accuracy that measure considers the correct classification out of all classes, where a high value of accuracy if preferred. This factor is explained in the following Eq. (4) [33]:

$$\text{Accuracy} = TP + TN / TP + FP + TN + FN \quad (3)$$

AP uses precision and recall creating an Area Under the Curve graph (AUC-PR) for the model. For every threshold, there is a different precision since the object detector will output a confidence score, which is then determined by the threshold on whether each bounding box is TP, TN, FP or FN. AP will take the precision and recall at every threshold, graph out a precision-recall plot, and thereafter take the area under the curve. The closer AUC is to 1, the better the model, and vice versa. It is better to have a high AUC as the model is good at predicting and distinguishing between classes it calculated based on the TPR (y-axis) versus the FPR (x-axis). AP is done separately for each class within the object detection model. Also, F1-measure that measure mainly computes the harmonic mean of precision and recall measuring them at the same time [33]. It has the below Eq. (4):

$$\text{F1-measure} = 2 \text{ x Recall x Precision / Recall + Precision} \quad (4)$$

To consolidate all these scores into one metric, mAP was introduced. mAP will take all the AP of each class (will have n number of scores for n number of classes) and take the mean of those AP to obtain one score. This score is the metric used to determine the overall performance of the object detection model.

2) Model validation: Model validation refers to the procedures and actions conducted to verify that a model is functioning as intended and aligns with its objectives and intended business applications. Typically, the validation

process involves the assessment performed by individuals who are not the model developers or owners, as their impartial perspective is valuable due to their non-technical background [31]. In the context of machine learning, validation may involve ML experts evaluating the labeling process to ensure its accuracy and reliability.

IV. Results and Discussion

This section aims particularly to answer the research questions which is: is there any efficient way to detect threats from live videos using DL algorithms? The other question was how accurate and efficient is using the MobileNet model to detect objects from live videos? The analysis has been conducted according to detailed steps that will be mentioned on the analysis Section 4.3. Comparison is done based on mean average precision (mAP) and frames per second (FPS) on the datasets collected as a part of research findings.

A. Data Preparation

1) Dataset collection: For the purpose of this study, a carefully curated dataset was created to support the research objectives. The dataset was generated and collected by the researcher from various sources, specifically chosen to include objects that are considered potentially dangerous or threatening. The dataset comprises five classes of objects: fire, guns, knives, arrows, and swords. These object classes were selected due to their relevance in public safety incidents and their potential to cause harm.

To obtain the necessary data, a range of sources was utilized. This included gathering footage from YouTube videos that contained CCTV recordings, educational videos demonstrating fighting skills, and demo videos. By extracting frames from these videos at a rate of 1 frame per second (1fps), the required images for the dataset were captured. The inclusion of CCTV footage is particularly valuable as it provides authentic data from public places where security cameras are commonly employed.

To annotate the dataset with the necessary information for object detection, the researcher employed the RoboFlow tool. RoboFlow is a widely used image annotation tool specifically designed for object detection tasks. Annotations provide crucial additional details about the images, such as the precise location and class of the object within each image. These annotations are vital for training machine learning models to accurately detect and classify objects. To enhance the diversity of the dataset and improve the performance of the machine learning models, various augmentations were applied to the images. Techniques such as rotation and contrast adjustment were employed to create variations of the original images.

By introducing these augmentations, the models were exposed to a wider range of data, enabling them to better handle different image conditions and variations. The careful curation of this dataset, encompassing the five specified object classes, serves as a crucial foundation for training machine learning models to accurately detect these objects and contribute to public safety improvements.

2) Data cleaning: In this research, the process of data cleaning and maintaining data quality played a crucial role in preparing the dataset for object detection models. After extracting images from videos, uninformative and vague frames were carefully discarded to ensure that only relevant and clear frames were included in the dataset. This step aimed to eliminate any noise or ambiguity that could hinder the performance of the models. The remaining frames with clear object visibility were then subjected to manual labeling for five object classes. Manual labeling involves human annotators carefully marking the objects of interest in each frame, providing accurate ground truth annotations. To ensure a balanced distribution of images per object class in the training, validation, and testing sets, the data classes were sampled strategically. This sampling process helps prevent bias towards specific object classes and ensures that the models are exposed to a diverse range of objects during training and evaluation. By performing data cleaning, manual labeling, and strategic sampling, the researchers improved the overall quality and representativeness of the dataset. This, in turn, enhances the reliability and generalizability of the object detection models, allowing them to effectively detect and classify objects in various real-world scenarios.

3) Data preprocessing: To ensure effective model training, data preprocessing techniques were applied to the dataset. One of the key preprocessing steps involved resizing the images to a standardized scale of 416x416 pixels, which is suitable for the MobileNet SSD model input. This resizing step helps to ensure consistency in the input size across all images, facilitating efficient model training. Data preprocessing serves to enhance the quality of the data and reduce training time. In addition to resizing, other preprocessing operations were applied to make the images compatible with the model requirements. These operations included scaling, cropping, and further resizing to bring all images into a consistent format prior to model training. By standardizing the images, the model can effectively process and analyze them. Data augmentation techniques were also employed to increase the diversity and veracity of the data, thereby enabling the model to learn better. Data augmentation involves applying various transformations to the images to create additional training samples. These transformations include rotation, blurring, and adjustments in contrast and orientation. By introducing such variations, the model becomes more robust and capable of handling different image conditions and variations that may be encountered in real-world scenarios. The combination of data preprocessing, including resizing and standardizing the images, along with data augmentation techniques, enhances the quality, variety, and quantity of the training data. This, in turn, contributes to the overall performance and generalization capabilities of the object detection model during training and subsequent inference tasks.

4) Data annotation: During this phase, the researcher utilized the RoboFlow annotation tool to facilitate the

annotation process for object detection. This tool allowed for the drawing of bounding boxes around the objects of interest in the images. By uploading videos into RoboFlow, images with varying frames per second (FPS) rates were extracted. Each image was then labeled with its corresponding class name and bounding box. For each image, bounding boxes were manually drawn around the instances of the objects to be detected. The researcher carefully assigned the correct class label from the predefined set of five classes: fire, gun, knives, arrows, or swords. This manual annotation process ensured accurate placement of the bounding boxes and correct labeling, as these annotations served as the ground truth for the subsequent model training phase. Once the dataset was annotated, the next step involved integrating it with the respective training frameworks. The annotations created by the researcher were utilized as the training data for the object detection models. These annotations, combined with the corresponding images, formed a labeled dataset that could be used to train the models and enable them to detect and classify objects accurately.

a) Data Quality Check

Below is the explanation of the used methods and steps taken to ensure data quality in object detection:

- Annotation Quality Control: Ensure accurate and consistent annotation of bounding boxes and class labels in the dataset. Annotators should follow clear guidelines and have a solid understanding of the objects of interest. This research has five object types (guns, swords, knives, arrows, and fire) in the dataset. Researchers have annotated each object carefully using RoboFlow. Bounding boxes were created with extreme care.
- Data Cleaning and Preprocessing: researchers removed duplicate images after annotations to build a good model. Also, we have eliminated corrupted images or annotations that might negatively impact training.
- Balanced Class Distribution: researchers ensured that the dataset has a balanced distribution of objects across classes to prevent bias towards dominant classes and improves the model's ability to detect all classes accurately. Normal videos are recorded at 30 fps, means 30 frames per second can be extracted and in this case, 1 frame/sec was extracted. There were some images that do not have any object, such images were deleted, and remaining images were annotated accurately. Exact number of samples before augmentation for all objects is provided in Table I below:
- Data Augmentation: researchers have applied data augmentation techniques such as random rotation between -15 degree to + 15 degree to increase the number of data samples.

TABLE I. CLASSES SAMPLES

Classes	Train	Test	Valid
Arrow	1286	186	360
Gun	1362	209	379
Sword	1370	221	394
Fire	1297	197	353
Knife	1312	198	369

- Validation and Anomaly Detection: researchers have eliminated images that have no object or object is not clearly visible, etc.
- Consistent Image Quality: researchers have ensured that images are of consistent quality and resolution. They have applied resizing for all images to be 416x416.
- Real-World Scenario Simulation: researchers have collected videos from demos, CCTV videos and social media to collect diverse and real-world scenarios to train a good model.
- Review Annotations for Ambiguity: researchers have reviewed annotations that are ambiguous or challenging for the model to detect, such as partially occluded objects or objects in cluttered scenes.
- Class Label Consistency: researchers have verified that class labels are consistent across annotations. With every annotation object name was specified with that annotation to make sure that each object has its correct name.
- Cross-Validation: Divide the dataset into training, validation, and test sets. Cross-validation can help assess how well the model generalizes by training on one subset and testing on another.
- Continuous Monitoring: Continuously monitor and update the dataset as needed. Over time, as the model's requirements change or new challenges arise, the dataset should evolve accordingly.
- Use External Data Sparingly: When using external data sources like stock images or online datasets, ensure that they are relevant and high-quality. External data should complement the dataset without introducing noise.
- Expert Review: researchers engaged experts' volunteers to review and validate the quality of the dataset, ensuring that the annotations and data align with the real-world scenarios.
- Feedback Loop: Establish a feedback loop with annotators to address questions, provide clarification on guidelines, and continually improve annotation quality.

5) Model evaluation

a) Results of Performance Metrics Comparison

In Table II researchers indicate the difference across different aspects. These results are achieved after training both models on same dataset. A dataset is divided into three parts which are: training, validation, and testing. For each class 70%, 20% and 10% images are used for train, valid test, respectively.

- Train set is provided to the model during training so model can learn pattern from this data.
- Validation set is used to evaluate model during training, this is unseen for model but during training results are analyzed through this unseen data. If model is not learning correctly then we tune the parameters of model to see if is performing good on train data and validation data.
- Test set is totally unseen that is used after correct training of model, it depicts the real-world testing of model on unseen data. If a model performs as good on testing data as it is for training, then model is considered to be reliable for that task.

TABLE II. PERFORMANCE METRICS RESULTS

Performance Metrics	MobileNet SSD	YOLOv5
mAR (Mean Average Recall)	0.9565	0.8450
mAP (Mean Average Precision)	0.9125	0.7549
IoU (Intersection over Union)	0.9045	0.8020
False Positive Rate	0.053	0.078
False Negative Rate	0.053	0.078
Inference Speed	~ 18ms/image (CPU)	~ 41ms/image (CPU)
Memory Usage	1.3 GB	7 GB
Model Size	6.6 mbs	15 mbs
Class-wise Performance	90%	83%
F1 score	0.92	0.81

To evaluate and measure the model performance AUC method were used. This method is used to check the ability of the model to detect accurately among the different classes. The higher the AUC, the better the performance of the model is. Fig. 7 illustrates the AUCs of the MobileNet model.

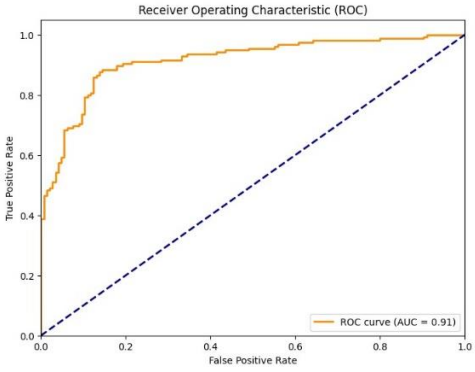


Fig. 7. ROC AUC curve.

The graph in Fig. 7 provides the following results:

- Since AUC = 0.91, the model is able to distinguish perfectly between all positive and negative class points.

The graph in Fig. 8 and 9 illustrate the results of the confusion matrix. The confusion matrix plot shows the predicted vs. true labels, and the values in each cell represent the percentage of the correct and incorrect classifications. Diagonal values are high as they show values for correct classification and off diagonal values are incorrect classification. As per the graphs the MobileNet model prove that it detects the classes efficiently over the Yolov5 model.

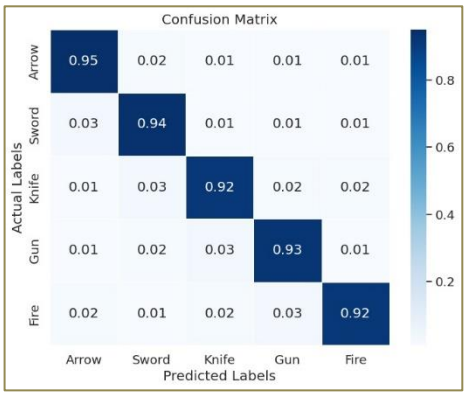


Fig. 8. Confusion matrix for mobilenet SSD model.

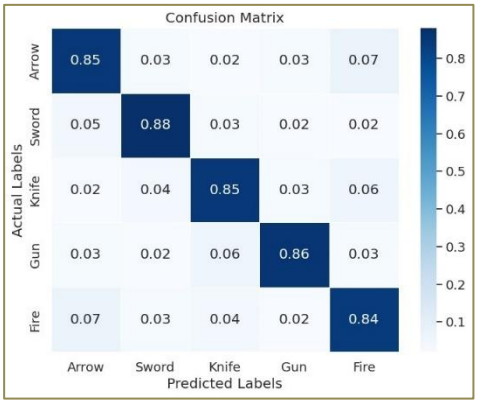


Fig. 9. Confusion matrix for YOLO model.

6) Model validation: In this research, the model validation was a meticulous process that contributed significantly to the success of the object detection models. The researcher dedicated time and efforts to ensure the annotations were accurate and comprehensive, as this foundation would directly impact the performance of the models in real-world scenarios. This validation step is carried out by ML experts who as their impartiality is crucial as the context of the labeling process, a ML expert is often involved in performing this step to validate the accuracy and quality of the labels assigned to the data [31]. There were experts in ML who volunteering to examine the validity of the model. The feedback from the experts were that all the dataset was accurately labeled since it was at first a manual step that takes a lot of time and efforts.

For the validity of the model, Table II examines the success rates, encompassing accuracy and F1-measure outcomes. The MobileNet SSD model exhibits an impressive success score of 0.92. Moreover, it achieves an accuracy rating of 96.5%. These findings collectively validate the superior performance of the

MobileNet model over the YOLO model in detecting threatening objects.

7) Discussion: In conclusion, MobileNetSSD performs better than YOLOv5 in the scenarios of detecting threatening objects due to its fast inference speed, memory efficiency, optimization for small objects, and training strategy as shown in Table II. Below are additional reasons why MobileNet model preferable over YOLO5 model. Certainly, here's the list of scenarios tailored to the context of the problem, considering the involvement of detection for five classes including: fire, gun, knives, arrows, and swords with a dataset of 2000 images per and a deployment on a mobile platform:

a) Mobile-Optimized Inference Speed: MobileNet SSD's fast inference speed is crucial for mobile deployments. It ensures that objects are detected rapidly on the mobile device, enhancing real-time detection capabilities.

b) Memory Efficiency for Mobile Devices: Deploying on mobile devices demands efficient memory usage. MobileNet SSD's architecture reduces memory requirements, allowing smooth operation on resource-constrained mobile platforms.

c) Real-Time Threat Detection on Mobile: MobileNet SSD's quick detection of objects like guns and knives is vital for real-time threat detection scenarios on mobile devices, such as identifying potential weapons in public spaces.

d) Streaming Video Analysis on Mobile: MobileNet SSD's fast inference speed aligns well with streaming video analysis on mobile devices. This is valuable for continuous monitoring using mobile cameras.

e) Optimization for Small Objects on Mobile: MobileNet SSD's specialization in detecting small objects, like arrows or knives, is advantageous for accurate detection on mobile screens, where these objects might appear relatively small.

f) Responsive Fire Detection on Mobile: Fast detection of fire instances using MobileNet SSD on mobile devices is critical for timely response to fire incidents, aiding firefighting efforts and safety protocols.

g) Edge Computing for Mobile: Deploying MobileNet SSD on mobile platforms extends the benefits of edge computing. The model's lightweight architecture is suitable for processing data on the device, reducing latency.

h) Mobile Surveillance Solutions: MobileNet SSD's deployment on mobile devices allows for portable surveillance solutions. Users can leverage their mobile phones for security monitoring, quickly detecting threats like fires or intruders.

i) Accurate Object Detection on Mobile: MobileNet SSD's optimization for small object detection ensures accurate identification of objects like arrows or swords on mobile screens, where details matter.

j) Reduced Data Transmission: MobileNet SSD's on-device detection reduces the need for transmitting sensitive data to remote servers, maintaining user privacy and potentially reducing data costs.

k) User-Friendly Mobile Applications: The combination of fast detection and accuracy makes MobileNet SSD suitable for developing user-friendly mobile apps that offer intuitive and effective object detection functionalities.

l) Cost-Effective Mobile Deployments: MobileNet SSD's low computational demands align with mobile platforms, making it a cost-effective choice for deploying object detection capabilities on mobile devices

V. Recommendations for Future Work

Several recommendations can be made for future study that will extend the present study's findings. Below are some possible recommendations for future studies:

1) Expand the scope of the study: The current study may have focused on a specific aspect or application of the topic. Future studies could expand the scope of the research to include other related areas, applications, or datasets.

2) Improve the performance of the model: The current study may have achieved good results with the model used, but, there may be other models or techniques that could improve performance further. Future studies could explore alternative models or techniques for the task and compare their performance.

3) The current study may have some limitations due to the dataset used. Future studies could address these limitations by using different datasets, models, or evaluation metrics.

4) Explore ethical considerations: The current study may not have explicitly addressed the ethical implications of the research. Future studies could explore the ethical considerations of the research and investigate ways to ensure that the technology is used ethically and responsibly.

Acknowledgment

My deepest gratitude and sincere appreciation go to all those who have contributed to the completion of this thesis. My supervisor, my family, and my friends. Their unwavering support, guidance, and encouragement have been invaluable throughout this academic journey. This thesis stands as a testament to the collective efforts and unwavering support from each and every one of you. Thank you for being a part of this remarkable journey and for helping me reach this significant milestone in my academic career. This research stands as a testament to the collective efforts and unwavering support from each and every one of you. Thank you for being a part of this remarkable journey and for helping me reach this significant milestone in my academic career.

References

[1] Z.-Q. Zhao, P. Zheng, S. Xu, and X. Wu, "Object detection with deep learning: A review," IEEE Trans. neural networks Learn. Syst., vol. 30, no. 11, pp. 3212–3232, 2019.

[2] Suharto, A. P. Widodo, and E. A. Sarwoko, "The use of mobilenet v1 for identifying various types of freshwater fish," in Journal of Physics: Conference Series, 2020, vol. 1524, no. 1, p. 12105.

[3] Z. Lin and W. Guo, "Cotton stand counting from unmanned aerial system imagery using mobilenet and centernet deep learning models," Remote Sens., vol. 13, no. 14, p. 2822, 2021.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Adv. Neural Inf. Process. Syst., vol. 28, 2015.

[5] N. Jaccard, T. W. Rogers, E. J. Morton, and L. D. Griffin, "Detection of concealed cars in complex cargo X-ray imagery using deep learning," J. Xray. Sci. Technol., vol. 25, no. 3, pp. 323– 339, 2017.

[6] Zeng, K. et al. (2022) 'FPGA-based accelerator for object detection: A comprehensive survey', The Journal of Supercomputing, 78(12), pp. 14096–14136. doi:10.1007/s11227-022- 04415-5.

[7] Hayat, S. et al. (2018) 'A deep learning framework using convolutional neural network for multi-class object recognition', 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC) [Preprint]. doi:10.1109/icivc.2018.8492777.

[8] Kanimozhi, S. et al. (2021) 'Key Object Classification for action recognition in tennis using cognitive mask RCNN', Proceedings of International Conference on Data Science and Applications, pp. 121–128. doi:10.1007/978-981-16-5348-3_9.

[9] Joshi, A.S. et al. (2020) 'Deep Learning Framework to detect face masks from video footage', 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN) [Preprint]. doi:10.1109/cicn49253.2020.9242625.

[10] AlAjlan, S.A. and Saudagar, A.K. (2020) 'Machine Learning Approach for threat detection on social media posts containing Arabic text', Evolutionary Intelligence, 14(2), pp. 811–822. doi:10.1007/s12065-020-00458-w.

[11] M. F. Shakeel, N. A. Bajwa, A. M. Anwaar, A. Sohail, and A. Khan, "Detecting driver drowsiness in real time through deep learning based object detection," in Advances in Computational Intelligence: 15th International Work-Conference on Artificial Neural Networks, IWANN 2019, Gran Canaria, Spain, June 12-14, 2019, Proceedings, Part I 15, 2019, pp. 283–296.

[12] Ren, S. et al. (2017) 'Faster R-CNN: Towards real-time object detection with region proposal networks', IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), pp. 1137–1149. doi:10.1109/tpami.2016.2577031.

[13] N. Jaccard, T. W. Rogers, E. J. Morton, and L. D. Griffin, "Detection of concealed cars in complex cargo X-ray imagery using deep learning," J. Xray. Sci. Technol., vol. 25, no. 3, pp. 323– 339, 2017.

[14] d S. Akcay, M. E. Kundegorski, C. G. Willcocks, and T. P. Breckon, "Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery," IEEE Trans. Inf. forensics Secur., vol. 13, no. 9, pp. 2203–2215, 2018.

[15] V. Riffo and D. Mery, "Automated detection of threat objects using adapted implicit shape model," IEEE Trans. Syst. Man, Cybern. Syst., vol. 46, no. 4, pp. 472–482, 2015.

[16] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," Int. J. Comput. Vis., vol. 77, pp. 259–289, 2008.

[17] A. Nurhopipah and A. Harjoko, "Motion Detection and Face Recognition for CCTV Surveillance System," IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 12, no. 2, pp. 107–118, 2018.

[18] B. Eamthanakul, M. Ketcham, and N. Chumuang, "The traffic congestion investigating system by image processing from CCTV camera," in 2017 International Conference on Digital Arts, Media and Technology (ICDAMT), 2017, pp. 240–245.

[19] N. Bordoloi, A. K. Talukdar, and K. K. Sarma, "Suspicious Activity Detection from Videos using YOLOv3," in 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1–5.

[20] A., A.B., P., P. and S., V. (2019) 'Detection of suspicious human activity based on CNN-DBNN algorithm for Video Surveillance Applications', 2019 Innovations in Power and Advanced Computing Technologies (i-PACT) [Preprint]. doi:10.1109/i-pact44901.2019.8960085.

[21] A. Younis, L. Shixin, S. Jn, and Z. Hai, "Real-time object detection using pre-trained deep learning models MobileNet-SSD," in Proceedings of 2020 the 6th international conference on computing and data engineering, 2020, pp. 44–48.

[22] Kumar, C.R. (2012) Research methodology. New Delhi: APH Publishing Corporation. page 366.

[23] Murthy, J.S. et al. (2022) 'ObjectDetect: A real-time object detection framework for advanced driver assistant systems using yolov5', Wireless Communications and Mobile Computing, 2022, pp. 1–10. doi:10.1155/2022/9444360.

[24] D. Biswas, H. Su, C. Wang, A. Stevanovic, and W. Wang, "An automatic traffic density estimation using Single Shot Detection (SSD) and MobileNet-SSD," Phys. Chem. Earth, Parts A/B/C, vol. 110, pp. 176–184, 2019.

[25] Sanjay Kumar, K.K. et al. (2020) 'A mobile-based framework for detecting objects using SSD-mobilenet in indoor environment', Intelligence in Big Data Technologies—Beyond the Hype, pp. 65–76. doi:10.1007/978-981-15-5285-4_6.

[26] Atik et al., (2022). Comparison of YOLO Versions for Object Detection from Aerial Images, International Journal of Environment and Geoinformatics (IJEGEO), 9(2):087-093 doi. 10.30897/ijegeo.1010741

[27] Diwan, T., Anirudh, G. and Tembhurne, J.V. (2022) 'Object detection using yolo: Challenges, architectural successors, datasets and applications', Multimedia Tools and Applications, 82(6), pp. 9243–9275. doi:10.1007/s11042-022-13644-y.

[28] Jiang, P. et al. (2022) 'A review of Yolo algorithm developments', Procedia Computer Science, 199, pp. 1066–1073. doi:10.1016/j.procs.2022.01.135.

[29] Yar, H. et al. (2023) 'A modified Yolov5 architecture for efficient fire detection in smart cities', Expert Systems with Applications, 231, p. 120465. doi:10.1016/j.eswa.2023.120465.

[30] Thuan, D. (2021) Evolution of YOLO Algorithm and Yolov5: The State-Of- The-Art Object Detection Algorithm.

[31] Liu, Wei & Anguelov, Dragomir & Erhan, Dumitru & Szegedy, Christian & Reed, Scott & Fu, Cheng-Yang & Berg, Alexander. (2016). SSD: Single Shot MultiBox Detector. 9905. 21-37. 10.1007/978-3-319-46448-0_2.

[32] Khurana, Y., 2019. Difference between Model Validation and Model Evaluation? [WWW Document]. Medium. URL https://medium.com/yogesh-khuranas- blogs/difference- between-model-validation-and-model-evaluation- 1a931d908240

[33] Narkhede, S., 2021a. Understanding Confusion Matrix [WWW Document]. Medium. URL https://towardsdatascience.com/ understanding-confusion-matrix- a9ad42dcfd62

[34] Rezatofighi, H. et al. (2019) 'Generalized intersection over union: A metric and a loss for bounding box regression', 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) [Preprint]. doi:10.1109/cvpr.2019.00075.

Developing an Improved Method to Remove Pectoral Muscle for Better Diagnosis of Breast Cancer in Mammography Images

Golnoush Abaei[1], Zahra Rezaei[2], Usama Qasim Mian[3],
Yasir Azhari Abdalgadir Abdalla[4], Nitin Mathew[5], Leong Yi Gan[6]
School of Computing Technologies, RMIT University, Melbourne, Australia[1]
Department of Computer Engineering, Marvdasht Branch, Islamic Azad University, Marvdasht, Iran[2]
School of Information Technology, Monash University Malaysia, Malaysia[3, 4, 5, 6]

***Abstract*—Mammography is a non-invasive method to study breast tissues for abnormalities. Computer-aided diagnosis (CAD) can automate the process of diagnosing malignant and benign tumors accurately. However, accurate results can be hampered by the presence of the pectoral muscle, which has a similar opacity to the breast tissue area. Detecting and removing pectoral muscles is not trivial due to various factors, and there are artifacts present near the pectoral muscle that can hamper proper segmentation. Given the significance of the topic, it is crucial to devise an accurate method for automatically detecting the muscle area in a mammography image and eliminating it from the rest of the image. This process of removing the pectoral muscle from the breast image can aid in precise segmentation and diagnosis of the tumor area, ultimately leading to faster diagnosis and better outcomes for patients. This study examined two segmentation algorithms, Level Set and Region Growing, for segmenting the pectoral muscle. An Improved Region Growing-based (IRG) algorithm was also proposed and showed promising results in automatically segmenting the pectoral muscle. All algorithms were tested on the MIAS dataset, and radiologists evaluated the results, showing an accuracy rating of up to 83% for IRG. The results indicated that IRG outperformed Level Set considerably due to many optimizations and modifications. IRG can be used as part of the preprocessing unit of an automated cancer diagnosis system.**

Keywords*—Breast cancer; preprocessing pectoral muscle segmentation; level set algorithm; region growing algorithm

I. Introduction

According to the World Health Organization (WHO), there were an estimated 2.3 million women diagnosed with breast cancer, and 685,000 died in 2020. Along with lung cancer, breast cancer is the most prevalent cancer worldwide, representing 12.3% of total diagnosed cancers in 2018. However, if the disease is caught early, treatments can be highly effective, with five-year survival probabilities of 90% and greater in advanced economies [1]. Unfortunately, this rate significantly drops to 66% in India and 40% in South Africa [2]. Mutebi and Anderson [1] mention that early detection is crucial for effective treatments and more so in the developing world since advanced-stage cancer treatments can be very costly and require advanced medical procedures along with a trained medical workforce to provide that treatment. These resources are not widely available in developing countries; hence, patients in these countries must get their cancers diagnosed early. To detect breast cancer early, WHO recommends yearly cancer screening for women at higher risk for developing this disease, which includes factors such as genetics, age, smoking, and drinking, among others. In addition, getting these screenings requires a trained radiologist to examine the mammography image carefully. Trained radiologists are scarce in many rural areas and often, an appointment with a specialist can be expensive for a big majority in the developing world. Furthermore, radiologists are also prone to intra/inter-observer variability errors.

One of the solutions to this problem is Computer-aided diagnosis (CAD) systems, which are highly recommended to assist radiologists in detecting breast tumors and outlining their borders. CAD systems usually use algorithms that include thresholding, region-based techniques, and edge detection techniques [3]. Mammography is an inexpensive and non-invasive method through which one can diagnose breast cancer in its early stages. As these images need interpretation by a radiologist, this may develop some problems due to fatigue, repetition, and the need for a great deal of attention to detail and other factors. Mammography can show changes in the breast up to two years before a physician can feel them. Computer-aided detection and diagnosis are considered to be one of the most promising approaches that may improve the efficiency of mammography.

In breast CAD, accurate breast segmentation is a crucial preprocessing step to speed up the subsequent processes without losing important anatomical information. However, breast and pectoral muscle segmentation is challenging, especially in scanned mammograms. This mostly happens due to the presence of some artifacts, such as duct tape and tags or it might be caused because of low contrast along the breast skin line and homogeneity between pectoral and breast tissues. Although many methods have been proposed for removing breast boundary and pectoral muscle and segmenting them, only a few have been evaluated quantitatively using all the images in the MIAS (Mammographic Image Analysis Society) database [4]. When performing mammography, the muscle area in the image can often appear similar to the tumor, making accurate segmentation difficult. To address this issue,

removing the muscle area before the segmentation stage is recommended. This study presents an improved method for accurately and quickly detecting and removing the pectoral muscle area from the mammography image, thereby preparing it for segmentation. Unlike existing methods, this method considers the separation of the muscle area during segmentation, avoiding the challenges posed using the information given about the tumor area, which may be unavailable in some datasets.

This empirical study is based on two well-known segmentation methods and proposed an improved robust approach based on a set of adjustments to these methods on the MIAS database [5] to assist CAD in its preprocessing step. This paper has studied and implemented two different types of popular algorithms, namely Level Set and Region Growing. Many novel findings and implementations of these algorithms have been discussed due to the direct comparison that has been done between these two algorithms, and their findings are compared along with suggestions on what type of mammogram each algorithm performs better or worse. Moreover, an Improved Region Growing algorithm (IRG) was proposed, which introduced the concept of dynamic thresholding. The threshold values were used based on testing to find what number of iterations would make the segmentation jump out in the breast tissue. This modification improved the accuracy of the algorithm; furthermore, it also provided insights on how this method can be used and further improved using machine learning techniques to predict the best threshold value for each image. This research also includes a web-based application that is based on the IRG algorithm, which can be used by anyone to access the API (Application Programming Interface), which can be integrated into classification systems as a preprocessing step. This facility will reduce the time required for future research projects focusing on classification since they can directly call the API and get the segmented images instead of writing segmentation algorithms themselves.

This paper is organized as follows. Section II contains the literature review of the proposed method to segment the pectoral muscle regions. This is followed by Section III, which contains the methodology, including design and implementation. Section IV highlights the results and discussion and finally, the paper is concluded in Section V.

II. Literature Review

Accurately removing breast pectoral muscle is challenging for researchers due to the difference in size, shape, and position of the breast in a mammogram [6]. As indicated by [7], there are four main types of problematic images within the most popular datasets used by researchers: MIAS [5] and DDSM [8]. These include images with tape artifacts, images with axillary fold, images with the invisible contour of the pectoral muscle, the level, and images with no pectoral muscle at all. Due to these issues, many different algorithms have been proposed to segment the pectoral muscle effectively from the mammogram. These classifications are visualized in Fig. 1.

The following section presents previous works proposed by other researchers for pectoral muscle segmentation. To make it easier for readers to follow this section, papers are categorized based on whether they belong to Intensity-based, Edge-based, or Deep learning-based algorithms.

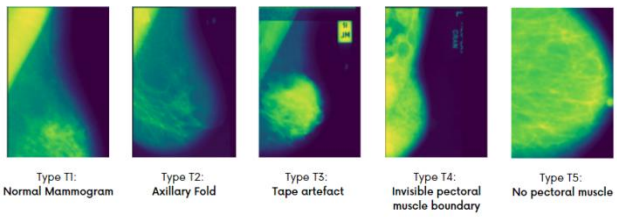


Fig. 1. Classification of the type of mammograms from the MIAS dataset

A. Intensity-based Algorithms

Intensity-based algorithms make use of the difference in the intensity in the mammography image. The pectoral muscle has a high intensity compared to the breast tissue. These techniques use these properties to segment the pectoral muscle out of the mammogram, assuming the pectoral muscle will have a higher intensity [9]. Such techniques include region growing, thresholding, and watershed.

Watershed algorithms are derived from mathematical morphology that segments the images into homogenous regions first introduced in 1978 [10]. Region growing uses the difference in the intensity of the pixels within an image. Region growing uses a starting seed value; from that, adjacent pixels are added depending on the homogeneity criteria assigned. Pixels are continually compared to the ones within the region and added until all adjacent pixels are too dissimilar [11].

Vikhe and Thool [12] used a thresholding-based segmentation technique using contrast enhancement and got an acceptable rate of 96.56%, verified by a certified radiologist. However, the result of this technique was affected by images with invisible contours between the pectoral muscle and the breast tissue.

Taifi et al. [13] used a watershed transformation technique to extract the pectoral muscle from the mammogram. The result was promising, with 90-99% accuracy and 86-99% for precision. There were, however, instances where this algorithm over-segmented the mammogram.

Gómez et al. [14] proposed a region-growing method with seed and threshold methods. The initial seed value used was (10,10). This algorithm's novelty is from its method of calculating the image threshold. The result was 91.92% for the mini-MIAS dataset. As with other intensity-based methods, this method struggles with the invisible contour between the pectoral muscle and the breast tissue.

Hazarika and Mahanta [15] proposed another region-growing-based segmentation method to remove pectoral muscle in mediolateral oblique view mammograms. This method uses a triangular region to estimate the area of the pectoral muscle, after which the region-growing algorithm is applied with automatic starting seed selection. Later, the segmented region is refined to increase the accuracy of the output. This method had an acceptable rate of 86.67% based on 150 images of the mini-MIAS dataset; the accuracy rate is sufficiently accurate.

The threshold-based segmentation methods, including the region growing, have performed relatively well; however, there are a few limitations attached to this method. For example, these methods over-segment or under-segment the images with unclear pectoral muscle and breast tissue boundaries, and they do not consist of any spatial information of the image [9].

B. Edge-based Algorithms

Edge-based segmentation methods use changes in the brightness of the images to identify different regions and segment out the pectoral muscle based on the sharp differences in the brightness of the breast tissue. Straight-line modelling of the pectoral muscle boundaries is used in this method to segment the breast muscle out of the mammography image.

Rampun et al. [4] used canny edge detection and contour growing techniques to segment the image. However, this method fails when the canny method cannot detect any edge between the regions and an invalid selection of the initial contour is made. Moreover, this method also overestimates the pectoral muscle boundary when artifacts are present in the image.

Level Set is also an edge detection technique introduced in 1988 by Osher and Sethian, which represents curves or surfaces as a level set of a higher dimensional hyper-surface [16]. The level set is suited to handle problems in which the evolving interfaces can develop sharp corners and cusps and change in topological features, topology, and images with a relatively high level of noise.

Li et al. [17] proposed a new level of set-based methodology on numerous medical imaging photographs. Comparing their results to a smooth model, the researchers concluded that their proposed method outperforms the smooth model in accuracy, efficiency, and robustness.

Zhou et al. [18] used a correntropy-based level set to segment the pectoral muscle in the mammogram and concluded that this method is considerably less time-consuming and complex and gives excellent results compared to the other state-of-the-art methods.

Anitha and Peter [19] proposed a Kernel-Based Fuzzy Level Set (KFLS) to preprocess the image and to segment the image into several clusters based on the breast structure. The results based on this method showed a high percentage of sensitivity and accuracy of 93.32 and 94.31, respectively.

The line/edge detection methods have advantages when handling noisy data, as they can easily be adjusted to noise. However, it requires large storage and more computational requirements.

C. Deep Learning-based Algorithms

Due to the variability of the shape, size, and type of breast in the mammograms, there has been a lot of interest in utilizing deep learning-based algorithms to segment the pectoral muscle.

Wang et al. [20] first applied some image normalization to the input image and then trained the model using 2000 digital images with a dice-similarity coefficient of 0.8879 based on 825 of those images. Ali et al. [21] are also using this technique with Gaussian and median filters. The accuracy rate achieved was around 97%. A similar technique is applied by Kim et al. [22]. They trained their model on the 322 images of the min-MIAS dataset with an accuracy rate of 95.88% accuracy.

Rampun et al. [7] used a Convolutional Neural Network (CNN) inspired by a holistically nested edge detection network to automatically model the characteristics of the pectoral muscle.

A hybrid breast cancer classification technique was proposed by [23]. Three deep learning models, including ResNet50, Inception-V3, and AlexNet, were used for feature extraction. The Term Variance feature selection is used to select the best features. The multiclass support vector machine was applied to classify the MIAS dataset.

While deep learning models report a higher accuracy rate, they require large datasets to be trained and a very high computational power for the neural networks to get trained. A large dataset for mammograms is not readily available, so this is a hurdle for deep learning techniques at this stage.

D. Further Readings and Research Potential

Wavelet-based algorithms are also used for muscle segmentation and are not discussed in the literature review but can be helpful for some readers. This technique uses a short-term Fourier transform. The spatial frequency of the image is identified by the wavelets. Ferrari et al. [24] proposed a wavelet technique; their method was tested on 84 MLO mammograms from the min-MIAS dataset with 0.58% false positive and 5.77% false-negative percentages, which indicates a good percentage of accurate segmentations.

The Wavelet technique has an advantage in the fact that all information required to segment the image is given by the wavelet decomposition. However, this technique does result in some lost information about the image in the process. There can be further improvements in the mentioned techniques for example, the choice of the initial seed in a region growing currently does not consider the shape and size of the pectoral muscle, which can be used to increase the accuracy rate further.

III. Methodology

This section covers the methodology of the major components in the study, namely the preprocessing steps, the Level Set and Region Growing algorithms, and some details about the development of the Web Application. The overall structure of the proposed method is shown in Fig. 2. The process of removing the pectoral muscle from a mammogram is broken down into three steps – preprocessing, pectoral muscle segmentation, and performance evaluation.

A. Preprocessing

Before a mammogram is fed into a segmentation algorithm, it will be preprocessed. Details of all the preprocessing steps can be found here. The first two preprocessing steps are the removal of empty space (bar) and automatic left flipping. This needs to be done because both the Level Set and Region Growing algorithms require the pectoral muscle region to be on the top-left corner of the mammogram. Otherwise, the pectoral muscle region cannot be detected. Fig. 3 and Fig. 4 show the

outcomes of the Level Set algorithm when the bar removal and automatic flipping preprocessing steps were not applied.

Another preprocessing step is the application of contrast on the mammogram. This increases the brightness differences between the pixels in the breast region and the pectoral muscle region, allowing both the segmentation algorithms to detect the pectoral muscle boundary more easily. Fig. 5 shows how the Level Set algorithm fails to detect the boundary of the pectoral muscle region (overshooting) when contrast is not applied on a mammogram.

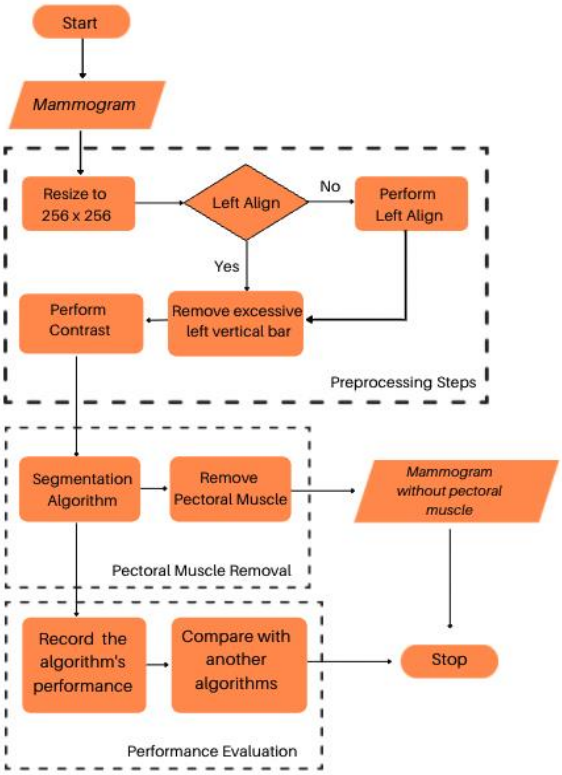


Fig. 2. The overall structure of the proposed method.

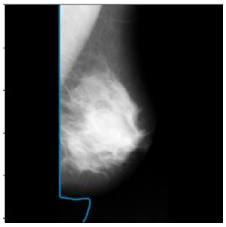


Fig. 3. Boundary found by the Level Set algorithm (denoted by the blue line) if the empty space on the left is not removed.

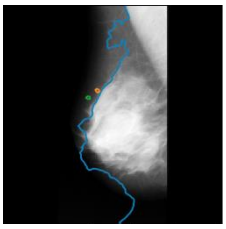


Fig. 4. Boundary found by the Level Set algorithm (denoted by the blue line) if the image is not automatically flipped to the left.

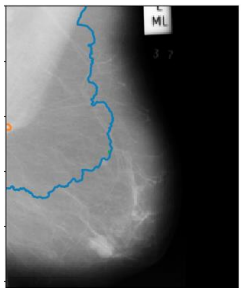


Fig. 5. Boundary found by the Level Set algorithm (denoted by the blue line) if contrast is not applied to the mammogram.

B. Segmentation

1) Level set algorithm: The Level Set image segmentation algorithm is widely used in the field of image processing and is now frequently employed in image segmentation. It is one of the active contour models that can handle complex topologies and capture boundaries and is specially used in images with intensity inhomogeneity, such as medical images [25] [26]. Researchers often prefer this method because it is flexible, easy to understand, and easy to employ. Level set models are based on the evolution of the zero level [16]. Let's assume that the closed interface $t:[0,\infty]\rightarrow Rn$ is an initial circle in 2-D/3D space when $t = 0$. In order to complete the evolution, a zero-level set function (φ) was constructed. In the general case, let 0 be a closed, disjoint, $(N - 1)$ dimensional initial hypersurface. Hence, φ can be defined by

$$xt,t=\mp r \quad (1)$$

where, r is the distance from x to the hyper-surface 0. φ is positive if x is outside of 0 and negative if it is inside. Since motion can be seen as:

$$xt,t=0 \quad (2)$$

Find the partial derivative of both sides of formula Eq. 2, by the chain rule, can get:

$$\varphi t+\Delta.xt \quad (3)$$

The speed of evolution is one of the crucial parameters; hence, it is defined as below:

$$V=xt.n \quad (4)$$

where, $n=\Delta\varphi\Delta\varphi$ is a normal vector or mean curvature, the final curve evolution equation is

$$\varphi t=V.\Delta\varphi \quad (5)$$

Level set methods can usually be divided into two main categories, which are edged-based and region-based.

C. Improved Region Growing (IRG) Algorithm

The Region-Growing Algorithm is an intensity-based segmentation algorithm. The algorithm consists of three main steps. First, a particular pixel is assigned as the initial seed, and a fixed threshold value is chosen. Second, the seed is then compared with its adjacent pixels at each iteration. If the difference between the initial seed and the adjacent pixel is within the threshold value, the adjacent pixel will merge with

the initial seed to form a region. Lastly, the algorithm will terminate when all the adjacent pixels are too dissimilar.

An additional step was implemented to improve the accuracy of the original algorithm in removing the pectoral muscle region. This involves checking if similar adjacent pixels are still found after running the algorithm for a predefined maximum number of iterations. If so, the algorithm will restart with a lower threshold value, in other words, stricter criteria for grouping similar pixels. This approach prevented the algorithm from removing the breast region; see more in the Segmentation Section. The flowchart of the Improved Region Growing (IRG) algorithm is shown in Fig. 6.

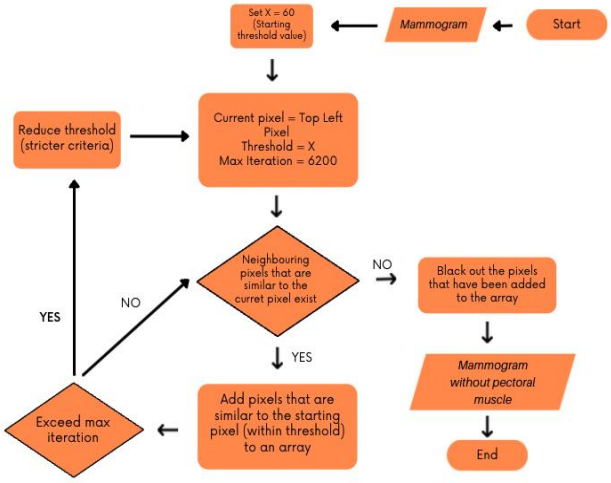


Fig. 6. Flow chart of the Improved Region Growing (IRG) algorithm to remove the pectoral muscle region from a mammogram

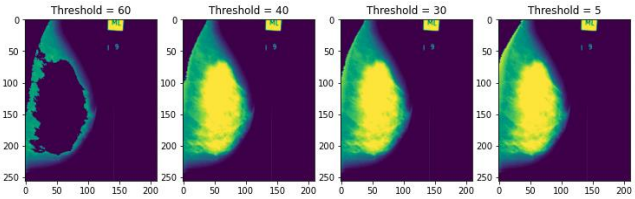


Fig. 7. The result of the improved region-growing algorithm varies with different thresholds used.

In the IRG algorithm, the threshold value is an important parameter as it determines the criteria, whether strict or loose, when grouping the pixels in the mammogram. Different thresholds would result in different results, as shown in Fig. 7. Segmentation for the mammogram (mdb003.pgm) works best when the threshold is set to 40. In this research, it was observed that different images require a different threshold to run well. Fig. 8 shows the various optimal thresholds needed by the different mammograms. Therefore, fixing one predefined threshold value before the algorithm runs would result in some mammograms not being segmented well. To solve this problem, the algorithm was modified by adding an additional step. The step involves restarting the algorithm with a lower threshold value (stricter criteria) once the algorithm has run for a predefined number of iterations. The logic behind this change is that if the algorithm runs for more than the predefined number of iterations, this means that the boundary found by the algorithm has crossed the pectoral muscle boundary, resulting in overshooting. By doing this, all mammograms will be segmented using the optimal threshold value.

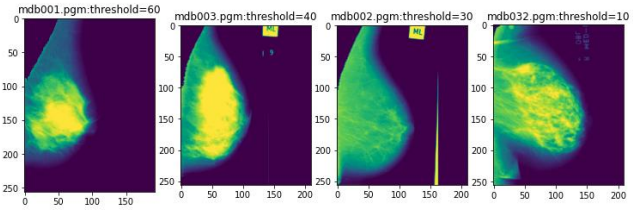


Fig. 8. Different performance results based on a variety of threshold values.

IV. Results and Discussion

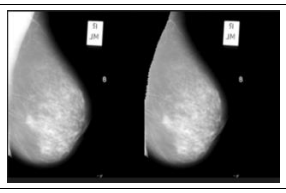
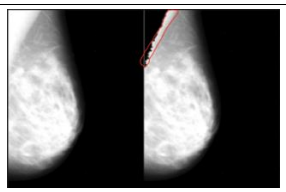
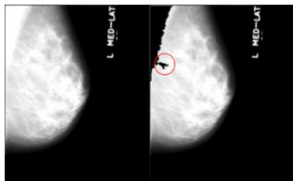
A. Evaluation of results by professional radiologists

To make sure the algorithm has removed the pectoral muscle from the mammogram completely and correctly, especially mammograms in which the pectoral muscle region is not easily seen, two radiologists were invited to evaluate the results. Both the segmentation algorithms are run with all 322 mammograms in the MIAS dataset. This section will first discuss the evaluation criteria and how the images in the MIAS dataset are classified into different categories. Then, the performance of the Level Set and Region Growing algorithm will be discussed.

B. Performance Evaluation

Each segmented mammogram can either be evaluated as acceptable or unacceptable. Table I describes the evaluation criteria. Pectoral removal is acceptable if the segmentation algorithm removes the entire pectoral muscle region from the mammography images. Pectoral removal is not acceptable if the algorithm does not remove the entire pectoral muscle region (see second row of Table I) or removes the entire muscle and some breast regions that may contain a tumor (see third row of Table I).

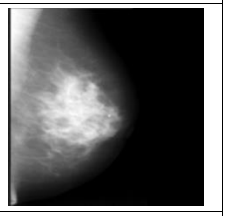
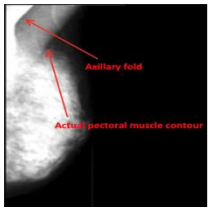
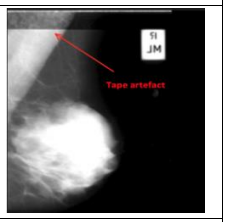
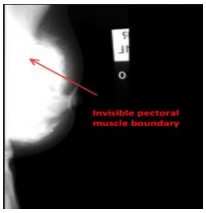
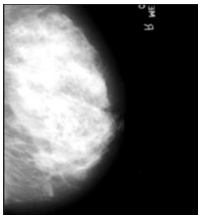
TABLE I. Evaluation Criteria

Evaluation	***Mammogram***	***Criteria***
Acceptable		The segmentation algorithm removes the pectoral muscle region from the mammogram completely.
Unacceptable		The algorithm does not remove the entire pectoral muscle region. Part of it has not been removed (red boundary)
Unacceptable		The algorithm removes the entire pectoral muscle region, and removes some of the breast regions that might contains tumor (red circle).

C. Classification of the MIAS dataset

To better understand the performance of each segmentation algorithm, all 322 mammograms in the MIAS dataset are classified into the five types as outlined in Table II. Different types of mammograms, along with their total numbers, brief descriptions, and examples, are presented in Table II.

TABLE II. CLASSIFICATION OF INPUT IMAGES OF THE MIAS DATASET

Type Code	*Count*	*Description & Image No.*	*Mammogram*
T1	210	Normal Mammogram. mdb187.pgm	
T2	84	Mammograms have an inner boundary (axillary fold) within the pectoral muscle region which results in high false positives when detecting the pectoral muscle boundary [7]. mdb039.pgm	
T3	14	Mammograms that contain tape artefact. mdb002.pgm	
T4	8	Mammograms in which the contour of the pectoral muscle region is invisible. mdb288.pgm	
T5	6	Mammograms in which the pectoral muscle region is not visible. mdb236.pgm	

D. Performance of the Segmentation Algorithms

As mentioned earlier, a professional radiologist was invited to evaluate the mammograms segmented by the Level Set algorithm and two professional radiologists to evaluate the mammograms segmented by the Improved Region Growing algorithm (IRG). The results of each algorithm were also evaluated by the development team. The results are presented in Tables III, IV, V, VI and VII. For each category of mammogram (as described in Table II), the tables show the number of mammograms classified into that category (count), followed by the number of segmented images classified by the evaluator as acceptable or unacceptable respectively. Finally, the acceptance rate for each category is recorded, and an overall acceptance rate is reported at the bottom.

1) Level set: Tables III and IV show the Level Set (LS) results of the radiologist and the team's evaluation respectively.

2) Improved Region Growing (IRG): For the results of the Improved Region Growing (IRG) algorithm, Tables V and VI show the results of the two radiologist's evaluations, and table VII shows the results of the team's evaluation.

TABLE III. EVALUATION OF LS METHOD BY RADIOLOGIST

Type Code	*Count*	*Acceptable*	*Unacceptable*	*Success Rate*
T1	210	113	97	54%
T2	84	6	78	7%
T3	14	5	9	36%
T4	8	2	6	25%
T5	6	3	3	50%
Overall Acceptance Rate				40%

TABLE IV. EVALUATION OF LS METHOD BY THE TEAM

Type Code	*Count*	*Acceptable*	*Unacceptable*	*Success Rate*
T1	210	115	95	55%
T2	84	3	81	4%
T3	14	5	9	36%
T4	8	2	6	25%
T5	6	3	3	50%
Overall Acceptance Rate				39.75%

TABLE V. EVALUATION OF THE IRG METHOD BY RADIOLOGIST 1

Type Code	*Count*	*Acceptable*	*Unacceptable*	*Success Rate*
T1	210	168	42	80%
T2	84	50	34	60%
T3	14	5	9	36%
T4	8	2	6	25%
T5	6	5	1	83%
Overall Acceptance Rate				71.42%

TABLE VI. EVALUATION OF IRG METHOD BY RADIOLOGIST 2

Type Code	*Count*	*Acceptable*	*Unacceptable*	*Success Rate*
T1	210	167	43	80%
T2	84	52	32	62%
T3	14	2	12	14%
T4	8	2	6	25%
T5	6	4	2	67%
Overall Acceptance Rate				70.40%

TABLE VII. EVALUATION OF IRG METHOD BY THE TEAM

Type Code	*Count*	*Acceptable*	*Unacceptable*	*Success Rate*
T1	210	161	49	77%
T2	84	50	34	60%
T3	14	3	11	21%
T4	8	0	8	0%
T5	6	3	3	50%
Overall Acceptance Rate				67.30%

E. Analysis

The IRG method performs better than the LS method in terms of the following aspects:

1) Removal of the pectoral muscle region from normal mammograms: The IRG method can generally remove the pectoral muscle more accurately than the LS method. Following the performance evaluation methodology, this means that the IRG method is more likely to remove the entire pectoral muscle region and is less likely to leave out part of the pectoral muscle. Fig. 9 and Fig. 10 show the performance differences between these two algorithms.

2) Removal of the pectoral muscle region from mammograms that contain axillary fold: The LS method performs poorly because the algorithm cannot properly segment a mammogram containing an axillary fold (Type T2). The axillary fold causes the algorithm to detect a false pectoral muscle contour, which results in incomplete pectoral muscle removal. In the radiologist's and team's evaluation, the algorithm only achieves a 7% and 4% acceptance rate, respectively. On the other hand, the IRG method is less likely than the LS method to detect a false contour. It achieves a higher acceptance rate when segmenting a Type T2 mammogram.

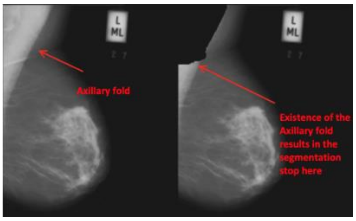


Fig. 9. An example of how the LS method fails to remove the entire pectoral muscle region.

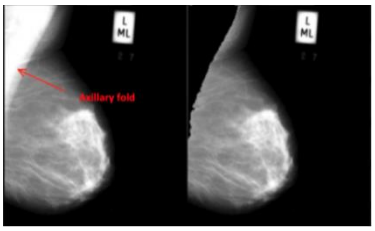


Fig. 10. An example of how the IRG method successfully removes the entire pectoral muscle region.

Fig. 11 compares the acceptable rate between the LS and IRG methods when segmenting the different types of mammograms.

Another observation that can be made is that both algorithms do not work well (acceptance rate less than 50%) with mammograms that contain tape artifacts and those that have invisible pectoral muscle contours (Type T3 and T5). This is because both segmentation algorithms are designed to identify a region based on the boundary found. First, the algorithm does not work well with type T3 mammograms because the boundary that is identified is the artifact boundary. Next, the type T5 mammogram will also not be segmented well by the algorithms as its boundary is not visible.

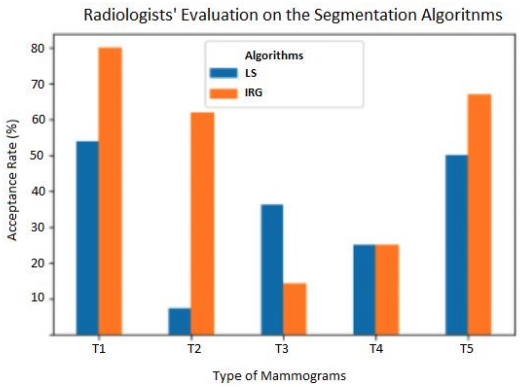


Fig. 11. Group bar chart comparing the result between the LS method and IRG method.

V. Conclusion

While the automatic pectoral muscle removal system is a complete system, it contains a few limitations on the preprocessing steps, segmentation algorithms, and web application. The result tabulation shows that none of the algorithms works well with mammograms containing the tape artifact (type T3). This shows that the existing preprocessing steps are not enough to ensure that both segmentation algorithms run well. The IRG method performs significantly better than the LS method. However, both algorithms have limitations. First, both algorithms do not work well with type T2 mammograms. For instance, the IRG method achieves an 80% acceptance rate on the normal mammogram (Radiologists' evaluation for T1), yet it only achieves a 62% and 60% acceptance rate on type T2 mammograms (Radiologists' evaluation for T2). Considering the number of T2 mammograms in the MIAS dataset (84 out of 322), it can be said that this type of mammogram is very common. Therefore, the low acceptance rate to segment this type of mammogram is a big limitation of the algorithm. Second, both algorithms fail when the mammogram contains an invisible pectoral muscle contour (type T4). While the LS method achieves a 25% acceptance rate, the IRG method performs poorly when segmenting the type T4 mammograms. Next, it is noticed that both segmentation algorithms cannot remove the pectoral muscle completely from a mammogram if it contains an axillary fold. Fig. 12 shows a type T2 mammogram in which the pectoral muscle is not removed completely.

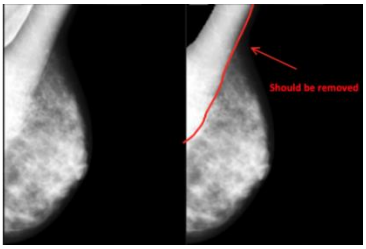


Fig. 12. A mammogram in which the pectoral muscle region is not removed completely.

One possible improvement that could be made to the IRG method to solve this problem is to place a triangle over the mammogram image to estimate the pectoral muscle region. The triangle can be drawn using the difference between the

intensity of the pectoral muscle region and the breast region, along with some mathematical formulas [15]. Having estimated the pectoral muscle boundary, the region growing algorithm can be modified to dynamically adjust the threshold value based on where the current pixel is. For example, a higher threshold value can be used when the current pixel lies within the estimated boundary. In other words, a looser criterion is used for grouping similar pixels if the current pixel is within the boundary. This will allow the algorithm to completely avoid the false pectoral boundary and group the pectoral muscle region. Fig. 13 shows how the algorithm could be improved by drawing an additional triangle before the algorithm runs.

Fig. 13. An example of how a triangle can be placed on top part of the mammogram.

The study has a constraint of not validating the algorithm's performance with other datasets. Additionally, utilizing deep networks can enhance the accuracy of the proposed technique by extracting improved features from segmented images. In the future, it is recommended to automatically crop the tumor region and utilize pre-trained networks.

ACKNOWLEDGMENT

The Initial part of this work was carried out when the first author was affiliated with Monash University Malaysia.

REFERENCES

[1] Mutebi, M., Anderson, B. O., Duggan, C., Adebamowo, C., Agarwal, G., Ali, Z., ... & Eniu, A. (2020). Breast cancer treatment: A phased approach to implementation. Cancer, 126, 2365-2378.

[2] DeSantis, C. E., Bray, F., Ferlay, J., Lortet-Tieulent, J., Anderson, B. O., & Jemal, A. (2015). International variation in female breast cancer incidence and mortality rates. Cancer epidemiology, biomarkers & prevention, 24(10), 1495-1506.

[3] Raghavendra, U., Acharya, U. R., Fujita, H., Gudigar, A., Tan, J. H., & Chokkadi, S. (2016). Application of Gabor wavelet and Locality Sensitive Discriminant Analysis for automated identification of breast cancer using digitized mammogram images. Applied Soft Computing, 46, 151-161.

[4] Rampun, A., Morrow, P. J., Scotney, B. W., & Winder, J. (2017). Fully automated breast boundary and pectoral muscle segmentation in mammograms. Artificial intelligence in medicine, 79, 28-41.

[5] Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., ... & Savage, J. (2015). Mammographic image analysis society (mias) database v1. 21.

[6] Rezaei, Z. (2021). A review on image-based approaches for breast cancer detection, segmentation, and classification. Expert Systems with Applications, 182, 115204.

[7] Rampun, A., López-Linares, K., Morrow, P. J., Scotney, B. W., Wang, H., Ocaña, I. G., ... & Macía, I. (2019). Breast pectoral muscle segmentation in mammograms using a modified holistically-nested edge detection network. Medical image analysis, 57, 1-17.

[8] Mustra, M., & Grgic, M. (2013). Robust automatic breast and pectoral muscle segmentation from scanned mammograms. Signal processing, 93(10), 2817-2827.

[9] Pawar, S. D., Sharma, K. K., Sapate, S. G., & Yadav, G. Y. (2021). Segmentation of pectoral muscle from digital mammograms with depth-first search algorithm towards breast density classification. Biocybernetics and Biomedical Engineering, 41(3), 1224-1241.

[10] Digabel, H.L., C, Iterative Algorithms. Proceedings of the 2nd European Symposium Quantitative Analysis of Microstructures in Material Science, Biology and Medicine, 1978. 66(1): p. 240-256.

[11] Gonzales, R.C. and P. Wintz, Digital image processing. 1987: Addison-Wesley Longman Publishing Co., Inc.

[12] Vikhe, P. S., & Thool, V. R. (2016). Intensity based automatic boundary identification of pectoral muscle in mammograms. Procedia Computer Science, 79, 262-269.

[13] Taifi, K., Ahdid, R., Fakir, M., Elbalaoui, A., Safi, S., & Taifi, N. (2017, May). Automatic breast pectoral muscle segmentation on digital mammograms using morphological watersheds. In 2017 14th International Conference on Computer Graphics, Imaging and Visualization (pp. 126-131). IEEE.

[14] Gómez, K. A. H., Echeverry-Correa, J. D., & Gutiérrez, Á. Á. O. (2021). Automatic pectoral muscle removal and microcalcification localization in digital mammograms. Healthcare Informatics Research, 27(3), 222-230.

[15] Hazarika, M., & Mahanta, L. B. (2018). A novel region growing based method to remove pectoral muscle from MLO mammogram images. In Advances in Electronics, Communication and Computing: ETAEERE-2016 (pp. 307-316). Springer Singapore.

[16] Osher, S., & Sethian, J. A. (1988). Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. Journal of computational physics, 79(1), 12-49.

[17] Li, C., Huang, R., Ding, Z., Gatenby, J. C., Metaxas, D. N., & Gore, J. C. (2011). A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI. IEEE transactions on image processing, 20(7), 2007-2016.

[18] Zhou, S., Wang, J., Zhang, M., Cai, Q., & Gong, Y. (2017). Correntropy-based level set method for medical image segmentation and bias correction. Neurocomputing, 234, 216-229.

[19] Anitha, J., & Peter, J. D. (2015). Mass segmentation in mammograms using a kernel-based fuzzy level set method. International Journal of Biomedical Engineering and Technology, 19(2), 133-153.

[20] Wang, K., Khan, N., Chan, A., Dunne, J., & Highnam, R. (2019, November). Deep learning for breast region and pectoral muscle segmentation in digital mammography. In Pacific-Rim Symposium on Image and Video Technology (pp. 78-91). Cham: Springer International Publishing.

[21] Ali, M. J., Raza, B., Shahid, A. R., Mahmood, F., Yousuf, M. A., Dar, A. H., & Iqbal, U. (2020). Enhancing breast pectoral muscle segmentation performance by using skip connections in fully convolutional network. International Journal of Imaging Systems and Technology, 30(4), 1108-1118.

[22] Kim, Y. J., Yoo, E. Y., & Kim, K. G. (2021). Deep learning based pectoral muscle segmentation on Mammographic Image Analysis Society (MIAS) mammograms. Precision and Future Medicine, 5(2), 77-82.

[23] Elkorany, A. S., & Elsharkawy, Z. F. (2023). Efficient breast cancer mammograms diagnosis using three deep neural networks and term variance. Scientific Reports, 13(1), 2663.

[24] Ferrari, R. J., Rangayyan, R. M., Desautels, J. L., Borges, R. A., & Frere, A. F. (2004). Automatic identification of the pectoral muscle in mammograms. IEEE transactions on medical imaging, 23(2), 232-245.

[25] Wang, Z., Ma, B., & Zhu, Y. (2021). Review of level set in image segmentation. Archives of Computational Methods in Engineering, 28, 2429-2446.

[26] Rezaei, Z., Selamat, A., Taki, A., Rahim, M. S. M., & Abdul Kadir, M. R. (2019). Systematic mapping study on diagnosis of vulnerable plaque. Multimedia Tools and Applications, 78, 21695-21730.

Applying Machine Learning Models to Electronic Health Records for Chronic Disease Diagnosis in Kuwait

Talal M. Alenezi[1]*, Taiseer H. Sulaiman[2], Amr M. AbdelAziz[3]
Faculty of Computers and Information, Assiut University, Assiut, Egypt[1]
Information Science Dept.-Faculty of Computers and Information, Assiut University, Assiut, Egypt[2]
Faculty of Computers and AI, Beni-Suef University, Beni-Suef, Egypt[3]

***Abstract*—The leading cause of death nowadays is chronic disease. As a result, personal wellbeing has received a considerable boost as a healthcare preventative strategy. A notable development in data-driven healthcare technology is the creation of a prediction model for chronic diseases. In this situation, computational intelligence is used to analyze electronic health data to provide clinicians with knowledge that will help them make more informed decisions about prognoses or therapies. In this study, various classification algorithms have been implemented namely, Decision Tree, K-Nearest Neighbors, Logistic Regression, Multilayer Perceptron, Naïve Bayes, Random Forest, and Support Vector Machines to examine the medical records of patients in Kuwait who had chronic conditions. For predicting diabetes, the support vector machines classifier was the best classifier for predicting diabetes and kidney chronic diseases. For diabetes, it achieved 88.5% accuracy, and 93.6% f1-score, while for kidney; it achieved 94.9% and 92.6% accuracy and f1-score respectively. For predicting heart disease, MLP was the best and achieved 84.7%, and 87.8% accuracy and f1-score respectively.**

***Keywords*—Chronic diseases; Electronic Health Records (EHR); machine learning; classification**

I. Introduction

Healthcare systems and governments from all around the world are very interested in finding ways to improve healthcare outcomes while lowering costs. Accurate patient risk assessments from both patients and clinicians are essential to achieving this improvement. The usage of electronic health records (EHRs) for digitally recording patient healthcare contacts has increased significantly in tandem with this focus on improving outcomes (Hsiao et al., 2014). As a result, disease risk prediction technologies that may use EHR data to evaluate patient risk have drawn a lot of interest [1].

There is now a lot of interest in applying machine learning techniques to create disease risk prediction tools from EHR data since machine learning algorithms have experienced significant growth in use. “What is the distribution of accuracy with which all EHR-coded medical occurrences may be anticipated?” is a logical place to start. We provide an answer to this query for coded diagnoses since it has not previously been addressed [2].

By using computational intelligence techniques, this study was able to extract pertinent information from electronic medical records that might be used by doctors to deliver effective care. Electronic health records (EHRs) are instantaneous, patient-centered records with secure access for authorized individuals. EHRs are distinguished by the capability of authorized clinicians to generate and manage health data in digital format, which may be shared with other doctors across various healthcare organizations. The Arabic region faces significant challenges due to chronic ailments. Chronic disease-related mortality has been linked to a sizable number of deaths, according to Kuwait's Ministry of Health [3]. Their statistics show that 41% of fatalities were attributable to heart disease, 15% to cancer, 3% to respiratory problems, and 3% to diabetes. Additionally, according to Kuwait's national plan report on chronic diseases for the years 2017 to 2025, cardiovascular disorders are the main cause of mortality there [4].

The goal of this study is to create a predictive analytics model that can predict the onset of three common chronic diseases diabetes, heart disease, and kidney disease early on using electronic health records that contain important information about patients from their hospital visits. A key component of prevention is prediction, which enables healthcare organizations to give top priority to patients with the greatest needs and make the best use of their scarce resources.

To attain the research objectives, the authors conducted a study that involved defining the methodology, study population, study sample, study tools, verifying the validity and reliability of the tools, and utilizing statistical treatment in analyzing the results. The study population comprised all healthcare stakeholders in Kuwait hospitals, and the study sample consisted of (30) medical sector workers, from which a random sample was selected. The study sample's frequencies and percentages were computed, and they are represented in the basic data that includes Gender: (43% males and 56.7% females), Age: (26.7% less than a year, 43.3% between 30 to 40 years, 16.7% between 40 to 50 years, and 13.3% over 50 years), Position: (16.7% manager, 40% attending physician, and 43.3% technician), Number of years of experience (33.3% less than three years, 6.7% between three years to less than six years, and 60% more than six years), and Computer skills: (3.3% weak, 36.7% medium, 60% good). The paper evaluates physicians' perceptions on the current management of chronic diseases in primary healthcare centers using electronic health

record systems, shedding light on the practical implications of using EHR for disease management.

From Table V in Appendix A, it can be observed that the general average for the first dimension, technological infrastructure environment, had a response rate (High), an arithmetic means of (2.66), a standard deviation of (.409), and phrase No. (27) appeared in the first order. (The need to employ e-health solutions to improve the quality of the healthcare sector.) Phrase No. 22 came in second place (There is a need to establish a well-designed primary health care network) with an arithmetic mean (2.87), a standard deviation (.346), and a response rate (High), and it was followed in the last position by phrase No. 23 (There is a need for the establishment of a well-designed primary health care network). (25) (Health informatics is already implemented in the State of Kuwait.) with an arithmetic mean (2.43), and a standard deviation (.679) with a response score (High), and the rest of the expressions came with a response score (High), The standard deviations for the expressions in the first dimension varying between (.681 - .305), which are regarded as low values. This implies that the study sample's reactions to these assertions were consistent. The emergence of the technical infrastructure environment can be explained by the fact that there are numerous opportunities for developing information systems in Kuwaiti hospitals, in addition to the possibility of directing a large investment towards developing the state's infrastructure relying on electronic systems in facilitating work in the medical field, beginning with data recording, and progressing to electronic solutions.

Table VI in Appendix B shows that the overall average for the second dimension, administration, and organization, had a (High) response rate, an arithmetic mean of (2.62), and a standard deviation of (.473). Phrase 31, (The need for high-quality services in the public and private sectors of healthcare.) came in first place with an arithmetic mean (2.93), a standard deviation (.254), and a response score (High), and phrase 32 came in second place (The need for a national health information management strategy to guide public health policies.) Assuming an arithmetic mean (2.27), The rest of the expressions received a response score (High) and a standard deviation (.828), while the standard deviations for the expressions in the second dimension varying between (.254-.828), indicating homogeneity of opinions of the study sample towards these statements. The need to establish a national plan for health information management to lead public health policies might explain the appearance of the administration and organizations with a high level of responsiveness. It also mentioned the need to enhance the quality of services in the field of health care in general, by focusing on health insurance and trying to reduce health expenditures. The occurrence of phrase No. (34) in the last order can be explained by the fact that, despite the availability of an integrated framework for information and technology, this framework is not primarily relied upon in the field of medical care in Kuwait.

The previous study's recommendations emphasized the need of using electronic medical records to register patients' health information rather than paper records. Furthermore, the study emphasized the need for predictive analytics models that can anticipate the emergence of chronic illnesses at an early stage utilizing electronic health information. The study concentrated on diabetes, heart disease, and chronic kidney disease. Heart disease is a major cause of death in the United States; it kills 610,000 of people each year, accounting for one in every four fatalities. [5]. Diabetes is recognized as a chronic illness with the world's fastest growing rate at the same time [6]. Diabetes affects 415 million people worldwide, accounting for 12% of total medical spending ($673 billion) [6]. Diabetes affected 29.1 million Americans in 2012, accounting for around 9.3% of the population [6]. About 700 million people worldwide are affected with Chronic Kidney Disease (CKD) each year, resulting in the deaths of almost 1.2 million people [7]. Cardiovascular diseases such as myocardial infarction, stroke, and heart failure are more prevalent in chronic renal disease patients [7]. Patients who have both illnesses have a worse prognosis [8].

The following are the primary contributions of the article:

- Using electronic health records instead of paper records.
- Using accessible datasets from patients' medical records, machine learning techniques are used to predict the existence of chronic illnesses.
- Examining medical records of all patients to ensure proper diagnosis of chronic disorders.
- Identifying new patients with comparable symptoms and illness development phases based on physician supervision and medical record analysis for a specific type of chronic disease.

The second half of the paper will go into similar research, while Section III will examine the datasets used in the study. The fourth part will offer a full description of the suggested technique. Following that, Section V will show the test findings and assess the proposed strategy. Finally, Section VI will give findings and recommendations for further research.

II. Related Work

Yaser et al. [9], has presented a research paper. The fundamental contribution of this work is the development of a medical recommendation system that employs the closest neighbor’s classification approach and the collaborative filtering methodology. The suggested strategy was evaluated based on two criteria: statistical correctness and efficacy in giving patient counseling. The results showed that the suggested method outperformed earlier approaches in diagnosing patients, which was ascribed to the use of the "close neighbors" strategy. The suggested technique exhibited great accuracy in patient diagnosis and gave suitable therapy recommendations. However, due to the restrictions of patient data privacy, getting exact hospital data remains a substantial barrier for the suggested strategy.

Chicco et al. [10], presented another paper. A dataset of 491 patients from the United Arab Emirates was analysed to find independent risk variables for CKD at stages 3-5. The authors utilised two strategies, one based on classic univariate biostatistics testing and the other on machine learning. The biostatistics tests revealed that 68.42% of clinical parameters were significant but lacked accuracy. As a result, the authors

used Random Forests to determine feature ranking. The study proved that computational intelligence could predict significant CKD development with or without time information, and that the most important clinical variables alter when the temporal component is incorporated. The benefit of this study is that it provides a full examination of risk variables connected with CKD at stages 3-5 and applies machine learning techniques to determine the most essential clinical characteristics. However, the scientists did not examine the therapeutic significance of the findings, instead focused on developing and strengthening computational intelligence technologies.

Hohman et al. [11], presented The MENDS project which has successfully proved its capacity to satisfy principles and criteria aimed at optimising the project and supporting the CDC's DMI. The project team used statistical approaches to obtain credible prevalence estimates from EHR data that appropriately represent the underlying populations at geographic target levels. The team was able to generate more reliable local, state, and national prevalence estimates by using the breadth of clinical data. The initiative also provided an opportunity for the DOH to expand its EHR-based surveillance beyond federal programme requirements. However, obstacles such as outreach, communication, paperwork, research, and training have hindered the project's timeframe. The team is also comparing prevalence figures from other national chronic illness surveillance systems to determine their validity. The capacity to combine standard public health monitoring techniques with EHR-based surveillance and gain relevant insights from the data is one of the project's advantages. The drawbacks include the technical work and expense necessary to comply with federal rules, as well as the additional outreach, communication, documentation, investigation, or training that is required.

Kumari and Seema [12], used four distinct classifiers to detect diabetes and heart disease: Nave Bayes, Decision Tree, Support Vector Machine (SVM), and Artificial Neural Networks (ANN). The results showed that SVM had the best accuracy rate for heart disease at 95.556%, while Nave Bayes had the highest accuracy rate for diabetes at 73.588%. The benefit of employing these classifiers is that they can help clinicians find successful therapies and best practices, resulting in better and more direct healthcare for patients. However, these classifiers have the disadvantage of requiring a significant quantity of data to reach accuracy and can be time-consuming to train.

Perotte et al. [13], have evaluated the usefulness of employing electronic healthcare data and ICD-10 pain-related diagnostic clusters to self-reported data, which is considered the gold standard, in identifying persons with Musculoskeletal Chronic Pain (MSCP). The study had a 61% response rate, and the findings indicated that the prevalence of MSCP was somewhat lower when detected by electronic health records (14.7% weighted prevalence) than when found with survey questionnaires (25.9% weighted prevalence). The study found a poor level of agreement between the two MSCP categories (kappa agreement statistic = 0.21). When compared to MSCP determined using self-report as the benchmark, using electronic health data exhibited a sensitivity of 30.9%, specificity of 91.0%, and positive predictive value of 54.5%. When age and gender were taken into consideration, patients with MSCP identified through electronic health records or self-report had higher levels of pain-related disability, pain severity, depressive symptoms, and long-term opioid use, compared to individuals with single-site chronic pain identified through the same method. This study demonstrates the use of electronic health records in identifying people with MSCP, potentially leading to a greater detection of people with significant chronic pain in studies that are population-based. The study's weaknesses, however, include a low survey response rate and a lack of agreement between the two MSCP categories.

Hazazi and Wilson [14] aimed to evaluate physicians' perceptions on the current management of Noncommunicable Diseases (NCDs) in Primary Healthcare Centers (PHCs), emphasising on the role of the Electronic Health Record (EHR) system, according to a separate article provided by. The study included semi-structured interviews with 22 Ministry of Health physicians who worked in chronic illness clinics at PHCs. While physicians recognised the benefits of using EHRs, such as improved accuracy in patient documentation and access to patient information, they also identified areas for improvement, such as the lack of a patient portal for patients to access their health information and the system's inability to facilitate multidisciplinary care. In general, doctors viewed the EHR system favorably, although its influence on patient care at chronic illness clinics remained limited.

Areej and Malibari [15] used EO-LWAMCNet, MSSO-ANFIS, T-RNN, and DLMNN algorithms to predict heart and kidney illnesses in real time. On two separate datasets, the suggested model has an accuracy of 93.5% and 94%. The suggested model's merits include its capacity to forecast chronic illnesses including heart and renal disease in real time, its optimal performance, and its short execution time. The suggested model, however, has numerous shortcomings, notably its high cost and inability to distinguish between positive and negative predictions.

III. DATA

In this study, an electronic healthcare record (EHR) dataset comprising information about patients in Kuwaiti hospitals was constructed. Each row in the dataset represents a single patient, and the columns indicate all the patient's attributes/features, as detailed in Table I. This dataset was created for the purpose of predicting diabetes, heart disease, and chronic kidney disease by combining all characteristics from original datasets into a single dataset file for use in training and testing prediction models, in addition to the personal information of all patients during all hospital visits. This EHR dataset will be published soon after the government approval. Diabetes illness data was obtained from the Irvine's Center for Machine Learning and Intelligent Systems, California University [16]. Rather than more than 60,000 individual patients, this study includes clinical data from over 100,000 unique interactions. The information was gathered from roughly 74 million patient visits involving 17 million people [16]. This study's data was gathered over a ten-year period, from 1999 to 2008, and includes many credits that correspond to the seasons of confirmation and release for diabetes patients. The records

include information on laboratory tests and procedures, diagnoses, and drugs given during the hospitalization.

The dataset utilized in this work for heart disease prediction was collected from the UCI Machine Learning Repository of Irvine C.A, University of California and was freely accessible online [17]. The cardiac disease datasets utilized in this study have identical properties and example designs. Only 14 of the 76 raw qualities in these datasets are considered essential, including anticipated property. The Cleveland Clinic Foundation dataset has 303 patient records, whereas the Hungarian Institute of Cardiology dataset contains 294 patient records. This study examined a dataset of 491 individuals from the United Arab Emirates published by Al-Shamsi et al. for chronic renal disease [18]. The data for this study was gathered at Tawam Hospital in Al-Ain (Abu Dhabi, United Arab Emirates) in 2008. The dataset offered 491 patients, 241 women and 250 males, with an average age of 53.2 years. Each patient was given a chart with 13 clinical variables that represented the results of laboratory tests and examinations as well as data from the dataset. The authors used multivariable Cox proportional hazards to identify the risk variables that cause CKD at stages 3-5. However, the analysis did not include a prediction phase, which may have recovered more relevant information or previously unknown patterns in the data.

TABLE I. NAME AND TYPE OF EACH FEATURE OF THE EHR DATASET

Feature Name	Type	Feature Name	Type
Encounter ID	Numeric	Examide	Numeric
Patient number	Nominal	Insulin	Numeric
Race	Nominal	Anemia	Nominal
Gender	Numeric	Creatinine hosphokinase	Nominal
Age	Numeric	Blood pressure	Numeric
Weight	Nominal	Serum creatinine	Numeric
Admission type	Nominal	Smoking	Nominal
Discharge disposition	Numeric	Specific gravity	Nominal
Admission source	Numeric	Albumin	Numeric
Time in hospital	Nominal	Red blood cells	Numeric
Payer code	Nominal	Pus cells	Nominal
Medical specialty	Numeric	Bacteria	Nominal
Number of lab procedures	Numeric	Blood Urea	Numeric
Number of procedures	Numeric	Sodium	Numeric
Number of medications	Numeric	Potassium	Numeric
Number of outpatient visits	Numeric	Hemoglobin	Numeric
Number of emergency visits	Numeric	White Blood Cell Count	Numeric
Number of inpatient visits	Nominal	Hypertension	Numeric
Diagnoses 1	Nominal	CoronaryArtery Disease	Nominal
Diagnoses 2	Nominal	Appetite	Nominal
Diagnoses 3	Numeric	A1c test result	Nominal
Number of diagnoses	Nominal	Change of medications	Nominal
Glucose serum test result	Nominal	Readmitted	Nominal

IV. METHODOLOGY

The task stated previously is to apply binary classification to predict the existence of diabetes, heart, and kidney chronic diseases using the data said earlier. This part will display the classification models applied in this study. The proposed method architecture is described in Fig. 1.

A. Data Acquisition

In this step, we gather electronic health record (EHR) information from medical facilities, healthcare centers, or publicly available datasets, while guaranteeing that the data is anonymized and compiles with applicable privacy regulations.

B. Data Preprocessing

Before we use the datasets for prediction, we analyze them to obtain their drawbacks. In the diabetes datasets, we observed that there are many features with missing value ratio, as shown in Fig. 2.

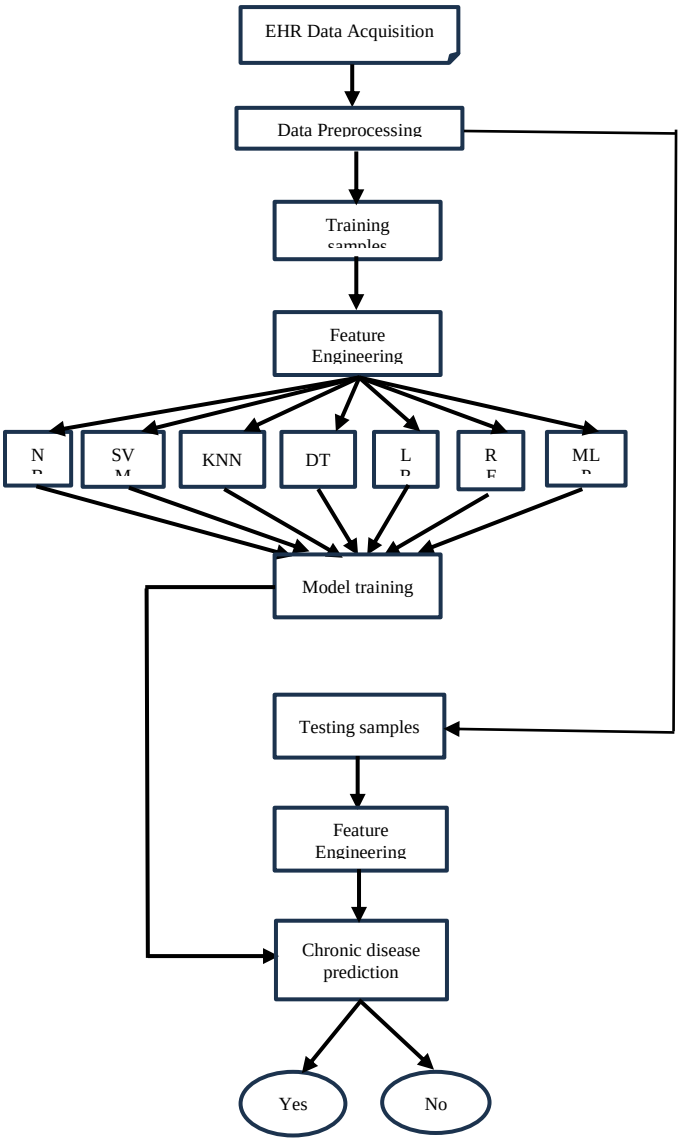


Fig. 1. Proposed method architecture.

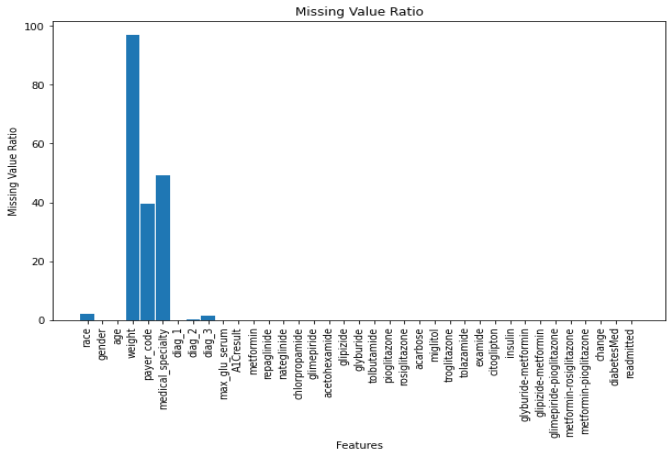


Fig. 2. Features with missing value ratio in diabetes dataset.

After applying feature reduction based on missing value ratio with ratio more than or equal 40%, "weight", "payer_code" and "medical_specialty" features have been dropped from dataset. Then we are dealing with features that have less than 40%, we have "diag_1", "diag_2", "diag_3" and "race" features. We have removed any row that have a missing value in any features and this through the following:

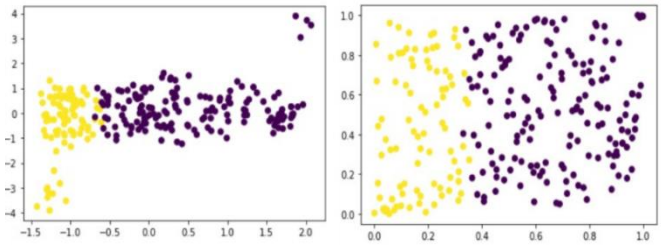
- Retrieving the index of the row where all three (diag_1, diag_2 and diag_3) values are missing and storing it in a set.
- Retrieving the index of the rows where only diag_1 value is missing and appending the same in the set.
- Retrieving the index of the rows where only diag_2 value is missing and appending the same in the set.
- Retrieving the index of the rows where only diag_3 value is missing and appending the same in the set.
- Retrieving the index of the rows where only race value is missing and appending the same in the set.
- Then drop all rows that have an index in the set.

Now we have a dataset without missing value. The next step is to remove columns (features) that have no relation with our output; in another word "featureless" such as "encounter_id" and patient_nbr".

So, we now have a clean dataset without any missing value and featureless columns but there are some columns still with text not numbers. In order to that, we have labeled these columns with integer numbers. Then to make our classifiers and model avoid overfitting, we make data regularization to all data.

Pearson's Correlation Coefficient helped us identify the connection between two quantities by assessing the intensity of their association. Also, it provides a metric ranging from -1 to +1 to quantify this relationship. A value of one signifies a strong correlation, while zero indicates no correlation. A heatmap is a visual representation of data in two dimensions, utilizing color to convey information. This graphical method assists users in visualizing both simple and complex data. This heat map is shown in Fig. 4 in Appendix C below. We observed the last row, "target" and noted its correlation scores with different features. "examide", "citoglipton" and "metformin-rosiglitazone" features are not correlated with "target". They don't contribute much to the model, so we dropped them.

To treat the outliers, we used Quantile Transformer. This technique converts the features into a uniform or normal distribution. Therefore, this conversion typically causes the most common values of a feature to be more widely distributed. Furthermore, this preprocessing scheme is considered robust as it effectively lessens the influence of minor outliers in the dataset. Fig. 3 shows the difference between applying Standard Scalar vs. Quantile Transformation. In Standard Scalar, Fig. 3(a), the Y-axis has eight units, whereas X-axis has only 3.5 units, indicating that Outliers have affected the scales. After applying Quantile Transformation, the Y-axis, and X-axis are equally scaled as shown in Fig. 3(b). The outliers are still present in the dataset, but their impact has been reduced.



(a) (b)

Fig. 3. (a) Standard scalar.(b) Quantile transformation.

We applied these preprocessing processes to heart and kidney chronic diseases datasets as they had the same problems, imbalance, and outliers.

C. Feature Engineering

Feature engineering involves identifying and extracting meaningful features from electronic health record (EHR) data that may be associated with chronic diseases. This process includes both feature extraction, which involves identifying relevant features, and feature selection, which involves selecting the most informative features for predicting the diseases.

D. Split Data

After Partitioning the dataset into appropriate training, validation, and testing subsets, using a suitable ratio, as follows, the training set contains 70% of the data, the validation and the test set are 15% each.

E. Classifiers

1) Naïve Bayes (NB): The "Naïve Bayes Classifier" refers to straightforward probability-based classifier which utilizes Bayes' theorem with dependable assumptions of independence. It anticipates that the proximity or lack of a particular class component will depend on the proximity or lack of any other element [19]. Conditional probabilities are necessary for the Naive Bayes calculation. It employs the Bayes hypothesis, a method for computing probability

through frequency counting of different qualities and attributes combinations in the collected data. The Bayes Theorem estimates the likelihood of an event, given the possibility of an earlier event. The Bayes hypothesis can be stated as follows [24]. When "b" addresses the current event for the needy and "a" addresses a previous occasion, Bayes' theorem can be formulated in the following manner:

$$Prob\ (a|b)\ = \frac{Prob\ (b|a) * Prob(b)}{Prob(a)} \qquad (1)$$

Here, Prob $(a|b)$ denotes the probability of "b", and Prob $(b|a)$ represents the probability of "b" given "a". The main benefit of the Naive Bayes classifier is the minimal training data it requires to compute mean and variance parameters essential for classification. Under the assumption of independent variables, Variances of for each class's variables need to be calculated only, rather than the overall variance. The reason for using the naïve bayes is that it is fast, scalable and useful for dealing with missing data.

2) Support Vector Machine (SVM): SVM is a supervised learning technique used for regression and classification tasks, known as Support Vector Regression (SVR) and Support Vector Classification (SVC), correspondingly. It is recommended for use with more diminutive datasets as the processing time for larger datasets is lengthy [20]. The fundamental principle of SVM involves determining the hyperplane which efficiently splits the dataset's features to distinct domains, where the points closest to the hyperplane are designated as support vectors, and their distance from the hyperplane is known as the margin. This algorithm aims to locate a hyperplane that minimizes the likelihood of misclassifying cases in the test dataset [21]. SVM is very useful when the data is not distributed regularly, and it doesn't suffer from overfitting.

3) K-Nearest Neighbor (KNN): KNN is a supervised learning algorithm that employs proximity to classify or forecast the grouping of a singular data point [22]. Although support vector machines (SVM) can be used for classification and regression tasks, it is commonly used as a classification technique, relying on the concept that analogous points are usually in proximity to one another. KNN is a lazy learning technique that doesn't generalize the training data points, resulting in a quick training phase. It predicts the classification of a new sample point using data from multiple classes. Additionally, KNN is non-parametric, making no assumptions about the data, and the model is derived directly from the data [24]. KNN was used since it doesn't need to be trained before producing predictions, and so new data can be supplied without disrupting the system's accuracy.

4) Decision tree: The classification process can be directed using decision trees [23]. They are arranged in a simple tree structure, with terminal nodes displaying the results of decisions and non-terminal nodes representing tests on one or more attributes. By using information gain and gain ratio measures as splitting models, C4.5 improved the ID3 decision tree induction algorithm. The decision tree construction algorithm can be summed up as follows: 1) Decide which attribute best distinguishes the values of the output attribute. 2) Make a distinct branch of the tree for each attribute value. 3) Create subgroups from the instances to correspond with the attribute values of the chosen node. For each subgroup, stop the attribute selection process if the following conditions are met: a) When all of the instances in a subgroup have the same value for an output attribute, the attribute selection process for the current path is stopped, and the branch on the current path is marked with the chosen value. b) There is only one node in the subgroup or no other distinguishing characteristics can be found. Label the branch with the output value seen in most of the remaining cases, just like in (a). 5) Repeat the procedure above for each subgroup created in (3) that has not been designated as terminal [24].

5) Random forest: The random forest algorithm, as explained in [25], is composed of multiple decision trees that are trained on a bootstrapped sample of the data set, where one-third of the data is reserved for testing as the out-of-bag sample. To introduce more diversity and reduce correlation among decision trees, feature bagging injects another instance of randomness. Before training, the random forest algorithm requires three crucial hyperparameters to be specified: node size, number of trees, and number of sampled features. The random forest classifier can be employed to address both regression and classification tasks, and the method of determining the prediction varies depending on the type of problem. For a regression task, the decision trees' predictions are averaged, while for classification tasks, the predicted class is determined through a majority vote, selecting the most frequent categorical variable [25].

6) Logistic Regression (LR): Logistic regression is a machine learning technique which is categorized as a supervised machine learning model. It is also regarded as a discriminative model that aims to differentiate between classes or categories [25]. Logistic regression is commonly used for prediction and classification problems such as Fraud detection, Disease prediction, and Churn prediction. This analytical approach in medicine can anticipate the probability of disease or illness for a particular population, enabling healthcare organizations to establish preventive care measures for individuals with a higher susceptibility to specific conditions [26].

7) Multi-Layer Perceptron (MLP): Three layers make up an MLP, a feed-forward neural network: an input layer, an output layer, and a hidden layer [25]. While the output layer completes various tasks like classification and prediction, the input layer receives the input signal for processing. There are numerous hidden layers that make up the computational engine of the MLP that lie between the input and output layers. Data flows forward from the input layer to the output layer, much like a feed-forward network in an MLP. The MLP neurons are trained with the backpropagation learning algorithm. Since MLPs can approximatively solve any continuous function, they can solve problems that cannot be

linearly separated. They are used in a variety of tasks including classification, pattern recognition, forecasting, and approximation. [27]. We used MLP as it is very flexible and can be used for learning mapping from inputs to outputs generally.

V. Experimental Results

To evaluate the proposed model, several experiments have been carried out using different parameters as shown in Table II. All experiments have been done using Python programming language by Jupiter notebook editor with some machine learning toolboxes (e.g. Scikit-learn, NumPy, and matplotlib).

TABLE II. Parameters of Some Proposed Models

Algorithm	Parameters
SVM	**Kernel functions** = {linear, sigmoid, RBF}
KNN	Number of neighbors (n) = {30, 40, 50}
MLP	Number of hidden layers (h) = {5, 10, 20}

The accuracy and f1-score are calculated to examine how classification is carried out. As a result of the classifier, four cases are taken into consideration.

- True Positives (TP): The sample size correctly assigned to that class.
- True Negatives (TN): The sample size correctly excluded from that class.
- False Positives (FP): The sample size wrongly excluded from that class.
- False Negatives (FP): The sample size wrongly assigned to that class.

The number of accurate and wrong classifications in each potential prediction of the classified variables is used to evaluate how effective the classification model is. The accuracy is calculated using the following formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

We have utilized three medical datasets related to diabetes, heart, and kidney disease. These datasets were all retrieved from the UCI machine learning library [15, 17, 18]. The objective is to compare all classifiers, including Naïve Bayes, KNN, SVM, DT, RF, LR, and MLP, and to categorize the diseases depending on parameter choices. The accuracy of several classifiers for diabetes and heart disease datasets is shown in Table III. After applying some experiments with different parameters shown in Table II, we decided on the best hyperparameters for SVM, KNN, and MLP classifiers. For SVM, three kernel functions have been utilized and the RBF kernel has achieved the highest accuracy on diabetes and kidney datasets. Three values for the number of neighbors were used for the KNN classifier and the best one was 40. In MLP, the best number of hidden layers was 10, which achieved the highest accuracy on the heart dataset.

In Table IV, a recognized confusion matrix is obtained to estimate four measures, accuracy, precision, recall, and f-score.

TABLE III. Accuracy of All Classifiers for Diabetes, Heart, and Kidney Datasets

Dataset	Classifier	Accuracy	Precision	Recall	F-score
Diabetes	Naïve Bayes	88.3%	85.5%	78.1%	83.5%
	SVM	**88.5%**	**92.4%**	**99.8%**	**93.6%**
	KNN	88.2%	57.3%	51.5%	49.5
	DT	82.2%	91.2%	85.8%	84.2%
	RF	88.4%	44.1%	51.5%	47.2%
	LR	88.3%	69.2%	52.3%	48.5%
	MLP	75.1%	39.5%	20.4%	26.4%
Heart	Naïve Bayes	83.1%	83.6%	82.5%	81.6%
	SVM	83.6%	83.8%	81.4%	81.5%
	KNN	83.1%	83.5%	80.6%	81.5%
	DT	80.9%	79.8%	78.8%	78.8%
	RF	84.2%	84.5%	84.4%	85.1%
	LR	82.6%	82.8%	83.6%	84.9%
	MLP	**84.7%**	**88.5%**	**87.4%**	**87.8%**
Kidney	Naïve Bayes	86.8%	82.6%	77.6%	74.7%
	SVM	**94.9%**	**94.5%**	**90.2%**	**92.6%**
	KNN	81.8%	85.7%	53.6%	54.8%
	DT	89.9%	88.5%	68.4%	82.3%
	RF	86.8%	84.6%	71.0%	75.3%
	LR	85.8%	87.4%	67.2%	71.1%
	MLP	89.8%	65.7%	68.5%	66.4%

TABLE IV. Confusion Matrix of SVM, DT, and MLP classifiers for Diabetes, Heart, and Kidney chronic datasets

Dataset	Classifier	Precision	Recall	F-measure	Class
Diabetes	NB	0.85	0.85	0.83	Yes
		0.86	0.89	0.84	No
	SVM	0.93	1.00	0.92	Yes
		0.92	1.00	0.93	No
	KNN	0.25	0.04	0.06	Yes
		0.89	0.99	0.93	No
	DT	0.85	0.87	0.82	Yes
		0.97	0.83	0.86	No
	RF	0.88	1.00	0.00	Yes
		0.00	0.00	0.94	No
	LR	0.50	0.01	0.03	Yes
		0.88	1.00	0.94	No
	MLP	0.40	0.20	0.27	Yes
		0.39	0.20	0.26	No
Heart	NB	0.85	0.87	0.86	Yes
		0.81	0.78	0.79	No
	SVM	0.83	0.90	0.86	Yes
		0.84	0.75	0.79	No
	KNN	0.83	0.90	0.86	Yes
		0.84	0.74	0.79	No
	DT	0.83	0.84	0.84	Yes
		0.78	0.77	0.77	No
	RF	0.85	0.89	0.87	Yes
		0.83	0.78	0.81	No
	LR	0.84	0.87	0.85	Yes
		0.81	0.77	0.79	No

	MLP	0.89 0.87	0.88 0.86	0.88 0.87	Yes No
Kidney	NB	0.77 0.88	0.50 0.96	0.61 0.92	Yes No
	SVM	0.94 0.95	0.80 0.99	0.86 0.97	Yes No
	KNN	1.00 0.81	0.10 1.00	0.18 0.90	Yes No
	DT	0.86 0.91	0.60 0.97	0.71 0.94	Yes No
	RF	0.82 0.88	0.45 0.97	0.58 0.92	Yes No
	LR	0.88 0.86	0.35 0.99	0.50 0.92	Yes No
	MLP	0.66 0.65	0.81 0.57	0.72 0.61	Yes No

The classification outcomes are presented as a matrix by the confusion matrix. It includes information for actual and predicted classes made using the classification framework. The cell represents the sample size that was classified as true when they were actually true (i.e., TP) and those classified as false while quiet (i.e., TN), respectively. The remaining two cells represent the number of incorrectly classified pieces. In fact, the cells denoting the sample size labeled false when they were actually true (i.e., FN) and the sample size labeled true when they were wrong (i.e., FP) [28]. After the confusion matrices are constructed, the following formulas can determine the precision, recall, and F-measure:

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{3}$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \tag{4}$$

$$\text{F-measure} = \frac{2 * TP}{(2 * TP + FP + FN)} \tag{5}$$

VI. Conclusion and Future Work

In this study, we generated electronic healthcare records with symptoms data sourced from three distinct datasets in the UCI machine learning repository for predicting diabetes, heart disease, and chronic kidney disease. Data of the patients were trained with various classifiers, including Naïve Bayes, SVM, Decision Tree, Random Forest, Logistic Regression, and Multilayer Perceptron. The results of this study indicate that data mining techniques can be effectively employed to identify and predict chronic diseases. The results indicate that the Support Vector Machine (SVM) algorithm was the most successful in predicting diabetes and kidney diseases, while the Multilayer Perceptron (MLP) algorithm was the optimal choice for predicting heart disease. The Naïve Bayes and Random Forest algorithms also demonstrated strong performance across all datasets.

References

[1] R. Kleiman, P. Bennett, P. Peissig, R. Berg, Z. Kuang, S. Hebbring, M. Caldwell, and D. Page. “High-Throughput Machine Learning from Electronic Health Records”. Proceedings of Machine Learning Research 85:1–24, Machine Learning for Healthcare, 2019.

[2] M. Ghaderzadeh, M. Aria. “Management of Covid-19 Detection Using Artificial Intelligence in 2020 Pandemic”. Proceedings of the 5th International Conference on Medical and Health Informatics. ICMHI 2021, May 14–16, 2021, Japan.

[3] ME. Hossain, A. Khan, MA. Moni, S. Uddin. “Use of Electronic Health Data for Disease Prediction: A Comprehensive Literature Review”. IEEE/ACM Trans Comput Biol Bioinform. 2021 Mar-Apr;18(2):745-758. doi: 10.1109/TCBB.2019.2937862. Epub 2021 Apr 6. PMID: 31478869.

[4] J. Al-Otaibi, E. Tolma, W. Alali, D. Alhuwail, S. Aljunid. “The Factors Contributing to Physicians’ Current Use of and Satisfaction With Electronic Health Records in Kuwait’s Public Health Care: Cross-sectional Questionnaire Study”. JMIR Med Inform.10(10): e36313 URL: https://medinform.jmir.org/2022/10/e36313 DOI: 10.2196/36313, 2022.

[5] W. Tsao, W. Aday, I. Almarzooq, A. Alonso, Z. Beaton, S. Bittencourt, “Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association”, American Heart Association, https://doi.org/10.1161/CIR.0000000000001052Circulation. 145:e153–e639. 2022.

[6] S. Suh, SO. Song, JH. Kim, H. Cho, WJ. Lee, BW. Lee. “Effectiveness of Vildagliptin in Clinical Practice: Pooled Analysis of Three Korean Observational Studies (the VICTORY Study)”. J Diabetes Res. 2017:5282343. doi: 10.1155/2017/5282343. Epub 2017 Aug 24. PMID: 29057274; PMCID: PMC5613692, 2017.

[7] JC. Lv, LX. Zhang. “Prevalence and Disease Burden of Chronic Kidney Disease”. Adv Exp Med Biol. 1165:3-15. doi: 10.1007/978-981-13-8871-2_1. PMID: 31399958. 2019.

[8] Ammirati, Adriano. “Chronic Kidney Disease. Revista da Associação Médica Brasileira”. 66. s03-s09. 10.1590/1806-9282.66.s1.3. 2020

[9] N, Yaser, L, Zhu, C. Junde, Z, Qiu, X. Yuan, D.N, Yahya, E. Sajad. “Diagnosis of Chronic Diseases Based on Patients’ Health Records in IoT Healthcare Using the Recommender System”. Wireless Communications and Mobile Computing. 2022. 1-14. 10.1155/2022/5663001.

[10] D. Chicco, C. A. Lovejoy and L. Oneto, "A Machine Learning Analysis of Health Records of Patients With Chronic Kidney Disease at Risk of Cardiovascular Disease," in *IEEE Access*, vol. 9, pp. 165132-165144, doi: 10.1109/ACCESS.2021.3133700. 2021.

[11] Hohman, Katherine H. DrPH, MPH; Martinez, Amanda K. MPH, MSN, RN; Klompas, Michael MD, MPH; Kraus, Emily M. PhD, MPH; Li, Wenjun PhD; Carton, Thomas W. PhD, MS; Cocoros, Noelle M. DSc, MPH; Jackson, Sandra L. PhD, MPH; Karras, Bryant Thomas MD; Wiltz, Jennifer L. MD, MPH; Wall, Hilary K. MPH. Leveraging Electronic Health Record Data for Timely Chronic Disease Surveillance: The Multi-State EHR-Based Network for Disease Surveillance. Journal of Public Health Management and Practice 29(2):p 162-173, March/April 2023. | DOI: 10.1097/PHH.0000000000001693.

[12] D. Kumari, S. Seema. “Predictive analytics to prevent and control chronic diseases”. 381-386. 10.1109/ICATCCT.2016.7912028. 2016.

[13] A. Perotte, R. Ranganath, JS. Hirsch, D. Blei, N. Elhadad. “Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis”. J Am Med Inform Assoc. 2015 Jul;22(4):872-80. doi: 10.1093/jamia/ocv024. Epub 2015 Apr 20. PMID: 25896647; PMCID: PMC4482276.

[14] A. Hazazi, A. Wilson. “Leveraging electronic health records to improve management of noncommunicable diseases at primary healthcare centres in Saudi Arabia: a qualitative study”. BMC Fam Pract. 2021 May 27;22(1):106. doi: 10.1186/s12875-021-01456-2. PMID: 34044767; PMCID: PMC8157615.

[15] Areej A. Malibari, An efficient IoT-Artificial intelligence-based disease prediction using lightweight CNN in healthcare system, Measurement: Sensors, Volume 26, 2023, 100695, ISSN 2665-9174, https://doi.org/10.1016/j.measen.2023.100695.

[16] B. Strack, JP. DeShazo, C. Gennings, JL. Olmo, S. Ventura, KL. Cios, JN. Clore. “Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records”. Biomed Res Int.;2014: 781670. doi: 10.1155/2014/781670. Epub 2014 Apr 3. PMID: 24804245; PMCID: PMC3996476.

[17] A. Tanvir, M. Assia, B. Sajjad, A. Muhammad, A. Raza. “Survival analysis of heart failure patients: A case study”. PLOS ONE. 12. e0181001. 10.1371/journal.pone.0181001, (2017).

[18] S. Al-Shamsi, D. Regmi, and R. D. Govender, ''Chronic kidney disease in patients at high risk of cardiovascular disease in the United Arab Emirates: A population-based study,'' PLoS ONE, vol. 13, no. 6, Jun. 2018, Art. no. e0199920.

[19] Z. Yang, J. Ren, Z. Zhang, Y. Sun, CH. Zhang, M. Wang, and L. Wang. A New Three-Way Incremental Naive Bayes Classifier. Electronics. 12. 1730. 10.3390/electronics12071730. (2023)

[20] T. Dai and Y. Dong, "Introduction of SVM Related Theory and Its Application Research," 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Shenzhen, China, pp. 230-233, doi: 10.1109/ AEMCSE50948.2020.00056. 2020

[21] Christopher J.C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery", Springer, 2(2), pp.121-167, 1998.

[22] M. A. Abdelaal, M. A. Fattah, and M. M. Arafa. "Predicting Sarcasm and Polarity In Arabic Text Automatically: Supervised Machine Learning Approach." Journal of Theoretical and Applied Information Technology 100.8 (2022).

[23] Garavand, Ali & Behmanesh, Ali & Aslani, Nasim & Sadeghsalehi, Hamidreza & Ghadezadeh, Mustafa. (2023). Towards Diagnostic Aided Systems in Coronary Artery Disease Detection: A Comprehensive Multiview Survey of the State of the Art. International Journal of Intelligent Systems. 1-19. 10.1155/2023/6442756. 2023

[24] A.G. Karegowda, V. Punya, M.A. Jayaram, A.S .Manjunath," Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5", International Journal of Computer Applications (0975 – 8887) Volume 45– No.12, May 2012.

[25] A. Garavand, C. Salehnasab, A. Behmanesh, N. Aslani, AH. Zadeh, M. Ghaderzadeh. "Efficient Model for Coronary Artery Disease Diagnosis: A Comparative Study of Several Machine Learning Algorithms". J Healthc Eng. Oct 18;2022:5359540. doi: 10.1155/2022/5359540. PMID: 36304749; PMCID: PMC9596250. 2022.

[26] Maalouf, Maher. "Logistic regression in data analysis: An overview". International Journal of Data Analysis Techniques and Strategies. 3. 281-299. 10.1504/IJDATS.2011.041335. 2011.

[27] T. Soliman A. Abd-elaziem. "A Multi-Layer Perceptron (MLP) Neural Networks for Stellar Classification: A Review of Methods and Results". International Journal of Advances in Applied Computational Intelligence. 3. 10.54216/IJAACI.030203. 2023.

[28] D.S. Kumar, G. Sathyadevi, and S. Sivanesh, "Decision Support System for Medical Diagnosis Using Data Mining", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011 ISSN (Online): 1694-081.

APPENDICES

A. *Appendix A*

TABLE V. FREQUENCIES, PERCENTAGES, MEANS, AND STANDARD DEVIATIONS OF THE RESPONDENTS' RESPONSES TO TECHNICAL INFRASTRUCTURE QUESTIONS

N.	phrase	Response				Average	Standard deviation	Phrase arrangement	Response degree
			Disagree	Neutral	Agree				
22	Kuwait is considered a high-income country that has advanced healthcare infrastructure	F	3	6	21	2.60	.675	5	High
		%	10.0	20.0	70.0				
23	There is a need for the establishment of a well-designed primary health care network.	F	0	4	26	2.87	.346	2	High
		%	0.0	13.3	86.7				
24	The use of computerized information systems in different healthcare facilities	F	2	5	23	2.70	.596	3	High
		%	6.7	16.7	76.7				
25	Health informatics is already implemented in the State of Kuwait.	F	3	11	16	2.43	.679	7	High
		%	10.0	36.7	53.3				
26	The existence of reliable registration, licensing, and authorization systems for medical information.	F	3	10	17	2.47	.681	6	High
		%	10.0	33.3	56.7				
27	The need to employ e-health solutions to improve the quality of the healthcare sector.	F	0	3	27	2.90	.305	1	High
		%	0.0	10.0	90.0				
28	The huge investment that is being allocated to develop the technical infrastructure of the State	F	3	5	22	2.63	.669	4	High
		%	10.0	16.7	73.3				
Overall average						2.66	.409	--	High

C. Appendix B

TABLE VI. FREQUENCIES, PERCENTAGES, MEANS, AND STANDARD DEVIATIONS OF THE RESPONDENTS' RESPONSES TO ORGANIZATION AND ADMINISTRATION QUESTIONS

N.	phrase	Response				Average	Standard deviation	Phrase arrangement	Response degree
			Disagree	Neutral	Agree				
29	Kuwait Vision 2035 pays great attention to the development of the national healthcare system.	F	3	10	17	2.47	.681	7	High
		%	10.0	33.3	56.7				
30	The advancement in health research capacities in the State of Kuwait needs more interest in information technology	F	1	3	26	2.83	.461	3	High
		%	3.3	10.0	86.7				
31	The need for high-quality services in healthcare public and private sectors.	F	0	2	28	2.93	.254	1	High
		%	0.0	6.7	93.3				
32	The need for a national health information management strategy to guide public health policies.	F	0	4	26	2.87	.346	2	High
		%	0.0	13.3	86.7				
33	The emergence of medical equipment manufacturers has appeared as a promising field in the State.	F	4	7	19	2.50	.731	6	High
		%	13.3	23.3	63.3				
34	The healthcare system in Kuwait depends on an integrated information and technology framework.	F	7	8	15	2.27	.828	8	High
		%	23.3	26.7	50.0				
35	The high priority that is given to the healthcare sector to improve the quality of services being provided to citizens	F	4	6	20	2.53	.730	5	High
		%	13.3	20.0	66.7				
36	The huge investment that is being allocated to develop the technical infrastructure of the State.	F	3	8	19	2.53	.681	4	High
		%	10.0	26.7	63.3				
Overall average						2.62	.473	--	High

D. Appendix C

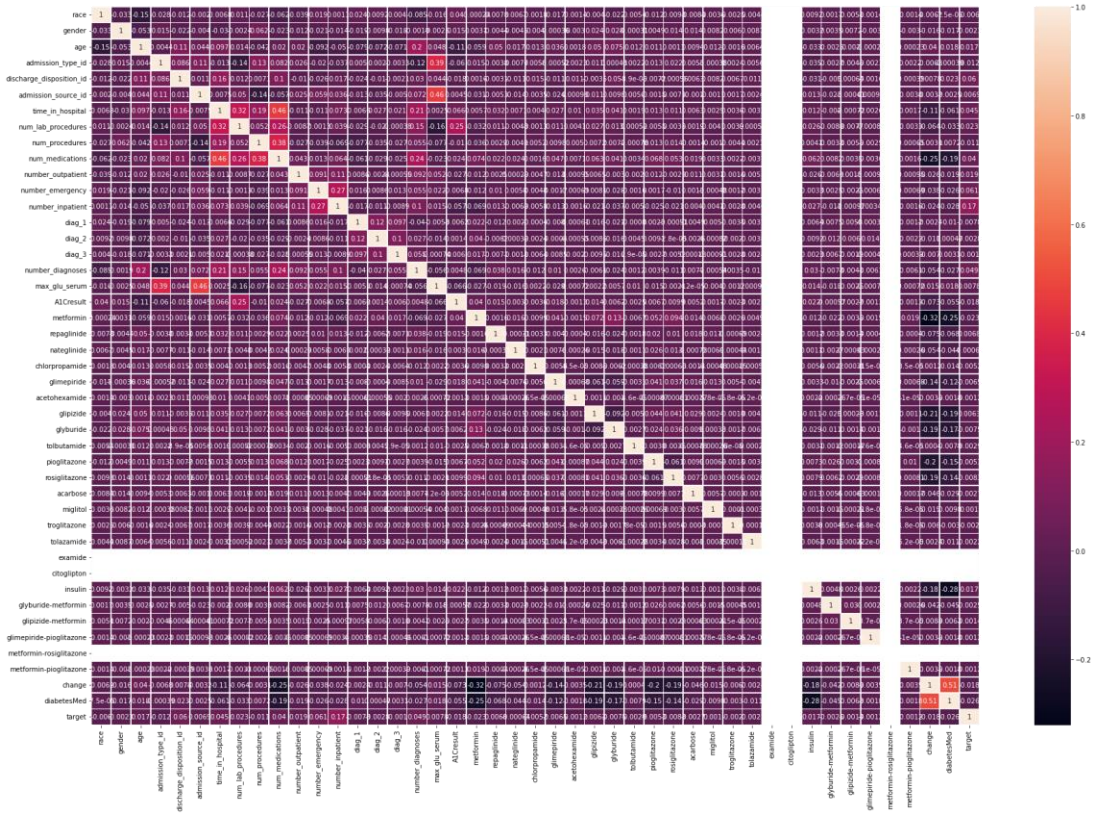


Fig. 4. Box Plot of features to see the outliers in diabetes dataset.

Separability-based Quadratic Feature Transformation to Improve Classification Performance

Usman Sudibyo, Supriadi Rustad, Pulung Nurtantio Andono, Ahmad Zainul Fanani
Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

***Abstract*—Feature transformation is an essential part of data preprocessing to improve the predictive performance of machine learning (ML) algorithms. Box-Cox transformation with the goal of separability is proven to align with the performance improvement of ML algorithms. However, the features mapped using Box-Cox transformation preserve the order of the data, so it is ineffective when used to improve the separability of multimodal distributed features. This research aims to build a feature transformation method using quadratic functions to improve class separability that can adaptively change the order of the data when necessary. Fisher score (Fs) measures the separability level by maximizing the Fisher's Criteria of the quadratic function. In addition to increasing the Fs value of each feature, this method can also make the feature more informative, as evidenced by the increasing value of information gain, information gain ratio, Gini decrease, ANOVA, Chi-Square, reliefF, and FCBF. The increase in Fs is particularly significant for bimodally distributed features. Experiments were conducted on 11 public datasets with two statistical-based machine learning algorithms representing linear and nonlinear ML algorithms to validate the success of this method, namely LDA and QDA. The experimental results show an improvement in accuracy in almost all datasets and ML algorithms, where the highest accuracy improvement is 0.268 for LDA and 0.188 for QDA.**

***Keywords*—Separability; feature transformation; quadratic function; fisher's criterion; fisher score**

I. INTRODUCTION

Machine Learning (ML) is an essential branch of artificial intelligence widely used for pattern recognition, image processing, text classification, intrusion detection systems, etc. [1]. However, the ML algorithm's success depends on the features' quality. The resulting model is also good if the features are good [2]. Therefore, improving features to suit ML algorithms' needs is an essential topic in feature engineering [3]. Feature transformation is one of the feature engineering techniques that can be used to improve features before being input into the ML algorithm [2], [4]. There have been many studies that discuss feature improvement through transformation techniques, where with the proper feature transformation, feature quality can be improved [5], [6].

Feature transformations developed to improve feature quality are generally grouped into three: first, feature transformations that only change the scale (e.g., Min-Max normalization, Z-Score normalization) [7], second nonlinear feature transformations that do not change the order of the data (e.g. log transform [8], square root transform [9], Box-cox transform, Yeo-Johnson transform [10]), and third, nonlinear feature transformations that can change the order of the data (e.g. kernel function in SVM) [11], [12]. Some machine learning algorithms are sensitive to scale differences, so normalization or standardization is needed to uniform the scale of features. Paper in [13] discusses the effect of various data normalization methods on support vector machine (SVM) algorithms and technical indicators to predict stock index price movements. The result is a slight increase in accuracy performance. Paper in [6] proposed mixed feature transformation methods such as CDF transformation and Symmetric log1p transformation, where feature transformation can substantially improve the performance of neural ranking models compared to directly using raw features. Paper in [14] compares the effect of Box-Cox transformation to improve two-dimensional images with advanced low-light image enhancement techniques. Paper in [15] addresses issues in nonlinear stochastic degradation modeling and prognostics from the Box-Cox transform (BCT) perspective, where BCT is used to transform nonlinear degradation data into near-linear data. Adaptive Box-Cox (ABC) transformation was introduced by [16], where adaptive parameter tuning is used to normalize data in various distributions that cannot be properly normalized using conventional data transformation algorithms, including log and square root transformations.

In general, feature transformation aims to change the data distribution to be close to Gaussian in order to improve ML performance, such as log transformation [8], [17], square root transformation [18], Box-Cox transformation [14], [19]–[23], and Yeo-Johnson transformation [10], [24]. However, the experimental results of Bicego & Baldo, 2016 [20] showed different results. Their findings show that ML classification accuracy improves when the data distribution is far from Gaussian and is more related to the class separability problem. This finding is corroborated by [21], [25]–[27], which state that, based on the fisher criterion, features with greater class separability are considered more informative and can improve ML classification performance. Based on these findings, ML classification performance can be improved when features have large separability. This transformation can improve class separability, even in cases where the original dataset is not linearly separable.

Bicego and Baldo's research above uses the Box-Cox transformation, where this transformation is monotonous because it cannot change the order of the data [20]. This condition causes the Box-Cox transformation not to produce maximum separability. In addition, the result of the Box-Cox transformation is determined by a parameter that is searched using the grid search method so that getting the best parameter of each feature associated with the maximum fisher value

requires high computational costs [21]. Bicego and Baldo's empirical analysis of the behavior of the Box-Cox transform for pattern classification opens up opportunities for the analysis of different nonlinear data pre-processing methods that can improve class separability.

This research proposes a quadratic feature transformation for preprocessing that is directly designed to maximize the class separability of each feature. The idea is to optimize the quadratic function parameters using Fisher's optimization criterion to obtain maximum class separability. In this way, each feature is transformed using a quadratic function to have a higher fisher score compared to the original feature's fisher score. The quadratic transformation process is performed at the preprocessing stage, making it flexible to be combined with various ML algorithms. Since this technique is directly geared towards maximizing the fisher score, while the mean and variance of each class can be easily calculated, it has the potential to be applied to multi-class data. Various feature quality test metrics, such as Information Gain [28]–[30], Gain ratio [31], Gini Decrease [32], Anova [33], [34], Chi-Square [35], ReliefF [36], [37], and Fast Correlation-Based Feature selection (FCBF) [38], [39], are used to test the feature quality of the proposed Box-Cox transformation and Quadratic transformation, before finally comparing their respective performance.

The classification algorithms used in this study are Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), which represent linear and nonlinear ML algorithms. Although LDA has proven to be an excellent classification and dimensionality reduction algorithm, it produces poor vector projections on multimodal data [40]–[42]. With feature transformation, it is expected that the multimodal influence on LDA can be reduced. QDA was chosen because it is one of the most commonly used classifiers in practice and is quite simple. In addition, QDA has been shown to improve performance when paired with the nonlinear feature extraction technique quadratic Fisher transformation [43].

The contributions of this research include:

1) Development of separability-based feature transformation to optimize the classification task of machine learning algorithms.

2) The chosen quadratic transformation is generally able to improve feature quality based on fisher score, information gain, chi-square, relief, and FCBC values on the dataset studied.

This paper is organized into several sections, starting with an introduction in Section I, followed by a brief explanation of fisher score and quadratic function in Section II, research methodology containing the dataset and the proposed method in Section III, results and discussion in Section IV, and then closed with a conclusion in Section V.

II. Preliminary Works

To prepare for a better understanding of this research, some feature transformation techniques, fisher score, and quadratic function definition and application will be introduced.

A. Feature Transformation Technique

In some literature, the use of the term feature transformation is often equated with feature engineering, feature extraction, and feature construction [44]. However, this study consistently uses the term feature transformation as a univariate feature engineering technique.

Feature transformations, which T. Verdonck et.all 2021[2] call feature engineering, are grouped into two, namely univariate and multivariate feature engineering techniques. Univariate feature transformations on continuous variables can improve symmetry, normality, or model fit, such as logarithmic transformation, Box-Cox transformation, and Yeo-Johnson transformation. Multivariate feature transformations aim to reduce the dimensionality of the data by creating new features that are linear combinations of the original variables, such as PCA, LDA, SVD, UV (non-negative) decomposition, and tensor decomposition.

The study of feature transformation has progressed quite well. The following studies are related to feature transformation. A feature transformation method based on Mutual Information (MI) is proposed by [45], where the Probability Density Function (PDF) of features in the class is assumed to be Gaussian. The gradient descent technique is used to maximize the mutual information between features and classes. Experimental results show that the proposed MI projection consistently outperforms other methods for various cases. Most of these Medical decision support systems (MDSS) focus on feature transformation-based methods and their integration with machine learning models for the prediction of risks associated with Heart failure (HF). However, the improvement in accuracy on test data is not followed by training data. This study proposes a more robust approach that integrates stacked autoencoder grids with neural network models to address the problem[46]. Most feature engineering in the input space relies on manually defined transformation functions. However, research [12] builds transformation functions automatically learned through autoencoders for latent representation extraction and multi-layer perceptron (MLP) regressors. The transformation function built in this way can also improve the performance of LSVM and JST, when embedded as a preprocessing step. A random projection-based feature transformation method using Metaheuristic Optimization Algorithm [47] is proposed to map data points from the original space to a new binary space, where the random projection process is formulated as an optimization problem. The transformation of features to binary space is needed when the system requires coarse quantization of measurements. A new guided FT method called minimax probabilistic feature transformation (MPFT) was proposed for multi-class datasets [48]. The idea of this method is based on trying to control the probability of correct classification of future test points as large as possible in the transformed feature space. Past tax default prediction by applying diverse feature transformation techniques and advanced machine learning approaches was proposed by [49]. A combination of feature transformations, such as logarithmic and square root transformations, is able to improve tax default prediction performance. A feature transformation method to improve classification performance referred to as weight-matrix

learning (WML), was proposed by[50]. The way this method works is that WML is identified as an off-center technique with a center of 0.5 similarity.

B. Fisher Score (Fs)

Fs belongs to the classical supervised feature selection filter method, which aims to score features based on the ratio of data scatter between classes and data scatter within classes. Fs is used to measure the class discriminant properties of each feature independently. Features with higher Fisher scores are more discriminant than features with lower scores [51]. Fisher score has been widely used for feature selection on gene microarray data [52]–[57]. This study uses the fisher score as a basis for improving separability because it is conceptually easy to understand, easy to implement on various functions, and Fs is an efficient approach to data dimensionality reduction [58]. Fisher score Fs(k) is used to measure the separability of classes at the kth feature of a dataset. Mathematically, Fs(k) is calculated using Eq. (1)[52].

$$Fs(k) = \frac{\sum_{i=1}^{c} n_i (\mu_i^k - \mu^k)^2}{\sum_{i=1}^{c} n_i (\sigma_i^k)^2} \tag{1}$$

$$\text{with} \quad \mu_i^k = \sum_{j=1}^{n_i} \frac{x_{ij}^k}{n_i} \quad ; \quad \mu^k = \sum_{j=1}^{N} \frac{x_j^k}{N}$$

$$\sigma_i^k = \sum_{j=1}^{n_i} \frac{(x_{ij}^k - \mu_i^k)^2}{n_i - 1} \tag{2}$$

where c, N, n_i, μ_i^k, μ^k, and σ_i^k respectively are the number of classes, the total number of samples, the number of samples of the i-th class, the mean value of the k-th feature of the i-th class, the mean value of the k-th feature for the whole class, and the variance of the k-th feature of the i-th class.

C. Quadratic Function

Throughout the literature review, no quadratic functions were found to be used for the purpose of increasing the separability of classes in features. However, there are many uses of quadratic functions for different purposes. Quadratic kernel-free non-linear support vector machine (QSVM) uses quadratic functions as decision boundaries that are able to separate data in a non-linear manner. The decision boundary is built from a multivariate quadratic function that can replace the kernel trick in SVM when faced with problems that cannot be separated linearly [59]. The QSVM method was successfully used for credit scoring models [60] and improved accuracy and efficiency. This method, called Quadratic Fisher Discriminant Analysis (QFDA), uses linear and quadratic basis functions to improve classification accuracy by considering data variance. This method aims to maximize the fisher criterion in the transformation space using a transformation matrix[61]. Before the transformation, each feature is squared, and the features are multiplied, resulting in a significant increase in the number of features and high computational cost. One disadvantage of the QFDA method is that it may only work well if the class mean values are equal or if the vital information for classification lies in the variance of the data rather than the mean value.

From a mathematical point of view, a quadratic function is a polynomial of degree 2. Its highest exponent on the independent variable (i.e., x) is 2. The general form of the parabolic quadratic function, as shown in Eq. (5), is defined as follows [62].

Definition (General form). For fixed constants b, c ∈ R and nonzero a ∈ R, the function f : R → R given by

$$f(x) = ax^2 + bx + c \tag{3}$$

is a (real) quadratic function written in general form.

Quadratic functions are capable of transforming data to be closer for data in the same group and further away for data from different groups. This ability is the basis for using quadratic functions for feature transformation.

III. Research Methodology

A. Datasets

This study used 11 datasets, which are presented in Table I. Ten datasets were downloaded from the Keel data repository (https://sci2s.ugr.es/keel/category.php?cat=clas), including Sonar, WDBC, Ringnorm, New Thyroid, Wisconsin, Parkinson's, Splice, Balance, Spectf, and Hayes Roth. One Hungarian dataset was downloaded from https://www.openml.org/. Table I consists of five columns containing dataset information (code, dataset, features, samples, and class) and one column containing different methods using the same data. The number of features is between 4 and 60, the sample size is between 160 and 7400, and the targets consist of two classes and three classes. The data type varies from numeric, categorical, binary, and there are no missing values in all datasets. The data information is summarized in Table I.

TABLE I. Profile Dataset and Accuracy Performance of State-of-the-Art Methods

Code	Dataset	Features	Samples	Class	Method
d-1	SONAR	60	208	2	2DCSCA[47]
d-2	WDBC	30	569	2	Hybrid GPFC[63]
d-3	Ringnorm	20	7400	2	EIDA[64]
d-4	New Thyroid	5	215	2	Hybrid GPFC[63]
d-5	Wisconsin	9	683	2	SSDA[65]
d-6	Parkinson	22	195	2	2DCSCA[47]
d-7	Hungarian	11	1190	2	RFSA -MCC[51]
d-8	Splice	60	3190	2	RMB-SSVM[66]
d-9	Balance	4	625	3	GPMO[67]
d-10	Spectf	44	267	2	RFSA -MCC[51]
d-11	Hayes-Roth	4	160	3	SSDA[65]

B. Proposed Method

The proposed method can find the best parameters of the quadratic function in Eq. (3) generated by maximizing the Fisher score [68], [69], which is the ratio of variance between classes and variance within classes. The result of maximizing the Fisher score is a closed-form solution of the quadratic function parameters described in Section III (B) (1), while the feature transformation procedure is described in Section III (B) (2).

1) Separability-based quadratic feature transformation: Based on the Fisher Score function of Eq. (1) and the quadratic function of Eq. (3), Eq. (4) is generated as the basis for obtaining the best parameters.

$$Fs(k) = \frac{\sum_{i=1}^{c} n_i^k \left[a(\theta_i^k - \theta^k) + b(\mu_i^k - \mu^k) \right]^2}{\sum_{i=1}^{c} \left[n_i^k \sum_{j=1}^{n_i^k} \frac{(a((x_{ij}^k)^2 - \theta_i^k) + b(x_{ij}^k - \mu_i^k))^2}{n_i^k - 1} \right]} \quad (4)$$

The Mean square of the k-th feature, i-th class.

$$\theta_i^k = \sum_{j=1}^{n_i^k} \frac{(x_{ij}^k)^2}{n_i^k} \quad (5)$$

The Mean square of the k-th feature

$$\theta^k = \sum_{j=1}^{N} \frac{(x_j^k)^2}{N} \quad (6)$$

Arg_max(Fs(k)) Eq. (4) yields the optimal parameters a, b, and c formulated in Eq. (7) to Eq. (9).

$$a^k = 2R^k \quad (7)$$

$$b^k = -Q^k \pm \sqrt{(Q^k)^2 - 4R^k P^k} \quad (8)$$

$$c^k = 0 \quad (9)$$

Where is

$$A_i^k = (\theta_i^k - \theta^k); B_i^k = (\mu_i^k - \mu^k) \quad (10)$$

$$P^k = \sum_{i=1}^{c} A_i^k B_i^k \sum_{i=1}^{c} v((x_i^k)^2) - \sum_{i=1}^{c} A_i^2 \sum_{i=1}^{c} c(x_i^k, (x_i^k)^2) \quad (11)$$

$$Q^k = \sum_{i=1}^{c} (B_i^k)^2 \sum_{i=1}^{c} v((x_i^k)^2) - \sum_{i=1}^{c} (A_i^k)^2 \sum_{i=1}^{c} v(x_i^k) \quad (12)$$

$$R^k = \sum_{i=1}^{c} (B_i^k)^2 \sum_{i=1}^{c} c(x_i^k, (x_i^k)^2) - \sum_{i=1}^{c} (A_i^k B_i^k) \sum_{i=1}^{c} v(x_i^k) \quad (13)$$

Variance of the k-th feature of i-th class

$$v(x_i^k) = \frac{\sum_{h=1}^{n_i^k} (x_{h,i}^k - \mu_i^k)^2}{n_i^k - 1} \quad (14)$$

The Squared variance of kth feature, i-th class

$$v((x_i^k)^2) = \frac{\sum_{h=1}^{n_i^k} ((x_{h,i}^k)^2 - \theta_i^k)^2}{n_i^k - 1} \quad (15)$$

Covariance between x_i^k and $(x_i^k)^2$ of the kth feature of the i-th class

$$c(x_i^k, (x_i^k)^2) = \frac{\sum_{h=1}^{n_i^k} (x_{h,i}^k - \mu_i^k)((x_{h,i}^k)^2 - \theta_i^k)^2}{n_i^k - 1} \quad (16)$$

2) Transformation procedure: Each feature is separately parameterized using Eq. (7) to Eq. (9). The following procedure is used for feature transformation:

a) Select the kth feature for which parameters are to be calculated.

b) Split the feature into training and testing data.

c) Square each element of the feature (x^k) and add it as a new feature (x^k)^2.

d) Using the training data, calculate the mean of the kth feature of the i-th class, the average of the kth feature, the variance of the kth feature of the i-th class, the squared variance of the kth feature of the i-th class, the covariance between x_i^k and $(x_i^k)^2$ of the kth feature of the i-th class, respectively, using Eq. (2), (14), (15), and (16) to obtain the optimal parameters, b, and c in Eq. (7) to Eq. (9).

e) Transform each feature element x_i^k , both training and testing, using Eq. (3).

C. Experimental Design

All 11 datasets were preprocessed, which included converting categorical data types to numerical and standardizing the datasets in the range between 1 and 2 [19]. Fig. 1 presents the experimental design where the data is split into training and testing using the Cross-validation technique with k = 10. The training data was used to obtain the quadratic function parameters as described in detail in Section III (B) (2), and the Box-Cox transformation parameter λ in the range of -5 to +5 as in Bicego and Baldo's study, which chose the best λ based on the highest fisher score. These parameters are used to transform training and testing data features based on quadratic and Box-Cox functions. The quality of the transformed training data features was measured using the Fisher score, Information gain, Gain ratio, Gini Decrease, Anova, Chi-square, ReliefF, and FCBF to ensure they were good inputs for training the two statistical machine learning algorithms LDA [70] and QDA[71]. The resulting ML model was tested with the transformed testing data to evaluate model performance. Model performance is measured by the accuracy (Acc) metric using Eq. (17).

$$Acc = \left(\frac{Tp + Tn}{Tp + Tn + Fp + Fn} \right) * 100 \quad (17)$$

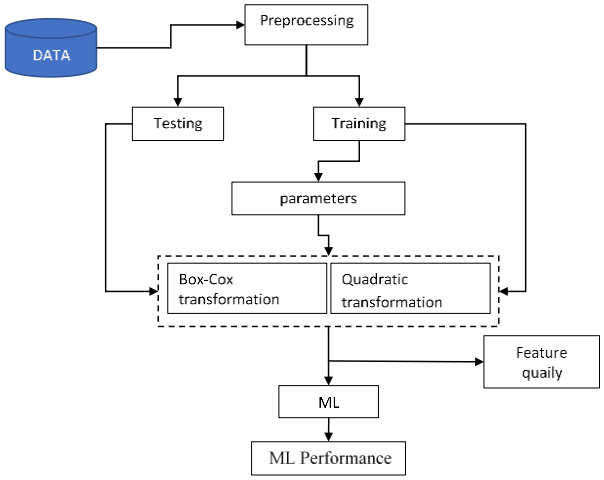


Fig. 1. Experimental design of quadratic and Box-Cox transformations.

where Tp, Tn, Fp, and Fn are components in the confusion matrix that respectively show the number of true positives, true negatives, false positives, and false negatives.

This study compares feature quality and ML performance before and after transformation to assess the efficacy of the proposed transformation, as well as comparisons with ML performance by other researchers. Seven metrics compare feature quality before and after transforming using quadratic and Box-Cox functions. The algorithm is trained using transformed training data and then tested using transformed testing data to determine ML performance. The results are compared with the performance before transformation. The accuracy of the proposed method is compared with the results of other researchers for the same dataset, including Huan Wan, et al 2018 [65], Jinsong Wang, et al 2020 [72], R Ksantini et al 2012 [73], Eslam Hamouda, et al 2021[47], Syed Muhammad Saqlain, et al 2019 [51], Jianbin Ma, et al 2019 [63], and Min Gan, et al 2021[74], and Peiyang Li, et al 2018 [75].

IV. Results and Discussion

A. Comparison of Feature Quality Before and After Transformation

The success of feature transformation is measured by comparing the quality of features before and after transformation. This research uses eight widely used feature quality measurement methods.

1) Fisher score: Since class separability is the basis for transforming features in this research, the higher the Fs, the better the feature quality. Fig. 2 presents the Fs of each feature from the d-3 (Ringnorm) dataset consisting of 20 features before and after transformation. As shown in Fig. 2, the Box-Cox and Quadratic transformations produce features with larger Fs than the original features, and the quadratic transformation produces better Fs improvement than the Box-Cox transformation on all features. Except for d-6 and d-9 datasets, the superiority of the quadratic transformation also occurs in other datasets, namely superiority in 41 out of 60 features, 25 out of 30 features, 3 out of 5 features, 9 out of 9 features, 8 out of 11 features, 38 out of 60 features, 37 out of 44 features, and 4 out of 4 features, respectively for d-1, d-2, d-4, d-5, d-7, d-8, d-10, and d-11. For the d-6 dataset of 22 features, the quadratic transform outperforms the Box-Cox in 11 features and vice versa in the other 11 features. For dataset d-9, the Box-Cox transform outperformed the quadratic transform for all features. In general, the quadratic transform produces more features with higher Fs for the whole dataset.

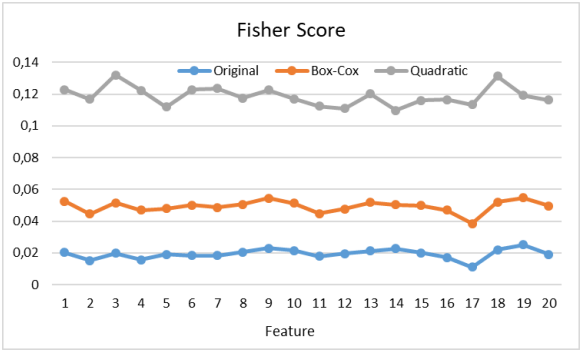


Fig. 2. Comparison of fisher score for d-3 (Ringnorm) dataset.

2) Information gain: Information Gain measurements for all datasets show that the Box-Cox transformation does not change the information gain value, meaning that the Box-Cox transformation does not increase the information gain value. On the other hand, except for datasets d-9 and d-11, the quadratic transformation generally produces features with greater information gain values, although this is not the case for every feature. Fig. 3 presents the information gain value of each feature for the d-3 (Ringnorm) dataset, which consists of 20 features. The increase in information gain in this dataset is observed for all features. An extreme case occurs in datasets d-9 and d-11, where the quadratic transformation does not increase the information gain value of each feature.

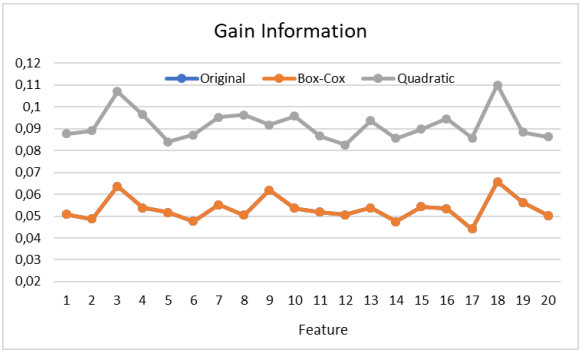


Fig. 3. Comparison of gain information for d-3 (Ringnorm) dataset.

3) Gain ratio: The Gain Ratio measurement shows similar results to the Information Gain, where the Box-Cox transformation does not change the Gain Ratio value. Except for d-9 and d-11 datasets that did not experience an increase in gain ratio in all features, the quadratic transformation generally increased the gain ratio value of a number of features in other datasets. The number of features that have

increased gain ratio values is different for each dataset, namely 29 out of 60, 13 out of 30, 20 out of 20, 2 out of 5, 4 out of 9, 8 out of 22, 3 out of 11, 15 out of 60, 18 out of 44 features, respectively for datasets d-1, d-2, d-3, d-4, d-5, d-6, d-7, d-8, d-10. Fig. 4 presents the Gain Ratio value of each feature for dataset d-3 (Ringnorm). It can be seen that the quadratic transformation increases the gain ratio of each feature.

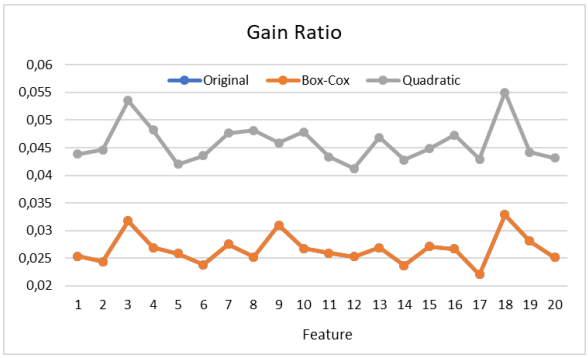


Fig. 4. Comparison of gain ratio for d-3 (Ringnorm) dataset.

4) Gini decrease: The Gini decrease measurement shows similar results to the Information Gain and Gain ratio, where the Box-Cox transformation does not change the Gini decrease value. Except for d-9 and d-11 datasets that did not experience an increase in gain ratio across all features, the quadratic transformation generally increased the Gini decrease value of a number of features in the other datasets. The number of features that have increased gain ratio values is different for each dataset, namely 27 out of 60, 13 out of 30, 20 out of 20, 2 out of 5, 5 out of 9, 9 out of 22, 5 out of 11, 15 out of 60, 17 out of 44 features, respectively for datasets d-1, d-2, d-3, d-4, d-5, d-6, d-7, d-8, d-10. Fig. 5 presents the Gini decrease value of each feature for dataset d-3 (Ringnorm). It can be seen that the quadratic transformation results in an increase in Gini decrease for all features.

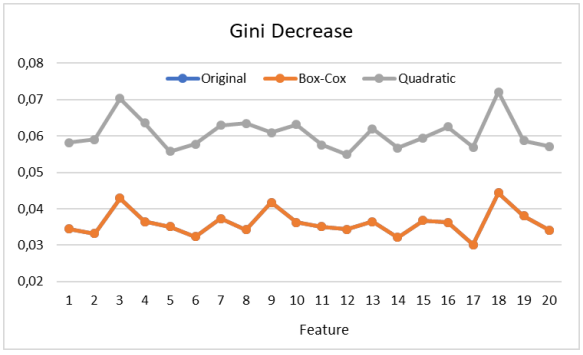


Fig. 5. Comparison of gini decreases for d-3 (Ringnorm) dataset.

5) Analysis of Variance (ANOVA): Except for the Hungarian and Balance datasets, ANOVA measurements show that the Box-Cox transformation improves most of the features in most datasets. The improvement in ANOVA values by Box-Cox transformation occurs in 58 out of 60, 25 out of 30, 20 out of 20, 3 out of 5, 7 out of 9, 19 out of 22, 60 out of 60, 31 out of 44, 4 out of 4 features for datasets d-1, d-2, d-3, d-4, d-5, d-6, d-8, d-10, d-11 respectively. The quadratic transformation gives slightly better ANOVA measurement results than the Box-Cox transformation. Except for dataset d-7, most datasets have an increase in ANOVA values. Even an increase in ANOVA values in each feature is observed in datasets d-3, d-8, d-9, and d-11. In the other six datasets, the increase in ANOVA values occurred in 59 out of 60, 27 out of 30, 3 out of 5, 8 out of 9, 19 out of 22, 30 out of 44 features, for datasets d-1, d-2, d-4, d-5, d-6, d-10, respectively. Fig. 6 presents the ANOVA values of each feature for dataset d-3 (Ringnorm) before and after undergoing Box-Cox and Quadratic transformations. It can be seen that the Quadratic transformation generally improves the ANOVA value better than the Box-Cox transformation.

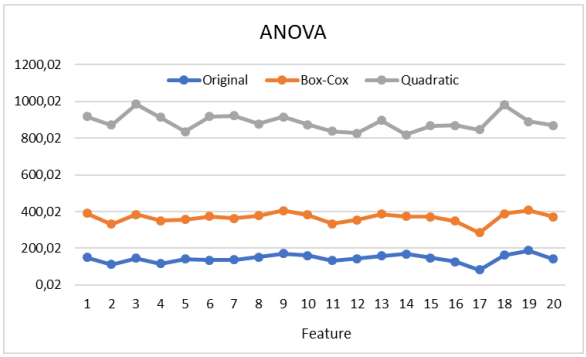


Fig. 6. Comparison of ANOVA for d-3 (Ringnorm) dataset.

6) Chi-Square: In Chi Square measurement in Box-Cox transformation, none of the features experienced changes in Chi Square value as produced by Information gain, Gain ratio, and Gini Decrease. For the quadratic transformation, except d-9 and d-11, the increase in Chi Square value occurred for 36 out of 60, 19 out of 30, 20 out of 20, 3 out of 5, 6 out of 9, 12 out of 22, 5 out of 11, 49 out of 60, 24 out of 44 features, respectively for datasets d-1, d-2, d-3, d-4, d-5, d-6, d-7, d-8, d-10. Fig. 7 presents the results of measuring the Chi-Square value of each feature for dataset d-3 (Ringnorm). It can be seen that the quadratic transformation results in an increase in the Chi-Square value of all features.

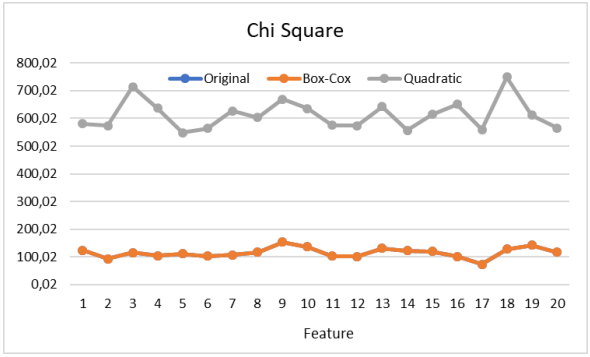


Fig. 7. Comparison of chi square for d-3 (Ringnorm) dataset.

7) ReliefF: Measurements using the ReliefF method provide varied results for both Box-Cox and Quadratic transformations, where in each dataset, features experience an increase in their ReliefF value. In Box-Cox transformation, the increase in reliefF value occurs in 35 out of 60, 28 out of 30, 2 out of 20, 3 out of 5, 8 out of 9, 19 out of 22, 4 out of 11, 29 out of 60, 2 out of 4, 25 out of 44, 3 out of 4 features, respectively for datasets d-1, d-2, d-4, d-5, d-6, d-7, d-8, d-9, d-10, d-11. The quadratic transformation generally gives slightly better ReliefF measurement results than the Box-Cox transformation. However, compared to the Box-Cox transformation, the number of features that have improved ReliefF values is less for datasets d-2, d-8, and d-10. However, the number of features that have improved ReliefF values is more in five datasets and the same in three other datasets. The increase in ReliefF value by Quadratic transformation occurs in 37 out of 60, 26 out of 30, 10 out of 20, 4 out of 5, 8 out of 9, 19 out of 22, 6 out of 11, 28 out of 60, 4 out of 4, 24 out of 44, 3 out of 4 features, respectively for datasets d-1, d-2, d-4, d-5, d-6, d-7, d-8, d-9, d-10, d-11. Fig. 8 presents the ReliefF value of each feature for dataset d-3 (Ringnorm) before and after undergoing Box-Cox and Quadratic transformations. On the d-3 dataset, the Quadratic transformation appears to increase the ReliefF value on ten features and decrease it on ten other features. In comparison, the Box-Cox transformation increases the ReliefF value on two features and decreases it on the other 18 features.

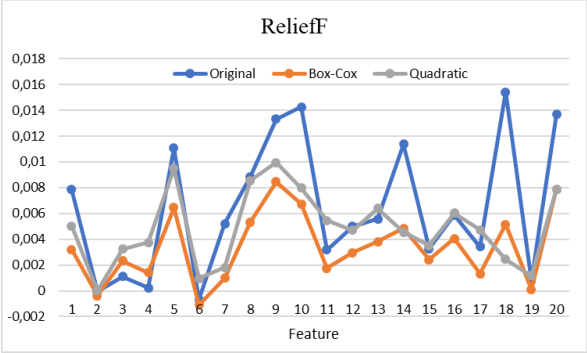


Fig. 8. Comparison of reliefF for d-3 (Ringnorm) dataset.

8) Fast Correlation Based Filter (FCBF): Measurement using FCBF gives very different results where Quadratic is better than Box-Cox regarding the number of features that have increased FCBF value. In quadratic transformation, out of 11 datasets, there are 10 datasets whose number of features has increased, where the increase in FCBF value occurs in 28 out of 60, 13 out of 30, 20 out of 20, 2 out of 5, 8 out of 9, 8 out of 22, 9 out of 11, 20 out of 60, 20 out of 44, 3 out of 4 features, respectively for datasets d-1, d-2, d-3, d-4, d-5, d-6, d-7, d-8, d-10, d-11. Except for datasets d-3, d-7, and d-11, the FCBF value from Box-Cox transformation has slightly increased in the other 8 datasets. The increase occurs in 2 out of 60, 3 out of 30, 3 out of 5, 1 out of 9, 2 out of 22, 2 out of 60, 4 out of 4, 1 out of 44 features, respectively, for datasets d-1, d-2, d-4, d-5, d-6, d-8, d-10. Fig. 9 presents the measurement results of each feature on dataset d-3 (Ringnorm), where all features have increased from 20 features.

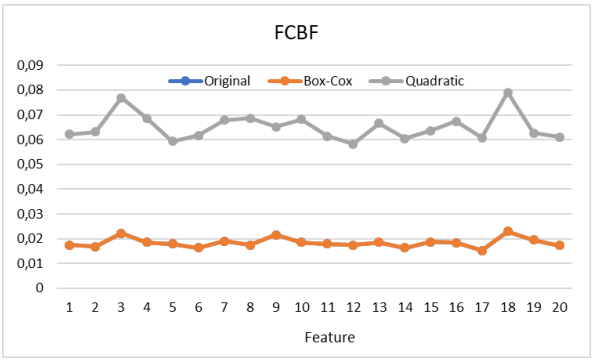


Fig. 9. Comparison of FCBF for d-3 (Ringnorm) dataset.

B. *Effects of Quadratic and Box-Cox Transformations on ML Performance*

This study uses two statistical-based machine learning algorithms, LDA and QDA. The performance results of each algorithm are presented in Tables II and III. Table II shows the comparison of LDA accuracy performance before and after Box-Cox and Quadratic transformations on 11 datasets. Experimental results on all datasets before and after quadratic transformation show an increase in accuracy performance. The highest increase in accuracy occurred on the Hayes Roth dataset, which was 0.268, and the lowest occurred on the Wisconsin dataset, which was 0.015. Experimental results on all datasets before and after the Box-Cox transformation also experienced an increase in accuracy performance. The highest increase in accuracy of 0.115 occurred on the Ringnorm dataset, and the lowest occurred on the WDBC dataset of 0.002. Compared to Box-Cox, Quadratic transformation generally produces a more significant increase in accuracy, ranging from 1.56% to 49.26%.

TABLE II. COMPARISON OF ACCURACY ALGORITHMS LDA BEFORE AND AFTER BOX-COX AND QUADRATIC TRANSFORMATION

dataset	Original	Box-Cox	Quadratic
Sonar	0.751	0.785	**0.834**
WDBC	0.956	0.958	**0.974**
Ringnorm	0.763	0.878	**0.941**
New Thyroid	0.934	0.939	**0.954**
Wisconsin	0.959	0.972	**0.974**
Parkinson	0.883	0.898	**0.908**
Hungarian	0.830	0.840	**0.848**
Splice	0.804	0.856	**0.884**
Balance	0.864	0.874	**0.912**
Spectf	0.746	0.765	**0.780**
Hayes Roth	0.544	0.606	**0.812**

Table III shows the comparison of QDA accuracy performance before and after Box-Cox and Quadratic transformations. Except for Ringnorm and Spectf datasets, where the accuracy is more or less the same, quadratic transformation improves QDA accuracy for all datasets. In comparison, Box-Cox transformation increases the accuracy on

seven datasets and decreases the accuracy on four datasets, namely Ringnorm, Balance, Spectf, and Hayes Roth. In general, both Box-Cox and Quadratic transformations improve accuracy on most datasets, whereas quadratic transformation improves accuracy on more datasets. Concurrent improvement in accuracy by Box-Cox and Quadratic transformations was observed in seven datasets, namely Sonar, WDBC, New Thyroid, Wisconsin, Parkinson, Hungarian, and Splice. On the Balance, Spectf, and Hayes-roth datasets, the Box-Cox transformation is shown to decrease accuracy, while the Quadratic transformation increases accuracy.

TABLE III. COMPARISON OF ACCURACY ALGORITHMS QDA BEFORE AND AFTER BOX-COX AND QUADRATIC TRANSFORMATION

dataset	Original	Box-Cox	Quadratic
Sonar	0.779	0.809	**0.818**
WDBC	0.956	0.961	**0.965**
Ringnorm	**0.979**	0.964	0.971
New Thyroid	0.967	0.985	**0.986**
Wisconsin	0.958	**0.972**	0.968
Parkinson	0.882	0.918	**0.923**
Hungarian	0.825	0.838	**0.844**
Splice	0.846	0.859	**0.889**
Balance	0.918	0.837	**0.963**
Spectf	**0.794**	0.791	**0.794**
Hayes Roth	0.649	0.609	**0.837**

C. Performance of LDA and QDA Based on Quadratic Transformation Compared to Other Methods

Table IV presents the performance comparison between LDA and QDA with methods from other researchers. It shows that the LDA algorithm in the 2nd column, six datasets have higher accuracy than other methods (OM) as listed in the 4th column of Table IV; they are Sonar, WDBC, Wisconsin, Parkinson, Hungarian, and Splice datasets. Compared to QDA, as shown in the 3rd column of Table 4, it indicates that QDA excels on seven datasets: Sonar, Ringnorm, New Thyroid, Parkinson, Splice, Balance, and Hayes Roth.

TABLE IV. COMPARISON OF THE OTHERS METHODS (OM) ACCURACY WITH QUADRATIC TRANSFORMATION

Dataset	LDA	QDA	OM
Sonar	**0.834**	0.818	0.768
WDBC	**0.974**	0.965	0.967
Ringnorm	0.941	**0.971**	0.953
New Thyroid	0.954	**0.986**	0.972
Wisconsin	**0.974**	0.968	0.970
Parkinson	0.908	**0.923**	0.862
Hungarian	**0.848**	0.844	0.845
Splice	0.884	**0.889**	0.868
Balance	0.912	**0.963**	0.939
Spectf	0.780	0.794	**0.827**
Hayes Roth	0.812	**0.837**	0.818

LDA and QDA outperformed other methods only on Sonar, Parkinson, and Splice datasets.

D. Discussion

The results of separability measurement using the Fisher score on each feature of the dataset before and after quadratic and Box-Cox transformation show that all features have increased separability. This result proves the success of the proposed method that aims to improve class separability. Compared to Box-Cox, the improvement in class separability of quadratic transformation is generally better. It happens because the quadratic transformation can change the order of the data that the Box-Cox does not have [20].

The results of measuring seven metrics show that the Box-Cox transformation cannot improve the quality of features in 4 matrices, namely Information gain, Gain ratio, Gini decrease, and Chi-Square. This finding indicates a relationship between transformations that do not change the order of the data and the ability of these transformations to improve the quality of the features of the four metrics above. This finding is reinforced by the results of the quadratic transformation on features that do not change the order of the data, where the feature quality also does not change. Thus, Box-Cox transformation is unsuitable for Information gain-based ML such as decision Tree and Random Fores. In contrast, quadratic transformation still has the opportunity to improve ML performance.

The quadratic transformation has been shown to improve separability, which in turn improves the performance of the LDA and QDA algorithms. Although the performance improvement varies for each dataset and algorithm used, it shows that the dataset's characteristics play a role. Likewise, the algorithm used where LDA provides a significant increase in accuracy compared to the QDA algorithm. It shows that the linear ML algorithm has an advantage over the QDA algorithm.

The quadratic transformation can work like the Box-Cox transformation in terms of transforming features to be more linearly separable between data groups, as shown by Bicego and Baldo. In some instances, the quadratic transformation can change the order of the data, which helps handle bimodal features, which Box-Cox lacks. The closed-form formulation to obtain the optimal quadratic parameters can be determined deterministically, whereas in Box-cox the optimal parameters are determined using Maximum Likelihood (MLE) or Grid search [14], [21]. It is clear that the computation time of the quadratic transform is more efficient.

Our research supports the findings of Bicego and Baldo that accuracy performance is more related to class separability than Gaussianity. We emphasize that from experimental results, using the quadratic transform where the separability effect is higher than Box-Cox also results in better accuracy. This result adds to the finding that Baldo's statement applies not only to the Box-Cox transform but also to the quadratic transform.

Based on the experimental results, we recommend the following for future research, namely:

- Quadratic transformation can detect the presence of bimodal distributed features by measuring the degree of change in the fisher score of features before and after

transformation. However, it needs to be researched more deeply on how much the right level of change is.

- Quadratic transformation can improve the Fisher score's ability to assess informative features.
- Overlapping features have a low Fisher score value. If this feature is transformed using a quadratic function, the new feature formed will also has a low Fisher score.
- Informative features with a high fisher score, when transformed using a quadratic function, the Fs value will not experience significant changes.

V. CONCLUSION

Feature engineering through quadratic transformation can be used to improve class separability. It can also be used to transform bimodal distributed data into unimodal. Unimodal features have a better fisher score than bimodal features, so if the dataset contains unimodal features, it can improve the performance of the ML algorithm. The proposed feature transformation method can improve the feature's fisher score and seven feature quality test metrics, i.e, Information Gain, Gen ratio, Gini Decrease, ANOVA, Chi-Square, ReliefF, and FCBF. Experimental results on 11 datasets using ML algorithms, namely LDA and QDA, show that feature transformation using quadratic functions can significantly improve the accuracy performance of ML algorithms.

REFERENCES

[1] Y. Wang, Z. Yang, and X. Yang, "Kernel-free quadratic surface minimax probability machine for a binary classification problem," Symmetry (Basel)., vol. 13, no. 8, 2021, doi: 10.3390/sym13081378.

[2] T. Verdonck, B. Baesens, M. Óskarsdóttir, and S. vanden Broucke, "Special issue on feature engineering editorial," Mach. Learn., no. 0123456789, 2021, doi: 10.1007/s10994-021-06042-2.

[3] M. Robnik Sikonja MarkoRobnik and friuni-ljsi IGOR Kononenko IgorKononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," Mach. Learn., vol. 53, pp. 23–69, 2003, [Online]. Available: http://lkm.fri.uni-lj.si/xaigor/slo/clanki/MLJ2003-FinalPaper.pdf

[4] A. Zheng and A. Casari, Feature Engineering for Machine Learning. 2018. doi: 10.1201/9781315181080.

[5] K. J. Kim and W. B. Lee, "Stock market prediction using artificial neural networks with optimal feature transformation," Neural Comput. Appl., vol. 13, no. 3, pp. 255–260, 2004, doi: 10.1007/s00521-004-0428-x.

[6] H. Zhuang, X. Wang, M. Bendersky, and M. Najork, "Feature Transformation for Neural Ranking Models," SIGIR 2020 - Proc. 43rd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., pp. 1649–1652, 2020, doi: 10.1145/3397271.3401333.

[7] W. Li and Z. Liu, "A method of SVM with normalization in intrusion detection," Procedia Environ. Sci., vol. 11, no. PART A, pp. 256–262, 2011, doi: 10.1016/j.proenv.2011.12.040.

[8] C. Feng et al., "Log-transformation and its implications for data analysis," Shanghai Arch. Psychiatry, vol. 26, no. 2, pp. 105–109, 2014, doi: 10.3969/j.issn.1002-0829.2014.02.

[9] C. A. Igbo and E. L. Otuonye, "The Result of a Square Root Transformation on the Error Part of the Additive Time Series Model," Am. J. Math. Stat., vol. 11, no. 4, pp. 73–93, 2021, doi: 10.5923/j.ajms.20211104.01.

[10] J. Raymaekers and P. J. Rousseeuw, "Transforming variables to central normality," Mach. Learn., no. May 2020, 2021, doi: 10.1007/s10994-021-05960-5.

[11] J. R. Berrendero and J. Cárcamo, Linear components of quadratic classifiers, vol. 13, no. 2. Springer Berlin Heidelberg, 2019. doi: 10.1007/s11634-018-0321-6.

[12] M. M. Elmorshedy, R. Fathalla, and Y. El-Sonbaty, "Feature Transformation Framework for Enhancing Compactness and Separability of Data Points in Feature Space for Small Datasets," Appl. Sci., vol. 12, no. 3, 2022, doi: 10.3390/app12031713.

[13] J. Pan, Y. Zhuang, and S. Fong, "The impact of data normalization on stock market prediction: Using SVM and technical indicators," Commun. Comput. Inf. Sci., vol. 652, pp. 72–88, 2016, doi: 10.1007/978-981-10-2777-2_7.

[14] A. Cheddad, "On Box-Cox Transformation for Image Normality and Pattern Classification," IEEE Access, vol. 8, pp. 154975–154983, 2020, doi: 10.1109/ACCESS.2020.3018874.

[15] X.-S. Si, T. Li, J. Zhang, and Y. Lei, "Nonlinear degradation modeling and prognostics: A Box-Cox transformation perspective," Reliab. Eng. Syst. Saf., vol. 217, p. 108120, 2022, doi: https://doi.org/10.1016/j.ress.2021.108120.

[16] H. Yu, P. Sang, and T. Huan, "Adaptive Box–Cox Transformation: A Highly Flexible Feature-Specific Data Transformation to Improve Metabolomic Data Normality for Better Statistical Analysis," Anal. Chem., vol. 94, no. 23, pp. 8267–8276, Jun. 2022.

[17] T. Zhan, "Log-based transformation feature learning for change detection in heterogeneous images," IEEE Geosci. Remote Sens. Lett., vol. 15, no. 9, pp. 1352–1356, 2018.

[18] F. Zhang, I. Keivanloo, and Y. Zou, "Data Transformation in Cross-project Defect Prediction," Empir. Softw. Eng., vol. 22, no. 6, pp. 3186–3218, 2017.

[19] C. C. Mason et al., "Bimodal distribution of RNA expression levels in human skeletal muscle tissue," BMC Genomics, vol. 12, no. February, 2011.

[20] M. Bicego and S. Baldo, "Properties of the Box–Cox transformation for pattern classification," Neurocomputing, vol. 218, pp. 390–400, 2016.

[21] L. Blum, M. Elgendi, and C. Menon, "Impact of Box-Cox Transformation on Machine-Learning Algorithms," Front. Artif. Intell., vol. 5, no. April, pp. 1–16, 2022.

[22] R. Van Der Heiden and F. C. A. Groen, "The Box-Cox metric for Nearest Neighbour classification improvement," Pattern Recognit., vol. 30, no. 2, pp. 273–279, 1997.

[23] M. M. Rahman, M. M. Hossain, M. K. Uddin, and A. K. Majumder, "Supervised Parametric Classification on Simulated Data via Box-Cox Transformation," Int. J. Adv. Sci. Tech. Res. Issue, vol. 3, no. 1, pp. 541–550, 2013.

[24] I. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry," Biometrika, vol. 87, no. 4, pp. 954–959, Dec. 2000, doi: 10.1093/biomet/87.4.954.

[25] D. Charte, I. Sevillano-García, M. J. Lucena-González, J. L. Martín-Rodríguez, F. Charte, and F. Herrera, "Slicer: Feature Learning for Class Separability with Least-Squares Support Vector Machine Loss and COVID-19 Chest X-Ray Case Study BT - Hybrid Artificial Intelligent Systems," in International Conference on Hybrid Artificial Intelligence Systems, 2021.

[26] Z. Liu, "Fast kernel feature ranking using class separability for big data mining," *J. Supercomput.*, vol. 72, no. 8, pp. 3057–3072, 2016.

[27] S. Li, H. Zhang, R. Ma, J. Zhou, J. Wen, and B. Zhang, "Linear discriminant analysis with generalized kernel constraint for robust image classification," *Pattern Recognit.*, vol. 136, p. 109196, 2023.

[28] M. A. Muharram and G. D. Smith, "Evolutionary feature construction using information gain and gini index," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3003, pp. 379–388, 2004.

[29] Y. Liu, X. Yi, R. Chen, Z. Zhai, and J. Gu, "Feature extraction based on information gain and sequential pattern for English question classification," *IET Softw.*, 2018.

[30] G. Zhang, J. Hou, J. Wang, C. Yan, and J. Luo, "Feature Selection for Microarray Data Classification Using Hybrid Information Gain and a Modified Binary Krill Herd Algorithm," *… Sci. Comput. Life …*, 2020.

[31] A. A. Nababan, O. S. Sitompul, and Tulus, "Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio," *J. Phys. Conf. Ser.*, vol. 1007, no. 1, 2018.

[32] H. Han, X. Guo, and H. Yu, "Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest," *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, vol. 0, pp. 219–224, 2016.

[33] H. Nasiri and S. A. Alavi, "A Novel Framework Based on Deep Learning and ANOVA Feature Selection Method for Diagnosis of COVID-19 Cases from Chest X-Ray Images," *Comput. Intell. Neurosci.*, vol. 2, 2022.

[34] K. J. Johnson and R. E. Synovec, "Pattern recognition of jet fuels: Comprehensive GC × GC with ANOVA-based feature selection and principal component analysis," *Chemom. Intell. Lab. Syst.*, vol. 60, no. 1–2, pp. 225–237, 2002.

[35] X. Jin, A. Xu, R. Bie, and P. Guo, "Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3916 LNBI, pp. 106–115, 2006.

[36] A. K. F. Dornaika, "A hybrid discriminant embedding with feature selection : application to image categorization," *Appl. Intell.*, 2020.

[37] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *J. Biomed. Inform.*, vol. 85, no. July, pp. 189–203, 2018.

[38] D. Gharavian, M. Sheikhan, and S. S. Ghasemi, "Combined classification method for prosodic stress recognition in Farsi language," Int. J. Speech Technol., vol. 21, no. 2, pp. 333–341, 2018.

[39] L. Gao, M. Ye, and C. Wu, "Cancer classification based on support vector machine optimized by particle swarm optimization and artificial bee colony," *Molecules*, vol. 22, no. 12, 2017.

[40] T. Luo, C. Hou, F. Nie, and D. Yi, "Dimension Reduction for Non-Gaussian Data by Adaptive Discriminative Analysis," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 933–946, 2019.

[41] Y. R. Guo, Y. Q. Bai, C. N. Li, Y. H. Shao, Y. F. Ye, and C. zi Jiang, "Reverse nearest neighbors Bhattacharyya bound linear discriminant analysis for multimodal classification," *Eng. Appl. Artif. Intell.*, vol. 97, no. October 2020, p. 104033, 2021.

[42] H. Wan, H. Wang, B. Scotney, J. Liu, and W. W. Y. Ng, "Within-class multimodal classification," *Multimed. Tools Appl.*, vol. 79, no. 39–40, pp. 29327–29352, 2020.

[43] M. A. Duarte-Mermoud, N. H. Beltrán, and M. A. Bustos, "Chilean wine varietal classification using quadratic Fisher transformation," *Pattern Anal. Appl.*, vol. 13, no. 2, pp. 181–188, 2010.

[44] H. Liu and H. Motoda, "Feature transformation and subset selection," *IEEE Expert*, vol. 13, no. 2, pp. 26–28, 1998.

[45] S. M. Bassir, A. Akbari, and B. Nassersharif, "An improved feature transformation method using mutual information," *Int. J. Speech Technol.*, vol. 17, no. 2, pp. 107–115, 2014.

[46] A. Noor, L. Ali, H. T. Rauf, U. Tariq, and S. Aslam, "An integrated decision support system for heart failure prediction based on feature transformation using grid of stacked autoencoders," *Measurement*, vol. 205, p. 112166, 2022.

[47] E. Hamouda, A. S. Abohamama, and M. Tarek, "Random Projection-Based Feature Transformation Using Metaheuristic Optimization Algorithm," *Arab. J. Sci. Eng.*, vol. 46, no. 9, pp. 8345–8353, 2021.

[48] Z. Deng, S. Wang, and F. L. Chung, "A minimax probabilistic approach to feature transformation for multi-class data," *Appl. Soft Comput. J.*, vol. 13, no. 1, pp. 116–127, 2013.

[49] M. Z. Abedin, G. Chi, M. M. Uddin, M. S. Satu, M. I. Khan, and P. Hajek, "Tax Default Prediction Using Feature Transformation-Based Machine Learning," *IEEE Access*, vol. 9, pp. 19864–19881, 2021.

[50] D. Yan, X. Zhou, X. Wang, and R. Wang, "An off-center technique: Learning a feature transformation to improve the performance of clustering and classification," *Inf. Sci. (Ny).*, vol. 503, pp. 635–651, 2019.

[51] S. M. Saqlain *et al.*, "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," *Knowl. Inf. Syst.*, vol. 58, no. 1, pp. 139–167, 2019.

[52] L. Sun, X.-Y. Zhang, Y.-H. Qian, J.-C. Xu, S.-G. Zhang, and Y. Tian, "Joint neighborhood entropy-based gene selection method with fisher score for tumor classification," *Appl. Intell.*, vol. 49, no. 4, pp. 1245–1259, 2019.

[53] C. Li and J. Xu, "Feature selection with the Fisher score followed by the Maximal Clique Centrality algorithm can accurately identify the hub genes of hepatocellular carcinoma," *Sci. Rep.*, vol. 9, no. 1, 2019.

[54] M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," *Genomics*, vol. 109, no. 2, pp. 91–107, 2017.

[55] M. N. K.P. and T. P., "Feature selection using efficient fusion of Fisher Score and greedy searching for Alzheimer's classification," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2021.

[56] J. Yang, Y. L. Liu, C. S. Feng, and G. Q. Zhu, "Applying the fisher score to identify Alzheimer's disease-related genes," *Genet. Mol. Res.*, vol. 15, no. 2, 2016.

[57] G. S. Saragih and Z. Rustam, "Support Vector Machine with Fisher Score Feature Selection to Predict Disease-Resistant Gene in Rice," *J. Phys. Conf. Ser.*, vol. 1108, no. 1, 2018.

[58] L. Sun, X. Zhang, Y. Qian, J. Xu, and S. Zhang, "Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification," *Inf. Sci. (Ny).*, vol. 502, pp. 18–41, 2019.

[59] I. Dagher, "Quadratic kernel-free non-linear support vector machine," *J. Glob. Optim.*, vol. 41, no. 1, pp. 15–30, 2008.

[60] Y. Tian, Z. Yong, and J. Luo, "A new approach for reject inference in credit scoring using kernel-free fuzzy quadratic surface support vector machines," *Appl. Soft Comput. J.*, vol. 73, pp. 96–105, 2018.

[61] M. A. Bustos, M. A. Duarte-Mermoud, and N. H. BeltrÁn, "Nonlinear feature extraction using fisher criterion," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 22, no. 6, pp. 1089–1119, 2008.

[62] S. S. Wirts, "Quadratic Formula: Revisiting a Proof Through the Lens of Transformations," 2020, [Online]. Available: http://arxiv.org/abs/2010.14251.

[63] J. Ma and G. Teng, "A hybrid multiple feature construction approach for classification using Genetic Programming," *Appl. Soft Comput. J.*, vol. 80, pp. 687–699, 2019.

[64] L. Li, L. Du, W. Zhang, H. He, and P. Wang, "Enhancing information discriminant analysis: Feature extraction with linear statistical model and information-theoretic criteria," *Pattern Recognit.*, vol. 60, pp. 554–570, 2016.

[65] H. Wan, H. Wang, G. Guo, and X. Wei, "Separability-Oriented Subclass Discriminant Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 409–422, 2018.

[66] X. Wang, S. Wang, Z. Huang, and Y. Du, "Condensing the solution of support vector machines via radius-margin bound," *Appl. Soft Comput.*, vol. 101, p. 107071, 2021.

[67] J. Ma and X. Gao, "Designing genetic programming classifiers with feature selection and feature construction," *Appl. Soft Comput. J.*, vol. 97, 2020.

[68] L. Ali, I. Wajahat, N. Amiri Golilarz, F. Keshtkar, and S. A. C. Bukhari, "LDA–GA–SVM: improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine," *Neural Comput. Appl.*, vol. 33, no. 7, pp. 2783–2792, 2020.

[69] R. Fu, M. Han, Y. Tian, and P. Shi, "Improvement motor imagery EEG classification based on sparse common spatial pattern and regularized discriminant analysis," *J. Neurosci. Methods*, vol. 343, no. June, p. 108833, 2020.

[70] Y. Aliyari Ghassabeh, F. Rudzicz, and H. A. Moghaddam, "Fast incremental LDA feature extraction," Pattern Recognit., vol. 48, no. 6, pp. 1999–2012, 2015, doi: 10.1016/j.patcog.2014.12.012.

[71] B. Ghojogh and M. Crowley, "Linear and Quadratic Discriminant Analysis: Tutorial," no. 4, pp. 1–16, 2019, [Online]. Available: http://arxiv.org/abs/1906.02590.

[72] J. Wang, J. Liao, and W. Huang, "A density-based maximum margin machine classifier," *Cluster Comput.*, vol. 23, no. 4, pp. 3069–3078, 2020, doi: 10.1007/s10586-020-03070-w.

[73] R. Ksantini and B. Boufama, "Combining partially global and local characteristics for improved classification," *Int. J. Mach. Learn. Cybern.*, vol. 3, no. 2, pp. 119–131, 2012, doi: 10.1007/s13042-011-0045-9.

[74] M. Gan and L. Zhang, "Iteratively local fisher score for feature selection," *Appl. Intell.*, vol. 51, no. 8, pp. 6167–6181, 2021, doi: 10.1007/s10489-020-02141-0.

[75] P. Li *et al.*, "Improved Graph Embedding for Robust Recognition with outliers," *Sci. Rep.*, vol. 8, no. 1, pp. 1–11, 2018, doi: 10.1038/s41598-018-22207-x.

Detecting Data Poisoning Attacks using Federated Learning with Deep Neural Networks: An Empirical Study

Hatim Alsuwat
Department of Computer Science, College of Computers and Information Systems
Umm Al-Qura University, Makkah, Saudi Arabia

***Abstract*—The advent of intelligent networks powered by machine learning (ML) methods over the past few years has dramatically facilitated various facets of human lives, including healthcare, transportation, and entertainment. However, the use of ML in intelligent networks raises serious concerns about privacy and security, particularly in the context of data poisoning attacks. In order to address these concerns, this research paper presents a novel technique for detecting data poisoning attacks in intelligent networks, focusing on addressing privacy and security concerns associated with the use of machine learning (ML) methods. The research combines federated learning and deep learning approaches to analyze network data in a distributed and privacy-preserving manner. The technique employs a federated neural network to identify malicious data by analyzing network traffic, leveraging the power of Bayesian convolutional neural networks for efficient and accurate detection. The research follows an empirical approach, conducting experimental analyses to evaluate the proposed technique's effectiveness in terms of network security and data classification. The results demonstrate significant performance, including high throughput, quality of service, transmission rate, and low root mean square error for network security. Furthermore, the technique achieves impressive accuracy, recall, precision and malicious data analysis for data detection. The findings of this research contribute to enhancing the security and integrity of intelligent networks, benefiting various stakeholders, including network administrators, data privacy advocates, and users relying on secure network communication.**

***Keywords*—Poisoning attacks; deep learning; network security; data classification; malicious data**

I. Introduction

In recent years, the widespread use of systems and the data they generate has increased significantly, thanks to rapid advancements in technology [1]. This has led to a surge in the velocity at which data is produced, enabling systems to access and utilize it without requiring detailed programming. As an application of artificial intelligence, machine learning (ML) techniques enable systems to produce meaningful results by learning data on their own [2]. These techniques are extensively applied in cybersecurity, where they are used to identify malware, malicious network traffic, and improper system behavior [3]. Commercial products, such as Exabeam, Fortscale, and E8 Security, leverage these and related ML techniques for cybersecurity.

However, to bypass such detection systems and compromise the security of critical areas by exploiting flaws in ML methodologies, attackers have resorted to deploying adversarial ML techniques. Adversarial ML is a strategy employed in the field of ML that aims to deceive methods using nefarious input in either training or decision-making time.

When building a machine learning algorithm, the first step is to collect data, such as a set of images for developing computer vision applications. Ideally, this data should be collected and labeled in a controlled and secure environment [4]. However, this is a time-consuming and costly operation that not all organizations and individuals can afford. Therefore, they sometimes collect data from the Internet or other untrusted sources. For example, when building security systems, users may download labeled data from external vendors, such as VirusTotal, for malware data annotation.

However, applying ML in Internet of Things (IoT) environments poses unique security challenges, as attackers may tamper with sensors and modify the training data. A poisoning attack, also known as a targeted misclassification or bad behavior assault, allows adversaries to significantly reduce overall performance, introduce backdoors and neural Trojans, and cause targeted misclassification or bad behavior [5].

The study of how adversarial approaches could exploit ML algorithms and the development of effective defenses against their exposure led to the creation of the discipline of adversarial machine learning (AML). AML has been extensively researched in various disciplines, including intrusion detection and picture categorization. However, IoT systems have not been thoroughly studied in this regard.

Although the prevalence of data-driven applications and our growing reliance on networked systems have many advantages, they have also raised serious security concerns [6]. Data poisoning attacks, in which malicious actors inject harmful data into the system to manipulate its behavior and compromise its performance, are a serious security risk. These attacks have the potential to have devastating effects, including incorrect decisions, privacy breaches, and possibly catastrophic outcomes in crucial systems like those that control finance, healthcare, and industry. Effective detection and mitigation of data poisoning attacks may not be possible with current security measures and data classification techniques. Furthermore, since sensitive data is frequently made available

to a single entity or server, centralized approaches to data analysis raise privacy issues. There is a growing interest in investigating decentralized and privacy-preserving techniques, like federated learning, which enables local data analysis while aggregating knowledge globally, to address these problems.

The main goal of this paper is to suggest a novel method for detecting data poisoning attacks with a focus on classifying malicious data using federated and deep learning techniques. The goal of the paper is to tackle the problem of spotting and countering malicious activity in networked systems while protecting data security and privacy. The proposed method enables decentralized data analysis and guarantees that sensitive data is stored and protected locally by using federated learning and a federated adversarial neural network. The analysis of harmful network data is further improved by using BCNN, producing more precise and trustworthy results.

The major contributions of this research study are as follows:

- The study suggests a novel method for identifying data poisoning attacks that focuses on the classification of malicious data. The suggested approach improves the capacity to recognize and counteract malicious activities within the network by utilizing federated and deep learning techniques.
- To analyze network data, spread across various participants, the research introduces the use of a federated adversarial neural network. With this strategy, sensitive information is stored locally, privacy is maintained, and effective analysis of malicious activity is still possible.
- The analyzed data is collected and processed using a cloud module. This federated learning system's participants can communicate with each other easily thanks to the cloud-based approach's efficient data handling.
- In this study, harmful network data is analyzed using a BCNN. The ability of BCNNs to capture model parameter uncertainty allows for more accurate analysis and classification of malicious data.
- The research makes use of real-world datasets, such as the Duchenne Smile Dataset, Product Dataset, and Sentiment Dataset, to show the effectiveness of the suggested attack approach.

The proposed method was chosen based on the distinct advantages of combining federated learning and deep learning approaches for effectively detecting data poisoning attacks in intelligent networks.

Acknowledging the limitations of existing methods in addressing data poisoning attacks in intelligent networks, such as scalability and sensitivity to biased data distributions, emphasizes the need for our proposed approach. By overcoming these constraints, our method offers a compelling alternative to enhance the effectiveness of detecting and mitigating such attacks.

The rest of this paper is organized into four sections. In Section II, we provide a comprehensive review of the existing literature, focusing on data poisoning attacks and their detection. Section III describes the system model used in this research and the architecture of our proposed technique. Section IV presents the results of our experimental analysis, which evaluates the effectiveness of our proposed technique. Finally, in Section V, we summarize our research and its contributions, discuss the implications of our findings, and identify areas for future work.

II. Literature Review

In recent years, the potential threats posed by adversarial machine learning (AML) have been widely studied by researchers. In this section, we provide a comprehensive review of the existing literature on data poisoning attacks and their detection techniques in machine learning.

In the realm of machine learning applications, the data generated for training and testing models is susceptible to manipulation by malicious actors who can gain control over a multitude of devices [7]. Biggio et al. [8] conducted the first systematic poisoning assault against the linear regression method by taking control of many devices, and introduced the TRIM algorithm, which is a more potent method than conventional methods for identifying poisoning spots on training data. Khalid et al. [9] highlighted potential AML attacks and training data poisoning risks, and provided examples of these assaults, including a less damaging training data poisoning attack. The study in [10] proposed a data poisoning attack that modifies labels of labeled data and affects machine learning systems' capacity to categorize data. To guarantee label clearing against this assault, they then put forth a defense method based on the k-nearest neighbors (K-NN) algorithm.

Adversarial attacks are a critical threat to the integrity of machine learning models, and adversaries can manipulate the data generated for these applications by commandeering multiple devices [11]. Within this context, the TRIM algorithm proposed in [12] has been shown to be a more effective and powerful technique for identifying poisoning points in training data, and it was the first systematic attack against the linear regression method. In [13], the author identified potential adversarial machine learning (AML) attacks and risks associated with training data poisoning, including a less harmful attack on training data. Additionally, [14] described a data poisoning attack that modifies the labels of labeled data and impairs the ability of machine learning systems to classify data. To combat this attack, the authors proposed a defense method based on the k-nearest neighbors (K-NN) algorithm. It is highly improbable for training data to represent all possible scenarios, and "adversarial areas" near the decision boundary are particularly vulnerable locations for machine learning models. As a result, adversaries may use trial and error or reverse engineering to uncover "adversarial samples" that are not covered by the training data. Indeed, adversaries can use trial and error or reverse engineering to uncover "adversarial samples" and deceive the model, endangering its integrity. These evasion attacks are experimental and frequently used [15]. For instance, creator [16] utilized a generative network

called Malware-GAN to make ill-disposed malware samples for a black-box classifier, causing the classifier to fail after the assault.

An attack is viewed as causal when an attacker approaches training data and is allowed to harm it. The author in [17] showed that a peculiarity identification strategy on network traffic that had been polluted by refuse traffic infusion enhanced the bogus negative rate to 28 percent for single preparation period harming and to more than 70% for multi-preparing period harming. They also presented a cure strategy that can reject harmful preparation information and is less vulnerable to exceptions for extensive inconsistency detection. The study in [18] proposed the RONI safeguard strategy, which was effective, but had limitations in that it must be tested and trained on spam email data. Furthermore, it could potentially dispose of important information from training data, requiring further examination.

The authors in [19] endeavored to address a limitation that had been encountered by several previous studies. In addition, the authors in [20] proposed a detrimental attack that has the potential to bypass current safeguards with ease. The attack was subjected to testing against a range of hypothetical adversaries, providing valuable insights into its efficacy. The study in[21] also suggested a detrimental attack that utilizes a generative approach to expedite the generation of manipulated data by leveraging the gradient of the model. These advances in adversarial machine learning highlight the urgent need for developing more robust and effective techniques to detect and mitigate these attacks.

In the literature, several studies have proposed techniques to detect data poisoning attacks and other adversarial attacks in machine learning. Data poisoning attack detection tools currently available have some drawbacks and restrictions. Some defense strategies, such as the K-NN algorithm, aren't robust enough to handle complex data poisoning attacks, leaving room for attackers to get around these strategies and successfully poison the training data. When used on contaminated training data, some detection techniques may produce a high percentage of false negatives, failing to accurately detect some cases of data poisoning. Additionally, some current solutions might be effective against the kinds of attacks but fall short when faced with fresh or unexpected attack patterns, which restricts their applicability in real-world situations. Additionally, some proposed techniques are less applicable to a variety of real-world datasets because they rely on data for testing and training, such as spam email data. Additionally, some techniques are impractical for use in real-world applications because of their inability to scale effectively to large datasets or distributed environments. Additionally, although BCNNs demonstrate promise in capturing uncertainty and identifying adversarial samples, their ability to estimate uncertainty may not be accurate enough to defend against all possible data poisoning attacks. Finally, the limited collaborative defense mechanisms used in current methods fail to fully capitalize on the benefits of federated learning and distributed learning for improved detection. Together, these flaws highlight the need for a more thorough and potent method to deal with the constraints imposed on current solutions. The proposed method integrates federated learning, adversarial neural networks, and BCNNs to close this research gap. It hopes to accomplish this by developing a more reliable and scalable method of identifying data poisoning attacks. The proposed method seeks to improve the defense against data poisoning attacks through careful experimental analysis, ultimately advancing the field of adversarial machine learning research.

III. System Model

This section presents a novel technique for detecting data poisoning attacks based on federated and deep learning techniques. The overview of the proposed approach is shown in Fig. 1. By randomly flipping labels in a section of the training dataset, the data poisoning process creates poisoned datasets with varying poisoning rates that include both legitimate and adversarial samples. The method uses an adversarial neural network integrated with a federated learning approach to counter these poisoning attacks. Participants (clients) in this collaborative setting use local datasets to jointly train a global model. As a result of the inclusion of adversarial elements in the learning process, the model is better equipped to fend off poisoning attacks during the federated learning procedure. To further improve model robustness, the proposed technique makes use of BCNNs. The ability of BCNNs to capture prediction uncertainty allows for more accurate detection of potential adversarial samples. Each of the poisoned datasets is used to train a separate BCNN during the phase of model training and evaluation. The effectiveness of both the global model and the BCNNs is then evaluated using results from a shared test dataset. Analyzing the BCNN predictions' levels of uncertainty on the test dataset is a step in the process of detecting data poisoning attacks. The method effectively identifies potential data poisoning attacks by establishing an uncertainty threshold. The detailed description of each step of the proposed approach is presented in the subsequent subsections.

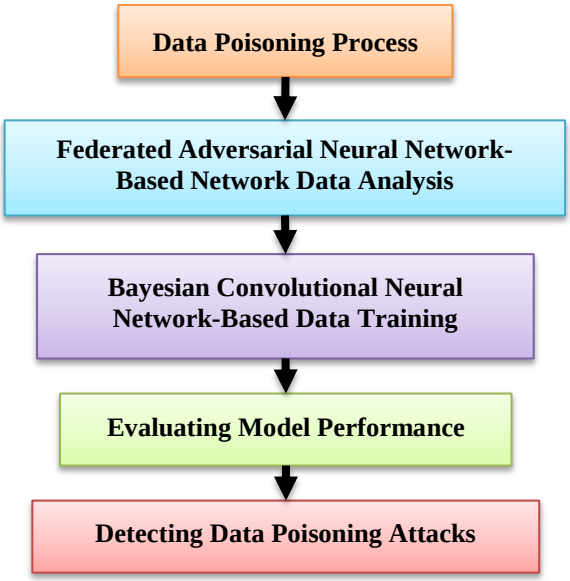


Fig. 1. Overview of the proposed approach.

A. Data Poisoning Process

To simulate the information poisoning attack on data classes, training datasets were created. To prevent a bias towards poisoning a significant amount of either normal data or attack data, normal observations were randomly selected, and most of the normal traffic observations were pooled. After randomizing the data, a Python script was used to determine the number of labels to flip based on a specified rate. To demonstrate the impact of data poisoning on classifiers, the data was poisoned at four rates (rpoison), as outlined in Algorithm 1.

Algorithm 1: Data Poisoning Process

Input: Sanitized Training Dataset (D^5)
Output: Poisoned Training Datasets $\left(D_5^P, D_{10}^P, D_{20}^{P_2}, D_{30}^P\right)$
Steps:
Randomise D^5 observations should not be biased by poisoning either normal or attack observations.
For every rate rpoison of data poisoning [0.05,0.1,0.2,0.3]
Evaluate a number of labels to flip Lpoison=D5*rpoison
Flip Lpoison labels within Ds
End for
Return poisoned training datasets $D_5^P, D_{10}^P, D_{15}^P, D_{30}^P$

B. Federated Adversarial Neural Network-based Network Data Analysis

Participants in federated learning may not always have the same learning objectives or method structures. The central server sends the most recent global model parameters to the chosen participants (mt) at the beginning of each communication round. Then, using the relevant local data, these participants go on to update and train their local models. Each participant uploads their updated model to the central server following the local training process. The central server then averages the models that were uploaded and incorporates the resulting information into the central model. This update procedure is implemented as shown in Eq. (1), ensuring that the central model gains from the group learning of all participants while maintaining the security and privacy of the data. The federated learning approach is a flexible and strong framework that can be applied to various scenarios because it allows participants to maintain their individuality when defining their learning objectives and selecting the best method structures.

$$M_{t+1} = M_t + \frac{1}{m_t}\sum_{k=1}^{m_t} u_t^k \tag{1}$$

In Eq. (1), u_t^k represents the method updates submitted by the k^{th} participant, and Mt represents the current global method at the t^{th} iteration. A federated learning system can achieve high accuracy when users download the same method with the same initialization, which is averaged by the central method with all valid uploads. We now introduce a new method, presented in Eq. (2), for training supervised federated learning models.

$$f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x_i), \psi(x) := \frac{1}{2n}\sum_{i=1}^{n} \|x_i - \bar{x}\|^2 \tag{2}$$

where, $\lambda \geq 0$ is a penalty specification, $x := (x_1, x_2, \dots, x_n) \in \mathbb{R}^{nd}$ are local methods, and $x := \frac{1}{n}\sum_{i=1}^{n} \bar{x}_i$ is average of local methods. Since Eq. (2) has a unique solution, which we designate by Eq. (3), F is strongly convex due to assumptions on fi that we will make.

Here in Eq. (2), $\lambda \geq 0$ is a penalty specification, $x := (x_1, x_2, \dots, x_n) \in \mathbb{R}^{nd}$ represents the local methods, and $x := \frac{1}{n}\sum_{i=1}^{n} \bar{x}_i$ is the average of the local methods. Since Eq. (2) has a unique solution, which we designate as Eq. (3), F is strongly convex due to the assumptions we make on fi.

$$x(\lambda) := (x_1(\lambda), \dots, x_n(\lambda)) \in \mathbb{R}^{\mathrm{nd}} \tag{3}$$

We further let $\bar{x}(\lambda) := \frac{1}{n}\sum_{i=1}^{n} x_i(\lambda)$. We now provide a statement regarding the new formulation's justification. Let's now examine the limit case $\lambda \to \infty$. The ideal local models should be forced to be mutually identical by this limit case while minimising the loss f, according to intuition. This limit situation will specifically be solved using Eq. (4).

We also define $\bar{x}(\lambda) := \frac{1}{n}\sum_{i=1}^{n} x_i(\lambda)$. We now provide a statement regarding the justification for the new formulation. Let us consider the limit case $\lambda \to \infty$. In this limit, the ideal local models should be forced to be identical to each other while minimizing the loss function f, according to intuition. This limit situation is specifically solved using Eq. (4).

$$\min\{f(x): x_1, \dots, x_n \in \mathbb{R}^d, x_1 = x_2 = \cdots = x_n\} \tag{4}$$

Eq. (4) is the equivalent global formulation. Therefore, we define xi(∞) as the optimal solution to Eq. (4) for each i, and let x(∞) := (x1(∞), . . . , xn(∞)).

For vectors $x = (x_1, \dots, x_n) \in \mathbb{R}^{nd}$ and $y = (y_1, \dots, y_n) \in \mathbb{R}^{nd}$, we define the standard inner product and norm as follows: $\langle x, y\rangle := \sum_{i=1}^{n} \langle x_i, y_i\rangle, \| x \|^2 := \sum_{i=1}^{n} \|x_i\|^2$. Note that the separable structure of f implies that $((\nabla f(x))_i = \frac{1}{n}\nabla f_i(x_i))$, i.e., $\nabla f(x) = \frac{1}{n}\left(\nabla f_1(x_1), \nabla f_2(x_2), \dots, \nabla f_n(x_n)\right)$.

Furthermore, note that f is L_f-smooth with with $L_f := \frac{L}{n}$ and μ_f-strongly convex with $\mu_f := \frac{n}{n}$. Clearly, ψ is convex by construction, and it is given that ψ is L_ψ-smooth with $L_\psi = \frac{1}{n}$. We can observe that $(\nabla\psi(x))_i = \frac{1}{n}(x_i - \bar{x})$, which, in turn, implies by Eq. (5), (6), and (7).

$$\psi(x) = \frac{n}{2}\sum_{i=1}^{n} \|(\nabla\psi(x))_i\|^2 = \frac{n}{2} \| \nabla\psi(x) \|^2 \tag{5}$$

$$\psi(x(\lambda)) \leq \frac{f(x(\infty)) - f(x(0))}{\lambda} \tag{6}$$

$$f(x(\lambda)) \leq f(x(\infty)) \tag{7}$$

For every $\lambda > 0$ and $1 \leq I \leq n$, we have by (8):

$$x_i(\lambda) = \bar{x}(\lambda) - \frac{1}{\lambda}\nabla f_i(x_i(\lambda)) \tag{8}$$

We have $\sum_{i=1}^{n} \nabla f_i(x_i(\lambda)) = 0$. By subtracting a multiple of the local gradient from the average model, the best local models Eq. (5) can be obtained. Note that at optimality, the local gradients always add up to zero. This is clearly true for $\lambda = 0$, but it is less clear that this is true for $\lambda = \infty$, or for any $\lambda > 0$.

Let P(z):=$\frac{1}{n}\sum_{i=1}^{n} f_i(z)$. Then, according to Eq. (9), x(∞) is the unique minimizer of P.

$$\| \nabla P(\bar{x}(\lambda)) \|^2 \leq \frac{2L^2}{\lambda}(f(x(\infty)) - f(x(0))) \quad (9)$$

If α ⩽ 2L , we then by (10) have the following.

$$\mathbb{E}\left[\|x^k - x(\lambda)\|^2\right] \leq \left(1 - \frac{\alpha\mu}{n}\right)^k \|x^0 - x(\lambda)\|^2 + \frac{2na\sigma^2}{\mu} \quad (10)$$

where, $\mathcal{L} := \frac{1}{n}\max\left\{\frac{L}{1-p}, \frac{\lambda}{p}\right\}$ and by (11), we have the following.

$$\sigma^2 := \frac{1}{n^2}\sum_{i=1}^{n}\begin{pmatrix} \frac{1}{1-p}\|\nabla f_i(x_i(\lambda))\|^2 \\ +\frac{\lambda^2}{p}\|x_i(\lambda) - \bar{x}(\lambda)\|^2 \end{pmatrix} \quad (11)$$

Let us determine the values of p and α that lead to the fastest rate for pushing the error within a $\left(\mathcal{O}(\varepsilon) + \frac{2na\sigma^2}{\mu}\right)$ - neighborhood of the optimum. In other words, we aim to achieve Eq. (12).

$$\mathbb{E}\left[\|x^k - x(\lambda)\|^2\right] \leq \varepsilon\|x^0 - x(\lambda)\|^2 + \frac{2\mathrm{nn}\,\sigma^2}{\mu} \quad (12)$$

The parameter $p^* = \frac{\lambda}{L+\lambda}$ reduces the predicted number of communications for attaining as well as the number of repetitions. The optimal expected number of communications is $2\frac{L+\lambda}{\mu}\log\frac{1}{\varepsilon}$, while best number of iterations is $2\frac{L+\lambda}{\mu}\log\frac{1}{\varepsilon}$. We employ relativistic average discriminator D_{Ra} to render the output image virtually identical to the original. According to Eq. (13), the objective functions are as follows:

$$\begin{aligned} \mathcal{L}_{Ra,D} &= -\mathbb{E}_x[\log(D_{Ra}(x))] \\ &\quad - \mathbb{E}_{x,v,}[\log(1 - D_{Ra}(G(x,v,c)))] \\ \mathcal{L}_{Ra\_G} &= -\mathbb{E}_{x,0,\mathcal{C}}[\log(D_{Ra}(G(x,v,c)))] \\ &\quad - \mathbb{E}_x[\log(1 - D_{Ra}(x))] \\ D_{Ra}(x) &= \mathrm{sigmoid}\big(H(x) - \mathbb{E}_{x,v,c}[H(G(x,v,c))]\big) \\ D_{Ra}(G(x,v,c)) &= \mathrm{sigmoid}\begin{pmatrix} H\big(G(x,v,c)\big) \\ -\mathbb{E}_x[H(x)] \end{pmatrix} \end{aligned} \quad (13)$$

The parameter $p^* = \frac{\lambda}{L+\lambda}$ reduces the predicted number of communications required to achieve the desired accuracy, as well as the number of repetitions needed. The optimal expected number of communications is $2\frac{L+\lambda}{\mu}\log\frac{1}{\varepsilon}$, while the optimal number of iterations is $2\frac{L+\lambda}{\mu}\log\frac{1}{\varepsilon}$.

To make the output image virtually identical to the original, we employ the relativistic average discriminator D_{Ra} . According to Eq. (13), the objective functions are as follows:

The output of the non-changed layer is denoted as H(•). The probability that certifies the real image as genuine is higher than the probability that certifies the generated image as genuine. This can be improved by minimizing the loss function $\mathcal{L}_{Ra,D}$.

To further reduce the loss, we subject the generator to a cycle consistency loss, which is described by Eq. (14) as follows:

$$\mathcal{L}_{cyc} = \mathbb{E}_{x,v,c}[\| x - G(G(x,v,c),v,1-c) \|_1] \quad (14)$$

To identify the source of the image, we add a helper classifier called D_{ind} on top of the discriminator network. According to Eq. (15), the loss function for the image attribution model is as follows:

$$\begin{aligned} \mathcal{L}_{\text{ind}} = \\ -\mathbb{E}_{x,t}\big[\log\big(D_{\text{ind}}(t = 01 \mid x)\big)\big] - \quad \mathbb{E}_{x,v,c,t}\big[\log\big(D_{\text{ind}}(t = 01 \mid \\ G(x,v,c))\big)\big] \end{aligned} \quad (15)$$

The picture producing model fundamentally affects the unraveling organization (c) since it is a common organization, and picture interpretation strategy utilizes essentially less examples than the picture age model does. We integrate the accompanying matched antagonistic misfortune condition (16) to more likely guarantee the fitting of the picture interpretation model:

Since the image generation model is a shared network, it significantly affects the decoding network (c). Moreover, the image attribution method uses significantly fewer samples than the image generation model. To better ensure the fitting of the image attribution model, we integrate the following paired adversarial loss condition as shown in Eq. (16).

$$\begin{aligned} \mathcal{L}_{pis} = -\mathbb{E}_{x_0,x_v}\left[\log\left(D_{pis}(x_0,x_v)\right)\right] - \mathbb{E}_{x,v\mathcal{C}}\big[\log\big(1 - \\ D_{pis}(x,G(x,v,c))\big)\big] \end{aligned} \quad (16)$$

In this scenario, D_{pis} is used to determine if two images belong to the same class. Our objective is to translate x_0 into an output image y that contains variation v, for input image x_0 and action (v, c = 01). Moreover, our goal is to remove variation v from input image xv using the action (v, c = 10).

To achieve this, we add an additional classifier called D_{var} on top of the discriminator network to identify different types of image variations. The classification loss during training of the discriminator network is given by Eq. (17).

$$\mathcal{L}^r_{\text{var}} = -\mathbb{E}_{x,v}[\log(D_{var}(v \mid x))] \quad (17)$$

Discriminator network may categorize real image x into variant type v by minimizing formula. Classification loss during training the generator network is as shown in Eq. (18)

The discriminator network can categorize the real image x into variant type v by minimizing the formula mentioned above. The classification loss during training of the generator network is given by Eq. (18).

$$\mathcal{L}^f_{\text{var}} = -\mathbb{E}_{x,v,c}[\log(D_{var}(v \mid x))] \quad (18)$$

The first condition in Eq. (18) states that the image produced by adding variation v to the input image x_0 should be accurately classified into class v. The second condition states that the image produced by removing variation v from the paired image xv should be classified into class v.

The values of x_m range from 0 to 1. The following Eq. (19) can be used to obtain the final output image.

$$x_{\text{out}} = x + (x_t - x) \odot x_m \tag{19}$$

$\odot$ element-wise product is located. By using Eq. (20), we add the next restriction for the mask x_m:

$$\mathcal{L}_{\text{mask}} = \left(\frac{1}{W}\sum_k |x_m[k]|\right)^2 \tag{20}$$

Here, W represents the number of pixels, and $x_m[k]$ refers to the k^{th} pixel of x_m. The formula shown above encourages minimizing alterations to the source image. Based on the foregoing discussion, the overall loss of the image translation model is given by Eq. (21):

$$\mathcal{L}_D = \mathcal{L}_{\text{Ra\_}D} + \lambda_{\text{pis}}\mathcal{L}_{\text{pis}} + \lambda_{\text{var}}\mathcal{L}_{\text{var}}^r + \lambda_{\text{ind}}\mathcal{L}_{\text{ind}}$$
$$\mathcal{L}_G = \mathcal{L}_{Ra\_G} - \lambda_{pis}\mathcal{L}_{\text{pis}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{var}}\mathcal{L}_{\text{tar}}^f + \lambda_{\text{ind}}\mathcal{L}_{\text{ind}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} \tag{21}$$

The hyperparameters $\lambda_{\text{pis}}, \lambda_{\text{var}}, \lambda_{\text{ind}}, \lambda_{\text{cyc}}$, and λ_{mask} control the relative significance of each term in Eq. (21) as outlined in Algorithm 2.

Algorithm 2: FANN

Input: K clients are indexed by k, C is client fraction, the T communication rounds are indexed by t, B is local minibatch size, E is number of local epochs, and η is learning rate, PGD Attack $A_{s,\epsilon,\alpha}$: where s, ϵ, α are number of PGD steps, perturbation ball size, step size, r is the adversarial ratio, q is the scale factor.

Output: The global model θ.

On server

Initialize θ_0

For every round $t = 1,2, \ldots \ldots T$ do

$$m \leftarrow \max(1, CK)$$
$$S_t \leftarrow (random\ set\ of\ m\ clients)$$

For every client $k \in S_t$ in parallel do

$$\theta_{t+1}^k \leftarrow client\ update\ (k, \theta_t)$$

End for

$$\theta_{t+1}^k \leftarrow FedWAvg\ (\theta_t, \{\theta_{t+1}^k\} k \in S_t)$$

End for

Return θ_{t+1}

Client update (k,θ)

$$E \leftarrow split\ the\ training\ data\ into\ batches\ of\ size\ B$$

For every local epoch I from 1 to E do

For batch $b \epsilon B$ do

$$n_{adv} \leftarrow r.B$$
$$b_{adv} \leftarrow (random\ set\ \in b\ of\ n_{adv}\ samples)$$
$$b_{nat} \leftarrow (set\ of\ B) - n_{adv} samples$$
$$b \leftarrow b_{bat} U b_{adv}$$

End for

End for

Return θ

C. Bayesian Convolutional Neural Network-based Data Training

BCNNs are a type of neural network that combines the CNN architecture and Bayesian inference principles to model uncertainty in deep learning tasks. In contrast to conventional CNNs, which provide point estimates of the model parameters, BCNNs estimate the model posterior distribution over the parameters, providing a principled method for dealing with model uncertainty. This feature is especially helpful when there is little or noisy data available, enabling more accurate predictions. BCNNs also allow for incorporating prior information and hypotheses, which can improve model performance in challenging real-world datasets. Additionally, BCNNs provide a natural method for model averaging, improving the generalizability of the model. In our work on identifying data poisoning attacks, BCNNs' uncertainty estimation is essential, as it can highlight areas of high ambiguity and possible adversarial inputs, resulting in a more accurate identification of such attacks.

Bayesian neural networks train a model by inferring the model posterior. However, accurate inference of the model posterior is computationally demanding, and even for moderately sized models, it can become intractable. Therefore, the model posterior is usually approximated. One popular and successful method for approximating the model posterior is variational inference. Fig. 2 provides an overview of the BCNN Architecture.

The BCNN architecture process is shown in Fig. 3. The "Start" symbol marks the beginning of the process at the top. Taking input data, which stand for the input set and the corresponding output set, respectively, is the first step. The next step in the flowchart is the "Feature Extraction" module. Utilizing techniques like convolution, non-linear transformations (relu), max-pooling, and local normalization, features are in this case extracted from the input data. The "Feature Selection" module is the next step in the flowchart after feature extraction. To further hone the extracted features, additional feature selection is carried out in this step using non-linear transformations (relu). The "Prediction" module is the next step in the process, where the final output probabilities are computed.

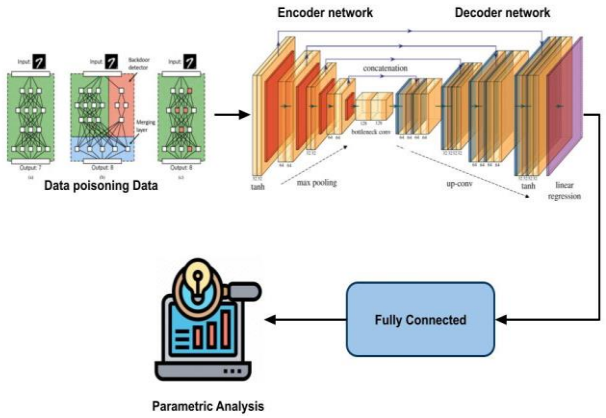


Fig. 2. Overview of the BCNN architecture.

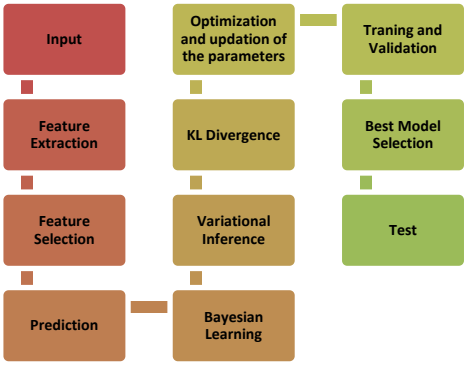


Fig. 3. Flowchart of BCNN architecture process.

The softmax operation provides a probability distribution for each output class C. The flowchart then moves to the "Bayesian Learning" phase, where variational inference is used to obtain the model posterior by approximating it with the variational distribution. The "KL Divergence" step computes the KL divergence, which is reduced through optimization of the model parameters W and b to increase the log evidence lower bound. The "Training and Validation" phase trains the model on the training set and assesses its performance on the validation set after each epoch. The "Best Model Selection" step chooses the model with the best validation performance. The "Test" step tests the chosen model on the test set to evaluate its final performance metrics. The "End" symbol marks the end of the process. The detailed process is presented in this section.

Given the input set $X = x_1, x_2, \ldots, x_N$ and the corresponding output set $y = y_1, y_2, \ldots, y_N$, the function $f(X) = y$ estimates the output y from the inputs X. Bayesian learning provides a principled approach to obtain the model posterior $p(f|X, y)$. To calculate the posterior, two components are required. First, a prior distribution $p(f)$ that captures a prior belief about the estimator functions. Second, a likelihood function $p(y|f, X)$ that indicates how likely it is for the model f to predict the output y given the observations X. More specifically, given an unseen data point $(x*, y*)$, the posterior is obtained by integrating over all possible estimator functions f that are parametric models with a parameter set θ, as shown in Eq. (22):

$$p(y^* \mid \mathrm{x}^*, \mathrm{X}, \mathrm{y}) = \int p(y^* \mid f)p(f \mid \mathrm{x}^*, \mathrm{X}, \mathrm{y})df = \int p(y^* \mid f)p(f \mid \mathrm{x}^*, \theta)p(\theta \mid \mathrm{X}, \mathrm{y})df d\theta \quad (22)$$

The integral in Eq. (22) is intractable because the distribution $p(\theta|X, y)$ is intractable. Therefore, the variational approach is to approximate $p(\theta|X, y)$ with a variational distribution $q(\theta)$. The candidate $q(\theta)$ should be as similar as possible to the original intractable distribution. The similarity between $p(\theta|X, y)$ and q(\theta) can be measured by the Kullback-Leibler (KL) divergence. Reducing the KL divergence is equivalent to increasing the log evidence lower bound based on the parameter set θ, as shown in Eq. (23):

$$\mathrm{KL}_{\mathrm{V1}} = q(\theta)p(F \mid \mathrm{X}, \theta)\log_p \int (\mathrm{y} \mid F)\, dF d\theta - \mathrm{KL}(q(\theta) \parallel p(\theta)) \quad (23)$$

Maximizing the KL divergence results in a variational distribution that approximates the posterior. The approximation $q(\theta)$ simplifies Eq. (23) to Eq. (24).

$$q(y^* \mid \mathrm{x}^*) = \int p(y^* \mid f)p(f \mid \mathrm{x}^*, \theta)q(\theta)df d\theta \quad (24)$$

During inference, the network parameters θ are sampled from $q(\theta)$. The feature extraction module at stage l, denoted as $g^{(l)}$, extracts the features $H^{(n)}$ as specified by Eq. (25).

$$H^{(n)} = g^{(l)}\left(H^{(l-1)}; W^{(l)}, b^{(l)}\right) = \text{normalize}\left(\text{pool}\left(\text{relu}\left(W^{(l)} * H^{(l-1)} + b^{(l)}\right)\right)\right) \quad (25)$$

The $*$ operator denotes convolution, which is one of the specific processes that go into feature extraction, along with non-linear transformations, max-pooling, and local normalization. After the convolution operation, a dot product is computed, which is followed by a non-linear transformation specified in Eq. (26) within the feature selection module f (l).

$$H^{(n)} = f^{(l)}\left(H^{(l-1)}; W^{(l)}, b^{(l)}\right) = \left(\text{relu}\left(W^{(l)} \cdot H^{(l-1)} + b^{(n)}\right)\right) \quad (26)$$

In Eq. (26), $H^{(l-1)}$ denotes the activation of the (l - 1) th hidden layer, and (.) denotes the dot product. To provide a probability distribution over every output class C, as represented in Eq. (27), the softmax operation is used as the final step in the prediction module.

$$p(C \mid X; W, b) = \text{softmax}\left(W^{(n)} \cdot H^{(l-1)} + b^{(n)}\right) \quad (27)$$

The DCNN model architecture is constructed by stacking the feature extraction, selection, and prediction modules, as shown in Eq. (28).

$$p(C \mid X; W, b) = \text{softmax}\left(f^{(5)}\left(f^{(4)}\left(g^{(3)}\left(g^{(2)}\left(g^{(1)}(X)\right)\right)\right)\right)\right) \quad (28)$$

During this optimization, local connections and weight sharing are implemented, resulting in a reduction in the number of parameters. Eq. (29) and Eq. (30) can be used to define 1-D and 2-D convolutional operations in a CNN, respectively:

$$\mathrm{o}_{i,k} = (\mathrm{x} * \mathrm{v})_{i,k} = \sum_{l,m} \mathrm{x}_{(i-1)s+m,l} \mathrm{V}_{m,l,k,} \quad (29)$$

$$\mathrm{O}_{i,j,k} = (\mathrm{X} * \mathrm{V})_{i,j,k} = \sum_{l,m,n} \mathrm{X}_{(i-1)s+m,(j-1)s+n,l} \mathrm{V}_{m,n,l,k} \quad (30)$$

In Eq. (29), x is the 1-D input, v is the convolutional kernel, and o is the output. Similarly, V and O are the corresponding kernel and output in Eq. (30), where X is the input of the 2-D convolutional operation. The number of data points skipped between two convolutional operations is referred to as the stride, denoted by s.

The data is split into training, validation, and test sets. The method is then trained on the training set, and after every epoch, the method is validated. After training, the model with the best validation kappa score is selected and evaluated on the test set.

Deep neural networks are capable of extracting features from raw input data. However, the quality and quantity of the training data are important requirements for achieving good performance. When the available data is limited, the network may not converge. In such cases, a pre-processing step can be applied to eliminate redundancy and reduce the feature dimensionality, which can help the network converge.

Consider a deep learning method with a model specification W. The training dataset consists of M samples, denoted as $S = (x1, y1), (x2, y2), (xM, yM)$, and so on. The model parameters are calculated using the Bayes formula, as shown in Eq. (31):

$$p(\mathrm{W} \mid \mathrm{S}) = \frac{p(\mathrm{S}|\mathrm{W})p(\mathrm{W})}{p(\mathrm{S})} = \frac{p(\mathrm{S}|\mathrm{W})p(\mathrm{W})}{\int_{\mathrm{W}} p(\mathrm{S}|\mathrm{W})p(\mathrm{W})\mathrm{dW}} \quad (31)$$

The prior distribution, denoted by $p(W)$, is based on an assumption, knowledge from the past, or experience. The likelihood function is $p(S|W)$, where $p(S)$ denotes the distribution of the training samples, and the predicted distribution of the model specification W is $p(W|S)$. However, the Bayes formula cannot be used directly to obtain the specification evaluation because it is challenging to calculate p(S). To address this issue, a new distribution, $q(W)$, is developed to approximate $p(\theta|S)$. The idea of Kullback-Leibler (KL) divergence can be used to calculate the difference between $q(W)$, and $p(W|S)$. Eq. (32) is used to express the KL divergence.

$$\begin{aligned} \mathrm{D_{KL}}\big(q(\theta) \parallel p(\theta \mid \mathrm{S})\big) &= \int_{\mathrm{W}} q(\mathrm{W}) \log \frac{q(\mathrm{W})}{p(\mathrm{WS})} \mathrm{dW} \\ &= \int_{\Theta} q(\mathrm{W}) \log \frac{q(\theta)\int_{\mathrm{W}} p(\mathrm{SW})p(\mathrm{W})\mathrm{dW}}{p(\mathrm{SW})p(\mathrm{W})} \mathrm{dW} \\ &= \int_{\Theta} q(\mathrm{W})p(\mathrm{S})\mathrm{dW} - \int_{\Theta} q(\mathrm{W}) \log \frac{p(\mathrm{SW})p(\theta)}{q(\mathrm{W})} \mathrm{dW} \quad (32) \end{aligned}$$

The goal of variational inference is to maximize the second term on the left-hand side of (33), which corresponds to $q(\theta)$, while minimizing the KL divergence.

$$\begin{aligned} q^*(\mathrm{W}) &= \underset{q(\mathrm{W})}{\operatorname{argmin}} \, \mathrm{DKL}\big(q(\mathrm{W}) \parallel p(\mathrm{W} \mid \mathrm{S})\big) \\ &= \underset{q(\mathrm{W})}{\operatorname{argmax}} \int_{\mathrm{W}} q(\mathrm{W}) \log \frac{p(\mathrm{S}|\mathrm{W})p(\mathrm{W})}{q(\mathrm{W})} \mathrm{dW} \quad (33) \end{aligned}$$

Assuming that q(W) is a joint Gaussian distribution and that each specification W_i in the specification matrix W follows an independent Gaussian distribution allows us to transform the variational problem into an optimization problem, as shown in Eq. (34):

$$(\mathrm{W}) = N(\mathrm{W}, \mu, \sigma^2) = \prod_i^{n_{W_i}} N(W_i, \mu_i, \sigma_i^2) \quad (34)$$

In Eq. (34), the mean value matrix and standard deviation are denoted by μ and σ, respectively. Determining the optimal values of the mean and standard deviation matrices, as shown in Eq. (35), will yield the ideal distribution $q(W)$:

$$\begin{aligned} \mu^*, \sigma^* = \underset{\mu_{1,\sigma},\sigma}{\operatorname{argmax}} & \sum_{k=1}^{n_\rho} \mathbb{E}_{q(\sigma_2 \varepsilon_k + u_k)}\big[\log\big(p(\sigma_k \varepsilon_k + u_k)\big)\big] \\ & + \sum_{k=1}^{n_\theta} \mathbb{E}_{q_k(\sigma_k \varepsilon_k + u_k)}\big[\log\big(q_k(\sigma_k \varepsilon_k + u_k)\big)\big] \\ & + \frac{N_N}{N} \sum_{j=1}^{N/N_B} N_B^N \sum_{i=1}^{N_n} \mathbb{E}_{q(\varepsilon)}\big[\log\big(p(\mathbf{y}_i \mid \mathbf{x}_i, \mu, \sigma, \varepsilon)\big)\big] \quad (35) \end{aligned}$$

IV. Experimental Analysis

In our evaluation, we present the compelling results of our proposed method for detecting data poisoning attacks in intelligent networks. Our approach consistently outperformed the referenced methods, achieving a significantly higher detection rate. The visualizations, including precision-recall curves and confusion matrices, vividly illustrate the superior performance and robustness of our method. These results provide strong evidence of the effectiveness and practical relevance of our approach in bolstering network security against data poisoning attacks.

A. Experimental Setup

The purpose of the experimental setup is to assess how well the suggested attack and defense strategies work. The Duchenne Smile Dataset, Product Dataset, and Sentiment Dataset are three real-world datasets used in the evaluation. Using customary cross-validation methods, these datasets are preprocessed and divided into training, testing, and validation sets. The suggested strategy is put into practice for the attack method using Python's NumPy and sklearn libraries. The datasets are subjected to the attack to evaluate its potential to undermine network security and jeopardize data classification. The suggested method is also put into practice for the defense method using Python's sklearn and NumPy libraries. The defense mechanism is applied to the datasets to test its efficacy in defending the network against threats and enhancing the accuracy and dependability of data classification. A comparison between the proposed methods and current methods, like K-Nearest Neighbors (KNN) and Malware-GAN, is done to ensure thorough evaluation. Throughput, Quality-of-Service (QoS), transmission rate, Root Mean Square Error (RMSE), accuracy, recall, precision, and malicious data analysis are just a few of the performance metrics that are measured and compared.

B. Dataset Description

This section provides a brief overview of the real-world datasets used in our experimental analysis in this section. These datasets are used to assess how well the attack and defense strategies we've suggested improve network security and data classification.

Duchenne Smile Dataset: The aim of this dataset is to determine whether a facial image contains a Duchenne or non-Duchenne smile. The task-creation and label-collection processes were performed using the Amazon Mechanical Turk platform. The dataset consists of 2,134 entries, with 64 regular employees producing 17,729 labels.

Product Dataset: The objective of this dataset is to determine whether two products are the same for each item in the dataset, which comprises pairs of items with descriptions. Participating employees were required to determine whether the two descriptions apply to the same item before providing their labels. This dataset contains 8,315 items, with 176 average workers providing 24,945 labels in total.

Sentiment Dataset: This dataset consists of a tweet about a specific firm for each item. The participating employees were tasked with determining whether the sentiment expressed in the tweet is favorable or unfavorable to the business. We created

1,000 objects using the AMT platform and collected labels from 85 regular workers. This dataset contains a total of 20,000 labels.

C. Performance Matrices

We use a set of performance metrics that cover various facets of the models' performance to assess the efficacy of our suggested attack and defense strategies.

Network throughput refers to the amount of data that can be successfully transported over a network in a certain period. It is measured in bits per second (bps) and can also refer to data packets per time slot or packets per second (pps). The aggregate throughput, also known as system throughput, is the total data rates sent to all network endpoints.

Quality-of-service (QoS) is a critical issue in wireless sensor applications, and each application has specific QoS requirements. Accuracy is one specification used to assess classification models, and it refers to the percentage of correct predictions made by a method. Recall and precision are measures of quantity and quality, respectively. A higher recall indicates that the method provides more relevant results, while a higher precision indicates that the method provides more relevant results than irrelevant ones. Precision is evaluated by dividing the total number of true positives (TP) by the total number of TP plus false positives (FP), while recall is calculated as the product of the number of TP divided by the sum of the TP and false negatives (FN).

Root means square error (RMSE) is a commonly used method for assessing the accuracy of forecasts, and it measures the Euclidean distance between the measured true values and forecasts. The standard deviation of residuals is also known as RMSE.

D. Comparative Analysis

Table I presents a comparative analysis between our proposed method and existing methods, based on network security and data classification. The analysis considers various parameters, including throughput, QoS, transmission rate, RMSE, accuracy, recall, precision, and malicious data analysis. The datasets analyzed include the Duchenne Smile Dataset, Product Dataset, and Sentiment Dataset.

Fig. 4 represents a comparative analysis between our proposed method and existing methods for network security. The graph shows that our proposed technique achieved a throughput of 96%, QoS of 83%, transmission rate of 89%, RMSE of 61%, accuracy of 95%, recall of 69%, precision of 79%, and malicious data analysis of 75%. In comparison, the KNN method achieved a throughput of 86%, QoS of 77%, transmission rate of 85%, RMSE of 55%, accuracy of 91%, recall of 65%, precision of 72%, and malicious data analysis of 69%, while the Malware-GAN method obtained a throughput of 94%, QoS of 79%, transmission rate of 88%, RMSE of 59%, accuracy of 93%, recall of 66%, precision of 75%, and malicious data analysis of 73%.

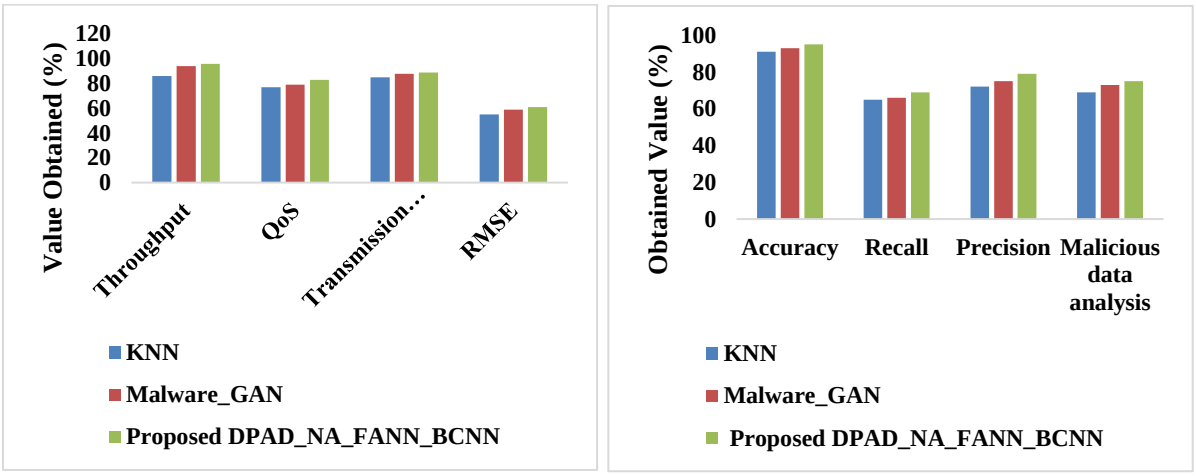
Significant performance differences are found when comparing the proposed method to the current network security methods. Compared to the KNN and Malware-GAN methods, our suggested technique outperformed them in all performance metrics. These findings suggest that when compared to the KNN and Malware-GAN methods, the proposed method is more effective and reliable in the context of network security analysis. The proposed method's efficiency in addressing network security issues is demonstrated by the higher throughput and transmission rate, better accuracy, and MDA.

Fig. 5 provides an analysis based on data classification between our proposed method and existing techniques. The graph shows that our proposed technique achieved a throughput of 95%, QoS of 85%, transmission rate of 93%, RMSE of 69%, accuracy of 96%, recall of 75%, precision of 85%, and malicious data analysis of 86%. In comparison, the KNN method achieved a throughput of 89%, QoS of 81%, transmission rate of 91%, RMSE of 63%, accuracy of 92%, recall of 71%, precision of 81%, and malicious data analysis of 79%, while the Malware-GAN method obtained a throughput of 95%, QoS of 85%, transmission rate of 93%, RMSE of 69%, accuracy of 96%, recall of 75%, precision of 85%, and malicious data analysis of 86%.

The proposed approach performs better than existing techniques for data classification, as shown by the comparative analysis between them. The outcomes show that the suggested method outperforms the KNN and Malware-GAN methods across the board. According to these findings, the proposed method performs data classification more effectively and efficiently than the KNN and Malware-GAN methods. The superiority of the suggested technique in handling data classification tasks is demonstrated by the higher throughput, transmission rate, accuracy, and MDA, along with better QoS and recall.

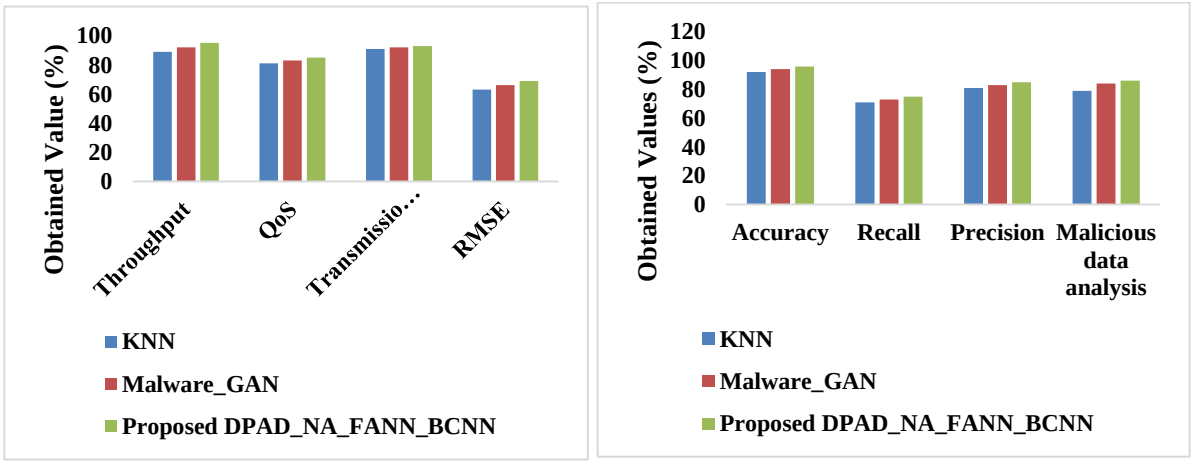
TABLE I. TABLE TYPE COMPARATIVE ANALYSIS OF PROPOSED AND EXISTING METHOD BASED ON NETWORK SECURITY AND DATA CLASSIFICATION

Techniques	Throughput	QoS	Transmission Rate	RMSE	Accuracy	Recall	Precision	Malicious Data Analysis
Case 1: Network security								
KNN	86	77	85	55	91	65	72	69
Malware_GAN	94	79	88	59	93	66	75	73
DPAD_NA_FANN_BCNN	96	83	89	61	95	69	79	75
Case 2: Data classification								
KNN	89	81	91	63	92	71	81	79
Malware_GAN	92	83	92	66	94	73	83	84
DPAD_NA_FANN_BCNN	95	85	93	69	96	75	85	86



(a) (b)

Fig. 4. Comparative analysis between proposed and existing method based on network security analysis: (a) comparison in terms of RMSE, transmission rate, QoS, and throughput (b) comparison in terms of MDA, precision, recall, and accuracy.



(a) (b)

Fig. 5. Comparative analysis between proposed and existing technique based on data classification: (a) comparison in terms of RMSE, transmission rate, QoS, and throughput; (b) comparison in terms of MDA, precision, recall, and accuracy.

We observe that our experimental analysis demonstrates the effectiveness of our proposed method in achieving high levels of network security and data classification performance. Our proposed technique outperforms the existing methods in terms of various parameters, including throughput, QoS, transmission rate, RMSE, accuracy, recall, precision, and malicious data analysis. The results indicate that our proposed method can significantly enhance the security and performance of wireless sensor networks.

Furthermore, the analysis of the Duchenne Smile Dataset, Product Dataset, and Sentiment Dataset reveals that our proposed method is robust and can be applied to various types of datasets. The high levels of accuracy, recall, and precision achieved by our proposed method indicate its potential for use in real-world applications, including in industries such as healthcare, e-commerce, and social media.

The results demonstrate the potential of our proposed method in enhancing the security and performance of wireless sensor networks and its ability to provide accurate and relevant results for data classification tasks. Further research can explore the use of our proposed method for other types of datasets and in different settings to evaluate its robustness and scalability.

The findings of this study highlight the effectiveness of the proposed technique in detecting and mitigating data poisoning attacks in intelligent networks. The achieved high levels of network security and accuracy in data detection demonstrate its practical value for network administrators, ensuring the protection of sensitive data and system integrity. The successful application of the technique contributes to advancements in network security and data analytics, while future research can focus on scalability and addressing potential vulnerabilities to further enhance its robustness. Overall, this study provides valuable insights for the implementation of secure and privacy-preserving intelligent networks.

V. Conclusion

This research proposes a novel technique for detecting data poisoning attacks based on deep learning, which combines

federated learning with adversarial neural networks. The proposed technique utilizes a Bayesian convolutional neural network to train the data analyzed by federated learning for detecting the presence of data poisoning attacks in the network. The experimental analysis is carried out based on network security and data classification, utilizing real-world datasets, such as the Duchenne Smile Dataset, Product Dataset, and Sentiment Dataset. The proposed technique outperforms the existing models in the literature, achieving high levels of performance in various parameters, including throughput, QoS, transmission rate, RMSE, accuracy, recall, precision, and malicious data analysis.

The results of this research indicate that the proposed technique can significantly enhance the security and performance of wireless sensor networks, contributing to a deeper understanding of data poisoning attacks and detection strategies in real-world contexts. Furthermore, this research contributes to the advancement of more effective outlier detection methods across a wider range of applications. Future work must address the challenge of preventing such attacks in a strengthened federated learning environment.

Overall, the proposed technique offers a promising approach to the detection of data poisoning attacks, which have become increasingly prevalent in wireless sensor networks. This research opens new avenues for future research in the field of wireless sensor networks, and the proposed technique holds potential for use in various industries, including healthcare, e-commerce, and social media.

REFERENCES

[1] F. Hanzely and P. Richtárik, "Federated learning of a mixture of global and local models," arXiv preprint arXiv:2002.05516, 2020.

[2] Y. Ding, Z. Tang and F. Wang, "Single-sample face recognition based on shared generative adversarial network," Mathematics, vol. 10, no. 5, pp. 752-758, 2022.

[3] T. Wang, H. Li, M. Noori, R. Ghiasi, S. C. Kuok et al., "Probabilistic seismic response prediction of three-dimensional structures based on bayesian convolutional neural network," Sensors, vol. 22, no. 10, pp. 3775-3790, 2022.

[4] G. Zhao, F. Liu, J. A. Oler, M. E. Meyerand, N. H. Kalin et al., "Bayesian convolutional neural network based MRI brain extraction on nonhuman primates," Neuroimage, vol. 175, no. 1, pp. 32-44, 2018.

[5] M. Joshaghani, A. Davari, F. N. Hatamian, A. Maier and C. Riess, "Bayesian convolutional neural networks for limited data hyperspectral remote sensing image classification," arXiv preprint arXiv:2205.09250, 2022.

[6] U. Zafar, M. Ghafoor, T. Zia, G. Ahmed, A. Latif et al., "Face recognition with bayesian convolutional networks for robust surveillance systems," EURASIP Journal on Image and Video Processing, vol. 2019, no. 1, pp. 1-10, 2019.

[7] I. M. Ahmed and M. Y. Kashmoola, "Threats on machine learning technique by data poisoning attack: a survey," in International Conference on Advances in Cyber Security, Penang, Malaysia, pp. 586-600, Springer, Singapore, August 2021.

[8] F. A. Yerlikaya and Ş. Bahtiyar, "Data poisoning attacks against machine learning algorithms," Expert Systems with Applications, vol. 189, no. 1, 118101, 2022.

[9] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild et al., "Dataset security for machine learning: data poisoning, backdoor attacks, and defences," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 2, pp. 1563-1580, 2022.

[10] H. Huang, J. Mu, N. Z. Gong, Q. Li, B. Liu et al., "Data poisoning attacks to deep learning based recommender systems," arXiv preprint arXiv:2101.02644, 2021.

[11] A. E. Cinà, K. Grosse, A. Demontis, B. Biggio, F. Roli et al., "Machine learning security against data poisoning: are we there yet?" arXiv preprint arXiv:2204.05986, 2022.

[12] L. Verde, F. Marulli and S. Marrone, "Exploring the impact of data poisoning attacks on machine learning model reliability," Procedia Computer Science, vol. 192, no. 2, pp. 2624-2632, 2021.

[13] J. Chen, X. Zhang, R. Zhang, C. Wang, and L. Liu, "De-pois: An attack-agnostic defense against data poisoning attacks," IEEE Transactions on Information Forensics and Security, vol. 16, no. 1, pp. 3412-3425, 2021.

[14] F. Nuding and R. Mayer, "Data poisoning in sequential and parallel federated learning," in Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics, Baltimore, MD, USA, pp. 24-34, 2022.

[15] H. Liu, D. Li and Y. Li, "Poisonous label attack: Black-box data poisoning attack with enhanced conditional DCGAN," Neural Processing Letters, vol. 53, no. 6, pp. 4117-4142, 2021.

[16] A. Milakovic and R. Mayer, "Combining defences against data-poisoning based backdoor attacks on neural networks," In IFIP Annual Conference on Data and Applications Security and Privacy, Newark, NJ, USA, pp. 28-47, Springer, Cham, 2022.

[17] S. Bhattacharjee, M. J. Islam and S. Abedzadeh, "Robust anomaly based attack detection in smart grids under data poisoning attacks," In Proceedings of the 8th ACM on Cyber-Physical System Security Workshop, Nagasaki, Japan, pp. 3-14, 2022.

[18] F. Razmi and L. Xiong, "Classification auto-encoder based detector against diverse data poisoning attacks," arXiv preprint arXiv:2108.04206, 2021.

[19] S. Farhadkhani, R. Guerraoui and O. Villemaud, "An equivalence between data poisoning and byzantine gradient attacks," In International Conference on Machine Learning, Honolulu, Hawai, pp. 6284-6323, PMLR, June 2022.

[20] Y. Mao, X. Yuan, X. Zhao and S. Zhong, "Romoa: Robust model aggregation for the resistance of federated learning to model poisoning attacks," in European Symposium on Research in Computer Security, pp. 476-496, Springer, Cham, October 2021.

[21] H. Zhang, J. Gao and L. Su, "Data poisoning attacks against outcome interpretations of predictive models," in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Virtual Event Singapore, pp. 2165-2173, 2021.

Strengthening AES Security Through Key-Dependent ShiftRow and AddRoundKey Transformations Utilizing Permutation

Tran Thi Luong
Academy of Cryptography Techniques, No. 141 Chien Thang Road, Tan Trieu, Thanh Tri, Hanoi, Vietnam

***Abstract*—AES (Advanced Encryption Standard) is a widely applied block cipher standard in the United States, used in various security applications today. Currently, there are numerous research endeavors aimed at making AES block ciphers dynamic to improve their security against contemporary strong attacks. The most common dynamic approach involves the dynamization of AES block transformations, including SubByte, ShiftRow, AddRoundKey, and MixColumn operations. The combination of these transformations has also been explored and proposed. However, to the best of our knowledge, the dynamic combination of AddRoundKey and ShiftRow transformations remains unexplored. Therefore, in this paper we introduce algorithms for generating key-dependent AddRoundKey and ShiftRow transformations based on permutations. Subsequently, these key-dependent transformations are applied to AES to create dynamic AES block ciphers. Security analysis and evaluation of NIST's statistical criteria are performed, and the entropy of AES and dynamic AES is assessed. From our findings, it is evident that dynamic AES block ciphers can significantly enhance AES security and meet stringent randomness criteria, similar to AES.**

***Keywords*—AES; ShiftRow; AddRoundKey; dynamic AES; key-dependent**

I. Introduction

Along with the strong development of information technology, security risks and attacks are increasing in complexity. Therefore, cryptographic primitives are being widely used in various security domains nowadays [1, 2, 3]. Substitution-Permutation Network (SPN) block ciphers [4, 5, 6] represent a prevalent category of block ciphers extensively applied in contemporary cryptographic scenarios. An SPN block cipher comprises three primary components: the substitution layer, which typically employs S-boxes [7–10]; the diffusion layer, commonly utilizing MDS matrices [11–14] (matrices derived from maximum distance separable codes); and the key addition layer.

AES [15, 16] belongs to the class of SPN block ciphers and serves as a block cipher standard established by NIST in 2001, originating in the United States. The AES round function incorporates three operations, namely key addition, substitution, and linear transformations. AES, one of the world's most widely used encryption algorithms, faces potential vulnerabilities that could be exploited by cryptanalysts. The simplicity of AES's mathematical structure and the threat of attacks such as algebraic attacks [17], linear attacks [8, 18], and differential attacks [18, 19], make its security a concern. Moreover, the advent of supercomputers and quantum computing poses further risks, necessitating increased key lengths for maintaining security. Therefore, researching various approaches to enhance the security strength of AES is crucial in the current scenario.

To enhance the resilience of block ciphers against modern, potent attacks, extensive research has been conducted to animate these cryptographic algorithms. Specifically, with the AES block cipher, there is a variety of approaches to dynamize the AES block cipher to enhance its security. Some of these methods center on incorporating S-boxes in AES that depend on a secret key [20–25], while others work on creating key-dependent MixColumn transformations for AES [26–28]. Notably, there are studies exploring the dynamization of both AES's S-boxes and MixColumn [29], or the dynamization of all three transformations: S-Boxes, MixColumn, and ShiftRow, which have also received attention [30–32]. Another current research direction is to make the XOR operation dynamic in AES [33, 34].

For the dynamic S-box approach in AES, in their work [20], the authors introduced a method for generating S-boxes that depend on the encryption key and possess favorable algebraic characteristics, including non-linearity, BIC, and SAC. Furthermore, an alternative approach to produce S-boxes that rely on the encryption key in AES was introduced in [21]. This method entails establishing a novel arrangement for the S-box through the use of a simulated key expansion algorithm. In [22], the authors introduced an innovative method for creating variable S-boxes by rearranging the S-box of AES. These adaptable S-boxes rely on a secret key and utilize an affine constant and an unconventional polynomial. For each additional key bit, a fresh S-box with rearranged values is produced, thus enhancing the intricacy of the algorithm. In [23], the authors presented an approach to create S-boxes that vary with the encryption key, employing a evolving approach. The evaluative experimentation of these key-dependent S-boxes was conducted, focusing on characteristics such as achieving a SAC, BIC, balanced output, non-linearity, and probabilities related to linear and differential approximation. In [24], the authors introduced four straightforward procedures for producing key-influenced S-boxes. To assess the quality of these S-boxes, they introduced eight standardized dissimilarity measurements. The authors outlined four methods for generating key-influenced S-boxes and scrutinized eight normalized dissimilarity measurements

employed to appraise the effectiveness of these key-dependent generation techniques. Furthermore, in [25], Murphy et al. introduced an approach for the differential cryptanalysis of key-influenced S-boxes, elucidating methods for performing cryptanalysis with the utilization of these S-boxes.

For the dynamic Mixcolumn approach in AES, in [26], the authors proposed a MixColumn transformation that depends on the encryption key, derived from the AES MDS matrix, using scalar multiplication on the rows of the matrix along with an extra m-bit key. An idea presented in [27] includes the creation of a diffusion layer that relies on the encryption key, achieved through scalar multiplication and immediate exponentiation. In [28], the authors introduced a collection of $n \times n$ binary matrices that can be employed to create dynamic matrices resembling AES and recursive MDS matrices.

For the approach of making multiple transformations dynamic in AES, in [29], the authors introduced dynamic S-boxes and novel MixColumn matrices that preserve favorable cryptographic characteristics when developing dynamic AES. In [30], an image encryption method rooted in symmetric cryptography was introduced. It utilizes transformations like MixColumns, ShiftRows, and SubByte, which are dynamically influenced by the encryption key. In [31], the authors presented a dynamic block cipher based on AES, in which AES parameters vary for each unique key. More precisely, the ShiftRows, the SubBytes, and MixColumns transformations adapt according to the key, leading to distinct behavior for every key. Extensive testing has verified the security of the proposed algorithm. In [32], a fresh and efficient AES algorithm, which is dependent on the key, is introduced. The authors have put forward an innovative approach to enhance the advanced encryption standard algorithm by employing dynamic sub-byte, mix-column, and shift rows operations to ensure secure communication. This novel work exhibits superior avalanche and strict avalanche effects when compared to the conventional AES algorithm.

In [33, 34], the authors introduced innovative techniques that employ key-dependent XOR tables utilizing 3D chaotic maps. The authors utilized XOR tables that rely on the initial confidential parameters. In [34], they established a fresh MDS matrix, however, regrettably, this matrix does not qualify as an MDS matrix. Furthermore, their approaches in [31, 32] still exhibit numerous weaknesses and shortcomings.

Based on our review of related works, to the best of our knowledge, we haven't come across any research that investigates the combination of animating both the ShiftRow and AddRoundKey transformations of AES. In this paper, we introduce algorithms for generating key-dependent AddRoundKey and ShiftRow transformations based on permutations. Subsequently, we apply these key-dependent transformations to AES to create a dynamic AES block cipher. We conduct security analysis and evaluate NIST's statistical criteria, as well as assess the entropy of AES and dynamic AES. Consequently, it becomes evident that dynamic AES block ciphers can significantly enhance the security of AES and meet rigorous randomness criteria, similar to AES.

The structure of the remaining part of the paper is as follows: Section II provides preliminaries. Section III introduces algorithms for generating key-dependent ShiftRow and AddRoundKey transformations based on permutations. Section IV adapts the AES block cipher using key-dependent ShiftRow and AddRoundKey operations. Section V is conclusion.

II. PRELIMINARIES

A. Introduction to Hadamard Matrices

A Hadamard matrix [35] of dimension d, with the initial row elements represented as $h_0, h_1, \dots, h_d$, can be designated in the following manner.

$$H = had(h_0, h_1, \dots, h_d)$$

Furthermore, a Hadamard matrix of size $2d \times 2d$ has the following form.

$$H = \begin{pmatrix} A & B \\ B & A \end{pmatrix}$$

where A and B are $d \times d$ matrices, d is even.

B. ShiffRow and AddRoundKey Transformations in AES

The ShiftRow operation in AES processes the state by left rotating the last three rows of the state with a varying number of rotation. Row 1 of the state remains unchanged, row 2 of the state left rotates by 1 byte, row 3 of the state left rotates by 2 bytes, and row 4 of the state left rotates by 3 bytes.

The AddRoundKey operation in AES performs a bitwise XOR between the state and a round key. The AddRoundKey operation is represented by a 4-bit XOR table as described in Table I.

TABLE I. THE 4-BIT XOR TABLE IN AES

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	0	3	2	5	4	7	6	9	8	11	10	13	12	15	14
2	2	3	0	1	6	7	4	5	10	11	8	9	14	15	12	13
3	3	2	1	0	7	6	5	4	11	10	9	8	15	14	13	12
4	4	5	6	7	0	1	2	3	12	13	14	15	8	9	10	11
5	5	4	7	6	1	0	3	2	13	12	15	14	9	8	11	10
6	6	7	4	5	2	3	0	1	14	15	12	13	10	11	8	9
7	7	6	5	4	3	2	1	0	15	14	13	12	11	10	9	8
8	8	9	10	11	12	13	14	15	0	1	2	3	4	5	6	7
9	9	8	11	10	13	12	15	14	1	0	3	2	5	4	7	6
10	10	11	8	9	14	15	12	13	2	3	0	1	6	7	4	5
11	11	10	9	8	15	14	13	12	3	2	1	0	7	6	5	4
12	12	13	14	15	8	9	10	11	4	5	6	7	0	1	2	3
13	13	12	15	14	9	8	11	10	5	4	7	6	1	0	3	2
14	14	15	12	13	10	11	8	9	6	7	4	5	2	3	0	1
15	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0

III. Proposing Algorithms to Generate Key-Dependent ShiftRow and AddRoundKey Transformations based on Permutation

A. Algorithm for Generating Key-Dependent ShiftRow

First, we analyze the diffusion capacity of active bytes (non-zero byte) through two rounds of AES. Fig. 1 represents the diffusion state of AES after the first round, with the initial state containing an active byte (indicated by the black cell).

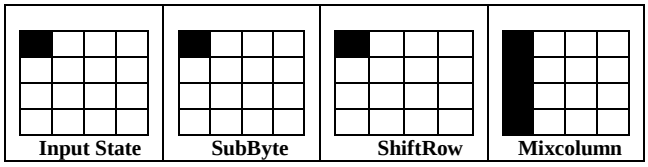


Fig. 1. Diffusion state of AES after the first round.

According to Fig. 1, after the SubByte and ShiftRow transformations in the first round, the number of active bytes in the state array remains at 1. Subsequently, during the MixColumn transformation, the number of active bytes becomes 4 due to the diffusion capabilities of the MDS matrix, and the number of active bytes is preserved when going through the AddRoundKey transformation. Thus, starting with 1 active byte, there will be 4 active bytes at the end of the first round.

Fig. 2 shows the diffusion state of AES after the second round.

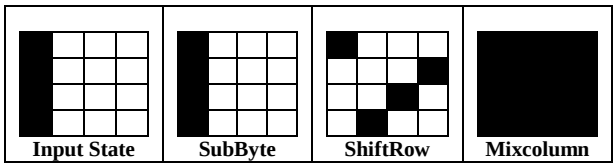


Fig. 2. Diffusion state of AES after the second round.

From Fig. 2, we can observe that the role of ShiftRow is extremely important in propagating active bytes from column 1 to all four columns of the state array. This allows, through the MixColumn transformation, the maximum number of active bytes to be achieved, which is 16 bytes in the state array.

Remark 1: The crucial point in designing the ShiftRow transformation is its capability to distribute active bytes from one column to all the remaining columns, ensuring that all four columns of the state array contain at least one active byte.

According to Fig. 2, the number of bytes rotated in each row of the state by the ShiftRow transformation in AES can be described as shown in Fig. 3.

Therefore, as long as the number of bytes rotated in each row is distinct and falls within the range [0, 3], active bytes will undoubtedly be propagated to all four columns of the state.

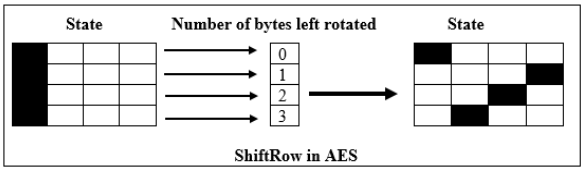


Fig. 3. The number of bytes rotated in each row of the state by the ShiftRow in AES.

Based on these observations, we propose the idea presented in Algorithm 1 to generate key-dependent ShiftRow transformations while ensuring the propagation of active bytes as described above.

Algorithm 1. Key-dependent ShiftRow Transformation Generation via Permutation

Input: A secret key K consists of k ($k \geq 128$) bits; A state S of size 4×4.

Output: The new ShiftRow operation for AES depends on the key K; A new state $\acute{S}$.

Step 1: Take the first two bits of the key K and convert them into an integer, denoted as a_0. Take the next two bits of K and convert them into an integer. If this integer is different from a_0, assign it to a_1; otherwise, shift the key K to the right by one bit until you obtain $a_1 \neq a_0$. Continue this process until you have four distinct integers: a_0, a_1, a_2, a_3.

Step 2: From the permutation (a_0, a_1, a_2, a_3) obtained in step 1, left rotate the rows of the state S as follows: left rotate a_0 bytes for row 1, left rotate a_1 bytes for row 2, left rotate a_2 bytes for row 3, left rotate a_3 bytes for row 4. The resulting state is denoted as $\acute{S}$.

Step 3: The left rotation operation as in step 2 is called *KD_ShiftRow*. The *KD_ShiftRow* operation will be used to replace the ShiftRow operation in AES.

Remark 2. Because there are 4! permutations of (0, 1, 2, 3), there will be $4! = 24$ key-dependent ShiftRow operations (*KD_ShiftRow*) that can be generated by Algorithm 1.

Example 1. If step 1 of Algorithm 1 results in the permutation $(a_0, a_1, a_2, a_3) = (2, 0, 3, 1)$, then the *KD_ShiftRow* operation obtained from Algorithm 1 will function as shown in Fig. 4.

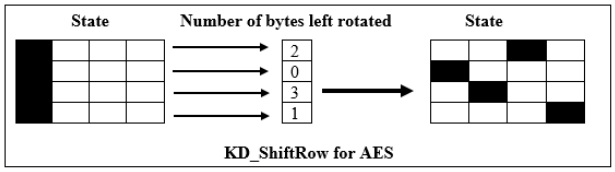


Fig. 4. An example of the KD_ShiftRow obtained from Algorithm 1.

B. Algorithm for Generating Key-Dependent AddRoundKey Transformations

From the original XOR table of AES (Table I), we have a remark regarding three essential Attributes that a XOR table must possess.

Remark 3. Three essential Attributes of an XOR table

- Attribute 1: Every row and column in the XOR table contains unique values within the range of 0 to 15.
- Attribute 2: The XOR table should exhibit symmetry along the principal diagonal, implying that $m\ XOR\ n = n\ XOR\ m$.
- Attribute 3: For any given m, n, and k elements within the XOR table where $m\ XOR\ n = k$, the following holds true: $m\ XOR\ k = n$ and $n\ XOR\ k = m$.

From the XOR table of AES, denote the 4×4 matrices as follows:

$$A_0 = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 3 & 2 \\ 2 & 3 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{pmatrix}, A_1 = \begin{pmatrix} 4 & 5 & 6 & 7 \\ 5 & 4 & 7 & 6 \\ 6 & 7 & 4 & 5 \\ 7 & 6 & 5 & 4 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} 8 & 9 & 10 & 11 \\ 9 & 8 & 11 & 10 \\ 10 & 11 & 8 & 9 \\ 11 & 10 & 9 & 8 \end{pmatrix}, A_3 = \begin{pmatrix} 12 & 13 & 14 & 15 \\ 13 & 12 & 15 & 14 \\ 14 & 15 & 12 & 13 \\ 15 & 14 & 13 & 12 \end{pmatrix}.$$

$$\text{Set } \boldsymbol{A} = \begin{pmatrix} A_0 & A_1 & A_2 & A_3 \\ A_1 & A_0 & A_3 & A_2 \\ A_2 & A_3 & A_0 & A_1 \\ A_3 & A_2 & A_1 & A_0 \end{pmatrix}.$$

Remark 4. The matrices A_i $(0 \le i \le 3)$ are Hadamard matrices. And the matrix $\boldsymbol{A}$ is also a Hadamard matrix in the form $\boldsymbol{A} = had(0, 1, 2, \dots, 15)$. Therefore, the XOR table of AES is created from the Hadamard matrix $\boldsymbol{A}$, where **row 0** and **column 0** (bolded) of this XOR table are ordered according to the first row of matrix $\boldsymbol{A}$, meaning in the order 0, 1, 2, …, 15.

We can denote: $\boldsymbol{A} = had(A_0, A_1, A_2, A_3)$.

From Remark 4, it can be seen that by permuting the matrices A_i, new Hadamard matrices $\boldsymbol{A}$ can be generated. Based on this idea, we propose Algorithm 2 to generate a new XOR table and key-dependent AddRoundKey transformation by permuting the A_i matrices in matrix $\boldsymbol{A}$.

Algorithm 2. Generating a new 4-bit XOR table and key-dependent AddRoundKey transformation based on permutation.

Input: A secret key K consisting of k $(k \ge 128)$ bits, the original XOR table of AES.

Output: A new 4-bit XOR table; Key-dependent AddRoundKey operation.

Step 1: Take the first two bits of the secret key K and convert these two bits into an integer, denoted as x_0. Take the next two bits of K and convert them into an integer. If this integer is different from x_0, assign this integer to x_1; otherwise, shift the key K to the right by one bit until $x_1 \ne x_0$ is obtained. Do the same for the remaining bits until four distinct integers are obtained: x_0, x_1, x_2, x_3.

Step 2: Construct a permutation-based Hadamard matrix using the permutation (x_0, x_1, x_2, x_3) obtained in step 1, which has the form: $\hat{\boldsymbol{A}} = had(A_{x_0}, A_{x_1}, A_{x_2}, A_{x_3})$.

Step 3: Construct a new 4-bit XOR table based on the Hadamard matrix $\hat{\boldsymbol{A}}$ such that the elements in the first row and the first column of this new XOR table follow the order of elements in the first row of matrix $\hat{\boldsymbol{A}}$.

Step 4: Reorder the rows and columns of the new XOR table so that both row 0 and column 0 of the new XOR table follow an increasing order from 0 to 15.

Step 5: Replace the regular bitwise XOR operation in the AddRoundKey transformation of AES with the new XOR operation determined by the new XOR table created in step 4. The result is the *KD_AddRoundKey* operation, which is used in place of the AddRoundKey operation in AES.

Remark 5. Since there are 4! permutations of (0, 1, 2, 3), there will be a total of $4! = 24$ new XOR tables generated by Algorithm 2, corresponding to the number of *KD_AddRoundKey* operations obtained.

Example 2. Suppose that step 1 of Algorithm 2 yields the permutation $(x_0, x_1, x_2, x_3) = (3, 2, 0, 1)$. The resulting Hadamard matrix is as follows:

$$\hat{\boldsymbol{A}} = had(A_3, A_2, A_0, A_1) = \begin{pmatrix} A_3 & A_2 & A_0 & A_1 \\ A_2 & A_3 & A_1 & A_0 \\ A_0 & A_1 & A_3 & A_2 \\ A_1 & A_0 & A_2 & A_3 \end{pmatrix}$$

In that case, the resulting new XOR table after Algorithm 2 is presented in Table II.

TABLE II. THE NEW XOR TABLE GENERATED FROM ALGORITHM 1

	12	**13**	**14**	**15**	**8**	**9**	**10**	**11**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
12	12	13	14	15	8	9	10	11	0	1	2	3	4	5	6	7
13	13	12	15	14	9	8	11	10	1	0	3	2	5	4	7	6
14	14	15	12	13	10	11	8	9	2	3	0	1	6	7	4	5
15	15	14	13	12	11	10	9	8	3	2	1	0	7	6	5	4
8	8	9	10	11	12	13	14	15	4	5	6	7	0	1	2	3
9	9	8	11	10	13	12	15	14	5	4	7	6	1	0	3	2
10	10	11	8	9	14	15	12	13	6	7	4	5	2	3	0	1
11	11	10	9	8	15	14	13	12	7	6	5	4	3	2	1	0
0	0	1	2	3	4	5	6	7	12	13	14	15	8	9	10	11
1	1	0	3	2	5	4	7	6	13	12	15	14	9	8	11	10
2	2	3	0	1	6	7	4	5	14	15	12	13	10	11	8	9
3	3	2	1	0	7	6	5	4	15	14	13	12	11	10	9	8
4	4	5	6	7	0	1	2	3	8	9	10	11	12	13	14	15
5	5	4	7	6	1	0	3	2	9	8	11	10	13	12	15	14
6	6	7	4	5	2	3	0	1	10	11	8	9	14	15	12	13
7	7	6	5	4	3	2	1	0	11	10	9	8	15	14	13	12

From the three attributes of an XOR table as mentioned in Remark 3, we prove the accuracy of the new XOR table generated by Algorithm 1 with the following proposition.

Proposition 1. The new XOR table generated by Algorithm 1 satisfies the three necessary attributes of an XOR table.

Proof.

Since the Hadamard matrix $\hat{\boldsymbol{A}}$ is essentially a permutation of elements within a row of the original XOR table of AES, Atrribute 1 of the new XOR table is satisfied.

Matrix

$$\acute{\boldsymbol{A}} = had\left(A_{x_0}, A_{x_1}, A_{x_2}, A_{x_3}\right) = \begin{pmatrix} A_{x_0} & A_{x_1} & A_{x_2} & A_{x_3} \\ A_{x_1} & A_{x_0} & A_{x_3} & A_{x_2} \\ A_{x_2} & A_{x_3} & A_{x_0} & A_{x_1} \\ A_{x_3} & A_{x_2} & A_{x_1} & A_{x_0} \end{pmatrix}$$

is a Hadamard one, and the matrices $A_{x_i}(0 \le i \le 3)$ are also Hadamard matrices, so the matrix $\acute{\boldsymbol{A}}$ is symmetric across the main diagonal. Thus, Attribute 2 of the new XOR table is satisfied.

The matrix $\acute{\boldsymbol{A}}$ is denoted as follows: $\acute{\boldsymbol{A}} = had(a_0, a_1, \dots, a_{15})$.

Comparing with the original XOR table of AES, it can be observed that the elements of the new XOR table have been replaced by a one-to-one mapping as follows: $0 \to a_0$, $1 \to a_1$, …, $15 \to a_{15}$. This substitution is applied to all elements of the original XOR table, including both row 0 and column 0.

According to Attribute 3 of the original XOR table: if $i\ XOR\ l = j$, it always holds that $i\ XOR\ j = l$ and $l\ XOR\ j = i$.

From the replacement operation, we also have the corresponding relationship: if $a_i\ XOR\ a_l = a_j$, it always holds that $a_i\ XOR\ a_j = a_l$ and $a_l\ XOR\ a_j = a_i$. Therefore, Attribute 3 of the new XOR table is satisfied.

On the other hand, rearranging the order of rows and columns of the new XOR table so that row 0 and column 0 of that XOR table are in increasing order from 0 to 15 will not affect the values in the XOR table. Therefore, all three properties are satisfied for the new XOR table.

C. Proposing the Combined Algorithm

In this section, we will propose a combined algorithm to generate both the key-dependent ShiftRow and AddRoundKey transformations for the AES block cipher.

Algorithm 3. Generating key-dependent ShiftRow and AddRoundKey transformations.

Input: A secret key K consisting of k ($k \ge 128$) bits; The original XOR table of AES.

Output: A new 4-bit XOR table; New key-dependent ShiftRow and AddRoundKey transformations for AES.

Step 1: Take the first two bits of key K and convert these two bits into integers, denoted as a_0. Take the next two bits of K and convert them into integers. If this integer is different from a_0, assign it to a_1; otherwise, shift key K to the right by one bit until you obtain $a_1 \neq a_0$.. Repeat this process until you have four distinct integers: a_0, a_1, a_2, a_3.

Step 2: From the permutation (a_0, a_1, a_2, a_3) obtained in step 1, for any arbitrary state S, perform a left rotation on the rows of the state S as follows: rotate row 1 to the left by a_0 bytes, rotate row 2 to the left by a_1 bytes, rotate row 3 to the left by a_2 bytes, rotate row 3 to the left by a_2 bytes. This left rotating operation is named *KD_ShiftRow* and will be used to replace the ShiftRow operation in AES.

Step 3: Following the same procedure as in step 1, with the next bits of the secret key K after step 1, we obtain a permutation (x_0, x_1, x_2, x_3).

Step 4: Construct a Hadamard matrix based on the permutation (x_0, x_1, x_2, x_3) obtained in step 3, in the form of: $\acute{\boldsymbol{A}} = had(A_{x_0}, A_{x_1}, A_{x_2}, A_{x_3})$.

Step 5: Construct a new 4-bit XOR table based on the Hadamard matrix $\acute{\boldsymbol{A}}$, with the first row and column of the new XOR table containing elements in the same order as those in the first row of matrix $\acute{\boldsymbol{A}}$. Rearrange the rows and columns of the new XOR table so that the first row and column follow an increasing order from 0 to 15.

Step 6: Replace the usual bitwise XOR operation in the AddRoundKey transformation of AES with the new XOR operation defined by the XOR table generated in step 5. The result is the *KD_AddRoundKey* operation, which is used in place of AddRoundKey in AES.

IV. Adapt the AES Block Cipher by Incorporating the Key-Dependent ShiftRow and AddRoundKey Transformations

A. Implementation of Experiments

Execute the combined algorithm in Algorithm 3 to obtain two key-dependent transformations: *KD_ShiftRow* and *KD_AddRoundKey*. Then, use these two transformations to replace the original ShiftRow and AddRoundKey in AES. The resulting dynamic AES algorithm is denoted as *ShiftAES*. Fig. 5 illustrates the diagram of encryption/decryption rounds in the *ShiftAES* algorithm.

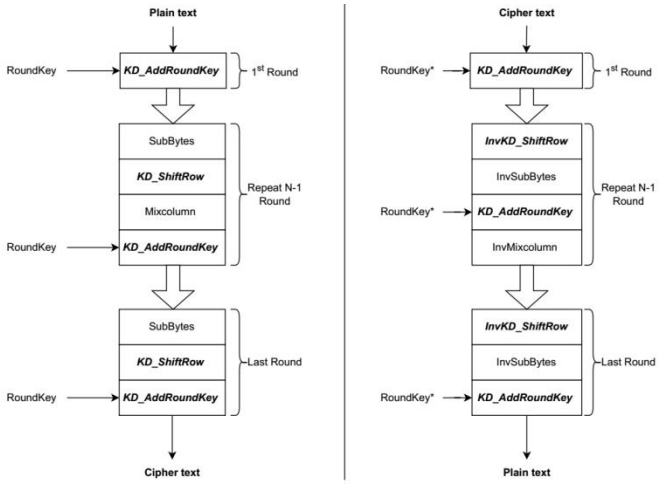


Fig. 5. Encryption / decryption round diagram of the ShiftAES algorithm.

For experiment, we select the AES-128 block cipher with a XOR table of 4-bit. This led to the development of a key-dependent dynamic block cipher algorithm, which we named ShiftAES-128 based on Algorithm 3. We implement these algorithms using C++ on an Asus K43SJ Laptop (Core i5-2430M, 500GB, HDD 6GB RAM, Nvidia Geforce GT 520M).

B. Security Analysis

In the realm of block ciphers, encompassing AES, the most potent threats manifest in the form of differential attacks [18, 19], linear attacks [18, 36], or their derivatives. In the case of differential attacks, attackers must rely on predefined differential patterns to execute their strategies. Our ShiftAES block cipher ensures the continuous alteration of differential patterns [17], contingent upon the encryption key. This

dynamic behavior significantly enhances the security of the ShiftAES block cipher, as it becomes exceedingly arduous for potential attackers to compute the differentials required for launching attacks.

The use of the key-dependent ShiftRow operation makes it considerably more challenging for attackers as it conceals the exact rotation values applied in AES's *KD_ShiftRow*, rendering decryption significantly more difficult. Moreover, the utilization of key-dependent AddRoundKey operations will heighten the intricacy of key-related attacks [37]. It disrupts the conventional property of XOR tables, where different inputs with the same difference yield an output difference of 0. The dynamic nature of the relationship between input differences, driven by the variable key, adds an additional layer of security to block ciphers that incorporate such elements.

Combining both KD_ShiftRow and KD_AddRoundKey makes the attacker's task significantly more challenging. In the case of AES, attackers know how ShiftRow and AddRoundKey work, they might collect a large number of plaintext/ciphertext pairs for differential or linear cryptanalysis attacks. This number of pairs could be very large, let's assume it's 2^{80}. However, with the *ShiftAES* algorithm, there can be 24 possibilities for *KD_ShiftRow* and 24 possibilities for dynamic XOR tables, which results in $24 \times 24 = 576$ possible combinations for ShiftRow and AddRoundKey used in *ShiftAES*. In this scenario, the number of plaintext/ciphertext pairs required for an attack is not just 2^{80} but 576×2^{80}. This number, at first glance, may not seem significantly larger, but in reality, it's extremely huge and has practical implications. So, for the dynamic AES block cipher, an attacker needs to collect plaintext/ciphertext pairs that are 576 times more than in the case of regular AES. This is not an easy task in practice. Furthermore, when the encryption/decryption switches to a different secret key, meaning different *KD_ShiftRow* and *KD_AddRoundKey* are used, the attacker might not have enough time to gather a sufficient number of plaintext/ciphertext pairs to carry out the attack.

Therefore, it can be seen that the proposed dynamic method can significantly increase the security of AES. Furthermore, as mentioned in the introduction, our proposed method is a novel approach, which may be highly beneficial for cryptography designers in designing secure dynamic SPN block ciphers.

C. Evaluating the Random Statistical Standards

In this section, we assess the random statistical standards according to NIST SP 800-22 [38] and Shannon Entropy criteria [39] of the AES and *ShiftAES* block cipher.

NIST SP 800-22 [38] has been developed to serve as the primary and extensively utilized means for evaluating the statistical randomness of random or pseudorandom number generators in the field of cryptography. NIST's set of randomness assessments encompasses a total of 15 tests, which are outlined below.

Random Excursions Test; Frequency Test within a Block; Test for the Longest Run of Ones in a Block; Approximate Entropy Test; Non-overlapping Template Matching Test; Binary Matrix Rank Test; Random Excursions Variant Test; Cumulative Sums (Cusum) Test; Frequency (Monobit) Test; Linear Complexity Test; Runs Test; Serial Test; Overlapping Template Matching Test; Discrete Fourier Transform (Spectral) Test; Maurer's "Universal Statistical" Test.

For a given input sequence, every test computes a respective p-value, which is then compared to the significance level $\alpha = 0.01$. Should $p_i \geq \alpha$, it is inferred that the sequence exhibits randomness, and conversely.

Entropy, also known as information entropy, is described as the degree of unpredictability concerning an individual's knowledge or the result of an experiment before it is observed, as well as the connected deterministic characteristics for forecasting its value. Higher entropy indicates increased uncertainty when forecasting an observation's value. Shannon entropy represents one category of information entropy developed by Shannon in [39].

1) Evaluate the shannon entropy change using the ENT tool: Shannon entropy is a vital indicator of randomness. In this section, we assess the alteration in Shannon entropy for the four datasets, namely LW, HW, AV1, and Rot, across each round of the DAES block cipher. This analysis aims to provide a comprehensive view of the randomness at each round. We utilized the ENT tool [40] to perform the evaluations and acquired the subsequent outcomes.

The results of the entropy evaluation are presented in Table III.

TABLE III. EVALUATION RESULTS OF ENTROPY FOR AES AND SHIFTAES

Rounds	AES	ShiftAES
1	4.994138	5.014244
2	7.077322	7.040298
3	8	7.999285
4	7.999989	7.999987
5	7.999990	7.999990
6	7.999989	7.999989
7	7.999990	7.999989
8	7.999990	7.999988
9	7.999989	7.999988
10	7.999989	7.999988

The entropy evaluation results show that after three rounds, the entropy of data encrypted by both original AES and ShiftAES is approximately 8 bits/byte. This implies that both block ciphers achieve randomness properties with three or more rounds.

2) The evaluation results according to NIST SP 800-22: We conducted an evaluation of the statistical tests for rounds 1, 2, ..., 10 of both AES and ShiftAES. Tables IV to VIII display the results of the randomness evaluation of AES and ShiftAES across 1, 2, 3, 4, and 10 rounds.

TABLE IV. RANDOMNESS EVALUATION OF AES AND SHIFTAES OVER 1 ROUND

Test	AES-1R	ShiftAES-1R
AppEnt	0	0
BlocFreq	0	0
Cusum 1	0	0
Cusum 2	0	0
FFT	0	0
Freq	0	0
Linear Complexity	0.631288	0.752788
LongRun	0	0
NonOverLap	148 p-values = 0	148 p-values = 0
OverLap	0	0
RanEx	8 p-values = 0	8 p-values = 0
RanEx Var	18 p-values = 0	18 p-values = 0
Rank	0	0
Run	0	0
Serial 1	0	0
Serial 2	0	0
Universal	0	0

TABLE V. RANDOMNESS EVALUATION OF AES AND SHIFTAES OVER 2 ROUNDS

Test	AES-2R	ShiftAES-2R
AppEnt	0	0
BlocFreq	0	0
Cusum 1	0	0
Cusum 2	0	0
FFT	0	0
Freq	0	0
Linear Complexity	0.204973	0.051540
LongRun	0	0
NonOverLap	148 p-values = 0	148 p-values = 0
OverLap	0	0
RanEx	8 p-values = 0	3 p-values >= 0.01, 5 p-values < 0.01
RanEx Var	18 p-values = 0	18 p-values >= 0.01
Rank	0	0
Run	0	0
Serial 1	0	0
Serial 2	0	0
Universal	0	0

TABLE VI. RANDOMNESS EVALUATION OF AES AND SHIFTAES OVER 3 ROUNDS

Test	AES-3R	ShiftAES-3R
AppEnt	0.925352	0
BlocFreq	0.831754	0.999999
Cusum 1	0	0
Cusum 2	0	0
FFT	0	0
Freq	0.993653	0
Linear Complexity	0.978496	0.876043
LongRun	0.041736	0
NonOverLap	148 p-values >= 0.01	148 p-values <= 0.01
OverLap	0	0
RanEx	3 p-values >= 0.01, 5 p-values < 0.01	8 p-values = 0
RanEx Var	2 p-values >= 0.01, 16 < 0.01	18 p-values = 0
Rank	0.611499	0.662867
Run	0.340678	0
Serial 1	0.437274	0.029134
Serial 2	0.113612	0.577590
Universal	0.000003	0

TABLE VII. RANDOMNESS EVALUATION OF AES AND SHIFTAES OVER 4 ROUNDS

Test	AES-4R	ShiftAES-4R
AppEnt	0.898152	0.826485
BlocFreq	0.970639	0.862578
Cusum 1	0.254482	0.046664
Cusum 2	0.230798	0.010182
FFT	0.339743	0.896743
Freq	0.363824	0.034995
Linear Complexity	0.540695	0.080683
LongRun	0.162664	0.675598
NonOverLap	148 p-values >= 0.01	148 p-values >= 0.01
OverLap	0.220930	0.628086
RanEx	8 p-values >= 0.01	8 p-values >= 0.01
RanEx Var	18 p-values >= 0.01	18 p-values >= 0.01
Rank	0.615078	0.995817
Run	0.225007	0.415387
Serial 1	0.987050	0.564141
Serial 2	0.996948	0.403646
Universal	0.571751	0.494948

TABLE VIII. RANDOMNESS EVALUATION OF AES AND SHIFTAES OVER 10 ROUNDS

Test	AES-10R	ShiftAES-10R
AppEnt	0.863909	0.002023
BlocFreq	0.415025	0.510530
Cusum 1	0.537509	0.143294
Cusum 2	0.927344	0.306088
FFT	0.307167	0.782720
Freq	0.604613	0.234997
Linear Complexity	0.002433	0.057640
LongRun	0.748478	0.465346
NonOverLap	148 p-values >= 0.01	148 p-values >= 0.01
OverLap	0.000320	0.133895
RanEx	8 p-values >= 0.01	8 p-values >= 0.01
RanEx Var	18 p-values >= 0.01	18 p-values >= 0.01
Rank	0.908691	0.312501
Run	0.369725	0.584987
Serial 1	0.759251	0.618985
Serial 2	0.412087	0.396930
Universal	0.674145	0.165542

The results show that both the original AES and *ShiftAES* achieve randomness when they pass all the tests. For rounds fewer than four, both AES and *ShiftAES* still fail in some tests. From round 5 to round 10, both AES and *ShiftAES* block ciphers pass all tests. It means that the p-values of those tests are all greater than or equal to 0.01 from round 5 to round 10. Therefore, both block ciphers exhibit randomness when using four rounds or more.

Combining the evaluation results based on entropy and statistical tests from NIST SP 800-22, we conclude that the *ShiftAES* algorithm achieves randomness with a number of rounds greater than or equal to four and is equivalent to the original AES algorithm.

V. Conclusion

In this paper, we propose algorithms for generating key-dependent AddRoundKey and ShiftRow transformations based on permutations. Subsequently, we apply these key-dependent transformations to AES to create a dynamic AES block cipher. We conduct a security analysis and evaluate the statistical standards of NIST, assess the entropy of both AES and the dynamic AES (ShiftAES). Consequently, it is evident that the dynamic AES block cipher can significantly strengthen the security of AES and meets statistical randomness criteria similar to AES. This result is significant both in theory and practice, providing cryptographic researchers with a new method to improve block cipher security. Our future research direction involves further development of other dynamic algorithms to strengthen the resilience of SPN block ciphers against cryptanalysis.

Statements and Declarations

Funding: No funding was received for conducting this study.

Financial interests: The author declare they have no financial interests.

Competing interests: The author have no competing interests to declare that are relevant to the content of this article.

Research Data Policy and Data Availability Statements: Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

[1] R. Bharathi, R. and N. Parvatham, "LEA-Siot: Hardware architecture of lightweight encryption algorithm for secure IoT on FPGA platform," International Journal of Advanced Computer Science and Applications, vol. 11, 2020.

[2] A. Maetouq, S. M. Daud, N. A. Ahmad, N. Maarop, N. N. A. Sjarif and H. Abas, "Comparison of hash function algorithms against attacks: A review," International Journal of Advanced Computer Science and Applications, vol. 9, 2018.

[3] H. Abroshan, "A hybrid encryption solution to improve cloud computing security using symmetric and asymmetric cryptography algorithms," International Journal of Advanced Computer Science and Applications, vol. 12, pp. 31–37, 2021.

[4] L. T. Keliher, "Linear cryptanalysis of substitution-permutation networks," *Queen's University, Kingston, Ontario, Canada*, 2003.

[5] A. M. Youssef, S. E. Tavares and H. M. Heys, "A new class of substitution permutation networks," in Proceedings of Third Annual Workshop on Selected Areas in Cryptography (SAC 96), Queens University, Kingston, Canada, vol. 96, pp. 132–147, 1996.

[6] A. S. Alanazi, N. Munir, M. Khan and I. Hussain, "A novel design of audio signals encryption with substitution permutation network based on the Genesio-Tesi chaotic system," Multimedia Tools and Applications, pp. 1–17, 2023.

[7] S. Beg, N. Ahmad, A. Anjum, M. Ahmad, A. Khan, F. Baig and A. Khan, "S-box design based on optimize LFT parameter selection: a practical approach in recommendation system domain," Multimedia Tools and Applications, vol. 79, pp. 11667–11684, 2020. https://doi.org/10.1007/s11042-019-08464-6.

[8] S. Mister and C. Adams, "Practical S-box design," Workshop on Selected Areas in Cryptography, SAC, vol. 96, pp. 61–76, August 1996.

[9] A. M. Youssef and S. E. Tavares, "Resistance of balanced s-boxes to linear and differential cryptanalysis," Information Processing Letters, vol. 56, pp. 249–252, 1995. https://doi.org/10.1016/0020-0190(95)00156-6.

[10] S. Farwa, T. Shah, N. Muhammad, N. Bibi, A. Jahangir and S. Arshad, "An image encryption technique based on chaotic S-box and Arnold transform," International Journal of Advanced Computer Science and Applications, vol. 8, 2017.

[11] B. W. Koo, H. S. Jang and J. H. Song, "On constructing of a 32×32 binary matrix as a diffusion layer for a 256-bit block cipher," International Conference on Information Security and Cryptology, Springer, Berlin, Heidelberg, pp. 51–64, December 2006.

[12] M. KUMAR, P. YADAV, S. K. Pal and A. PANIGRAHI, "Secure and Efficient Diffusion Layers for Block Ciphers," Journal of Applied Computer Science & Mathematics, vol. 11, pp. 24, 2017.

[13] H. N. Noura and A. Chehab, "Efficient binary diffusion matrix structures for dynamic key-dependent cryptographic algorithms," Journal of Information Security and Applications, vol. 68, pp. 103264, 2022.

[14] G. Tuncay, F. B. Sakallı, M. K. Pehlivanoğlu, G. G. Yılmazgüç, S. Akleylek and M. T. Sakallı, M. T, "A new hybrid method combining search and direct based construction ideas to generate all 4× 4 involutory maximum distance separable (MDS) matrices over binary field extensions," PeerJ Computer Science, vol. 9, e1577, 2023.

[15] J. Daemen and V. Rijmen, "AES proposal: Rijndael (version 2)," NIST AES website, 1999.

[16] J. Daemen and V. Rijmen, "The design of Rijndael," New York: Springer-verlag, vol. 2, 2002.

[17] L. R. Knudsen and C. V. Miolane, "Counting equations in algebraic attacks on block ciphers," International Journal of Information Security, vol. 9, pp. 127–135, 2010. https://doi.org/10.1007/s10207-009-0099-9.

[18] H. M. Heys and S. E. Tavares, "The design of product ciphers resistant to differential and linear cryptanalysis," Journal of Cryptology, vol. 9, pp. 1–19, 1996.

[19] X. Lai, J. L. Massey and S. Murphy, "Markov ciphers and differential cryptanalysis," Advances in Cryptology—EUROCRYPT'91: Workshop on the Theory and Application of Cryptographic Techniques Brighton, UK, April 8–11, 1991 Proceedings 10, Springer Berlin Heidelberg, pp. 17–38, April 1991.

[20] A. Y. Al-Dweik, I. Hussain, M. Saleh and M. T. Mustafa, "A novel method to generate key-dependent s-boxes with identical algebraic properties," Journal of Information Security and Applications, vol. 64, p. 103065, 2022. https://doi.org/10.1016/j.jisa.2021.103065.

[21] K. Kazlauskas and J. Kazlauskas, "Key-dependent S-box generation in AES block cipher system," Informatica, vol. 20, pp. 23–34, 2009.

[22] P. Agarwal, A. Singh and A. Kilicman, "Development of key-dependent dynamic S-boxes with dynamic irreducible polynomial and affine constant," *Advances in Mechanical Engineering*, vol. 10, 2018. https://doi.org/10.1177/1687814018781638.

[23] A. Ejaz, I. A. Shoukat, U. Iqbal, A. Rauf and A. Kanwal, "A secure key dependent dynamic substitution method for symmetric cryptosystems," PeerJ Computer Science, vol. **7**, e587, 2021.

[24] G. Vaicekauskas, K. Kazlauskas and R. Smaliukas, "A novel method to design S-boxes based on key-depdendent permutation schemes and its quality analysis," International Journal of Advanced Computer Science and Applications, vol. 7, pp. 93–99, 2016.

[25] S. Murphy and M. J. B. Robshaw, "Key-dependent S-boxes and differential cryptanalysis," Designs, Codes and Cryptography, vol. 27, pp. 229–255, 2002.

[26] G. Murtaza, A. A. Khan, S. W. Alam and A. Farooqi, "Fortification of AES with dynamic mix-column transformation," Cryptology ePrint Archive, 2011.

[27] T. T. Luong, "Building the dynamic diffusion layer for SPN block ciphers based on direct exponent and scalar multiplication," Journal of Science and Technology on Information security, vol. 1, pp. 38–45, 2022.

[28] M. R. M. Shamsabad and S. M. Dehnavi, "Dynamic MDS diffusion layers with efficient software implementation," International Journal of Applied Cryptography, vol. 4, pp. 36–44, 2020. https://doi.org/10.1504/IJACT.2020.107164.

[29] T. Xu, F. Liu and C. Wu, "A white-box AES-like implementation based on key-dependent substitution-linear transformations," Multimedia Tools and Applications, vol. 77, pp. 18117–18137, 2018. https://doi.org/10.1007/s11042-017-4562-8.

[30] V. Sawant, A. Solkar and R. Mangrulkar, "Modified Symmetric Image Encryption Approach Based on Mixed Column and Substitution Box," Journal of Applied Security Research, pp. 1–34, 2022. https://doi.org/10.1080/19361610.2022.2150498.

[31] A. Altigani, S. Hasan, B. Barry, S. Naserelden, M. A. Elsadig and H. T. Elshoush, "A polymorphic advanced encryption standard–a novel approach," IEEE Access, vol. 9, pp. 20191–20207, 2021. DOI: 10.1109/ACCESS.2021.3051556.

[32] T. Manoj Kumar and P. Karthigaikumar, "A novel method of improvement in advanced encryption standard algorithm with dynamic shift rows, sub byte and mixcolumn operations for the secure communication," International Journal of Information Technology, vol. 12, pp. 825–830, 2020.

[33] A. I. Salih, A. Alabaichi and A. S. Abbas, "A novel approach for enhancing security of advance encryption standard using private XOR table and 3D chaotic regarding to software quality factor," ICIC Express Letters Part B: Applications, An International Journal of Research and Surveys, vol. 10, pp. 823–832, 2019. DOI: 10.24507/icicelb.10.09.823.

[34] A. I. Salih, A. M. Alabaichi and A. Y. Tuama, "Enhancing advance encryption standard security based on dual dynamic XOR table and mixcolumns transformation," Indonesian Journal of Electrical Engineering and Computer Science, vol. 19, pp. 1574–1581, 2020. DOI: 10.11591/ijeecs.v19.i3.

[35] M. Sajadieh, M. Dakhilalian, H. Mala and B. Omoomi, "On construction of involutory MDS matrices from Vandermonde Matrices in GF (2^q)," Designs, Codes and Cryptography, vol. 64, pp. 287–308, 2012.

[36] M. Matsui, "Linear cryptanalysis method for DES cipher," Advances in Cryptology—EUROCRYPT'93: Workshop on the Theory and Application of Cryptographic Techniques Lofthus, Norway, Proceedings 12, Springer Berlin Heidelberg, pp. 386–397, May 1993.

[37] J. Guo, L. Song and H. Wang, "Key structures: Improved related-key boomerang attack against the full AES-256," In Australasian Conference on Information Security and Privacy, Cham: Springer International Publishing, pp. 3–23, November 2022.

[38] L. E. Bassham III, A. L. Rukhin, J. Soto, J. R. Nechvatal, M. E. Smid, E. B. Barker, and et al, "Sp 800-22 rev. 1a. a statistical test suite for random and pseudorandom number generators for cryptographic applications," National Institute of Standards & Technology, 2010.

[39] C. E. Shannon, "A mathematical theory of communication, Bell System Technical Journal 27," pp. 379–423 and 623–656, 1948.

[40] W. John, "ENT A Pseudorandom Number Sequence Test Program", 2008. Retrieved September, 3, 2022.

Selection of Unmanned Aircraft Development Model in Indonesia using the AHP Method

Agus Bayu Utama[1]*, Siswo Hadi Sumantri[2], Romie Oktovianus Bura[3], Gita Amperiawan[4]
Aeronautics Technology Research Center, National Research and Innovation Agency (BRIN), Jakarta, Indonesia[1]
Defense Science Doctoral Program, Republic of Indonesia Defence University, Jakarta, Indonesia[1, 2, 3, 4]

***Abstract*—Countries worldwide are attempting to acquire or create Class 3 unmanned aircraft as part of their armies' primary weapons systems. The development of medium altitude long endurance (MALE) unmanned aircraft in Indonesia forms part of the national strategic program. Based on documentation studies, three alternative MALE-class unmanned aircraft development models were identified. This study aims to determine the most appropriate unmanned aircraft development model for the MALE class for Indonesia's current situation. This will aid decision-making by the government and stakeholders related to the drone development model. The analytical hierarchy process (AHP) method was used to analyze the decision-making for the selection of an unmanned aircraft development model. The study began with a questionnaire survey of 11 experts from various institutions. The results show that the priority criterion should be the benefits obtained, followed by the opportunity and budget criteria, and, finally, the risk. The consortium model, which had the highest score of 0.548, is the most suitable for Indonesia's development of MALE-class unmanned aircraft. The results of the study are expected to provide useful input for AHP researchers, government institutions, and stakeholders.**

***Keywords*—Analytical Hierarchy Process (AHP); decision-making; development model; medium altitude long endurance (MALE); unmanned aircraft**

I. Introduction

Countries worldwide are attempting to acquire or create Class 3 unmanned aircraft as part of their armies' primary weapons systems. More and more nations are trying to compete in the race to create and offer ever-more advanced drones for sale on the international market. Long-simmering conflicts and rivalries are expected to change due to the emergence of a race for drones [1]. The primary rivalries in this arms race are between the United States of America, China, Russia, South Korea, and the European Union [2]. Unmanned aircraft are divided into three classes based on maximum take-off weight, namely Class 1 (<150 kg), Class 2 (150–600 kg), and Class 3 (>600 kg) [3]. Countries around Indonesia such as China, India, Pakistan, Singapore, and Thailand already have Class 3 unmanned aircraft. Meanwhile, countries that are growing and intend to obtain them include Australia, Bangladesh, Japan, Malaysia, the Philippines, South Korea, Taiwan, and Indonesia [4].

Examples of the deployment of unmanned aircraft in warfare include the United States of America's use of the Predator MQ-1 against Osama bin Laden in Afghanistan beginning in 2002. The Turkish military used Bayraktar TB2s against Syrian Army targets in March 2020 [5]. During the Nagorno–Karabakh war of 2020, Azerbaijan employed the Bayraktar TB2 against the Armed Forces of Armenia [6]. More recently, Russia has reduced the military power of Ukraine. Drones have played a significant role in the former's operations and Russia has improved the aiming speed for long-range indirect fire through the use of multiple drone types that fly at varying altitudes [7]. The conflict in Ukraine is swiftly bringing future trends for drone use into view [8]. The use of an Unmanned Aerial System (UAS) in a hostile environment reduces risk [9]. Drones can be used to solve a variety of problems related to the needs of the users, such as illegal fishing, illegal immigrants, piracy, floods, forest fires, terrorism, and military infiltration of other countries [10].

According to Presidential Regulation of the Republic of Indonesia, number 109 for the year 2020, numerous inventions and research projects are being executed efficiently to strengthen the country's technological independence. The national strategic program includes the development of medium altitude long endurance (MALE) unmanned aircraft in Indonesia. The Black Eagle unmanned aerial vehicle (UAV) system is part of several governmental research and innovation initiatives [11]. Based on documentation studies, three alternative development models for MALE-class unmanned aircraft have been identified, namely: 1) The BRIN model based on a letter from the Deputy for Research and Innovation Utilization of the National Research and Innovation Agency (BRIN) to the Secretary General of the Ministry of Defense dated 23-11-2021 regarding an explanation of the follow-up to the MALE UAV program, 2) A consortium model based on Minister of Research, Technology and Higher Education Regulation no. 38 of 2019 concerning the 2020-2024 national research program, the Agency for the Assessment and Application of Technology (BPPT) as coordinator of the combat PTTA development program, and based on a cooperation agreement (PKS) between the Ministry of Defense, Indonesian National Army Air Force (TNI AU), BPPT, Bandung Institute of Technology (ITB), PT Dirgantara Indonesia (PTDI), and PT LEN dated 08-21-2017, and 3) An international cooperation model based on the invitation letter and minutes of the coordination meeting for the implementation of the offset procurement of the Bayraktar TB2 UAV from the Director General of Defense Ministry of Defense dated 01-03-21.

The consortium-based MALE class UCAV development program started in 2017 but has been discontinued since 2020 due to the establishment of the National Research and Innovation Agency (BRIN). This study aims to determine the

most appropriate unmanned aircraft development model for the MALE class for Indonesia's current situation. This will aid decision-making by the government and stakeholders related to the drone development model. In this case, an analytical hierarchy process (AHP) based on the criteria provided for selecting current models can aid in making a superior choice.

The following section contains a literature review on UAV design, opportunities, product budgets, benefits, risks, and the AHP method. Section III presents the methodology: system design, the AHP method, sampling and data collection, research instruments and measurements, and methods of analysis. The findings of the analysis are discussed in Section IV. Finally, the study will end with conclusions and recommendations for future research in Section V.

II. Literature Review

Manned aircraft and UAV designs share certain similarities and contrasts. They both consist of a design procedure, limitations, and core UAV parts (autopilot, ground station, communication, sensors, and payload). A UAV designer must be well versed in the most recent UAV advancements, modern technologies, the lessons learned from previous failures, and the diversity of UAV design possibilities, and they must understand the environment, the requirements, and the design problems, as well as how to integrate complex, multi-disciplinary systems [12]. A step-by-step layout design study was conducted using precision tools and computational simulations to define the essential layout parameters and choose the ideal airframe-engine combination [13].

When Inertial Navigation System data are not sufficiently precise to minimize drift from a planned trajectory, Synthetic Aperture Radar (SAR) can assist in UAV navigation [14]; [15]. Stability and maneuverability are key trade-offs in aircraft construction for civil or military use. The UAV and its handling characteristics are both addressed in the design of the flight control system [16]. While every technological product carries the risk of malfunction or operational error, pilots can respond to a failure by using the best mitigation that has been designed, such as by disabling automatic power control or completing the operation manually [17].

The initial model in the B787 Dreamliner program represented a disruptive technology product innovation within the industry. The program fundamentally altered the supply chain of the industry partnership model [18]. Many strategies and thematic lenses have arisen in the academic community to address various questions in innovation management [19]. Similarities exist in the new product development innovation process within the stages of identifying issues and opportunities, creating and processing ideas, market projections, business analysis, visualization, and execution, as well as expressing the model in service organizations [20]. The current research contains various outstanding challenges concerning the discovery of technology opportunities [21]. Only an appropriate product structure will allow the benefits of product portfolio management to be fully realized [22]. With a restricted budget, a government could determine the appropriate level of subsidy to boost the sales of remanufactured products. The best subsidy is offered when a budget is limited [23]. The creation of a project selection and evaluation tool may potentially be applied to a wide range of research, technology, and investment decisions [24].

An organization that creates a new product invests time and money in the hope that the product will provide a sufficient return on investment. An effective and efficient risk management strategy must be selected to match the specific product development case. Firms that develop products use risk identification, assessment, and mitigation to support their risk management procedures and decision-making [25], where the risks are detectable and manageable. Product development industries are also assisted in incorporating sustainability into strategic, tactical, and operational decision-making [26].

The AHP method is effective for analyzing a complex problem involving the selection of an alternative as a decision from several choices. A problem's complexity stems partially from the presence of numerous influencing criteria. AHP offers a means of breaking down a complicated unstructured scenario into multiple components in a hierarchical order by assigning a subjective value to the relative relevance of each variable and identifying which variable has the highest priority in terms of impacting the outcome of the situation [27]. The AHP method developed by Professor Thomas L. Saaty is widely used for business purposes in companies, government interests, and research.

The AHP method has been widely used in several research fields, for example: The AHP method enables for the evaluation and rating of the device's design elements, resulting in a more reasonable and comprehensive device design [28], The VAHP model is a helpful decision-making tool for transportation planners, policymakers, and other stakeholders in the industry [29].

III. Methodology

The research method employed in this study combined the survey method by distributing questionnaires to experts in the field of UAV development with the AHP method to determine the global priority of the alternatives offered. The research flow is shown in Fig. 1. A literature review was conducted in step one, the criteria for the alternatives presented were determined in step two, a questionnaire was developed in step three and distributed to expert respondents in step four, the data were analyzed in step five, and conclusions were drawn from the research findings in step six. The research questionnaire can be seen in Appendix 1.

A. System Design

In this study, four important criteria were determined, namely: 1) Opportunity (Op): With government policies, sufficient and qualified human resources, and infrastructure for product development, the opportunity is enormous, 2) Budget (Bu): Budgets are needed for initial investment, expert and employee salaries, tool and material purchases, and testing costs, 3) Benefits (B): These include speedier technological mastery, achieving defense industry independence, avoiding embargoes, creating jobs, and producing foreign exchange, and 4) Risk (Ri): The identified risks would manifest in forms such as the program not running smoothly, the goal time being delayed, the product failing, or there being no partner. Fig. 2

shows the AHP structure chart with the existing goals, criteria, and alternatives.

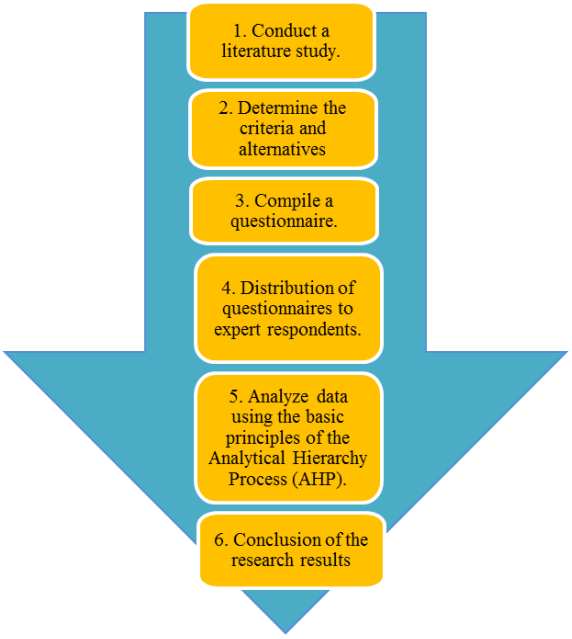


Fig. 1. Research flow.

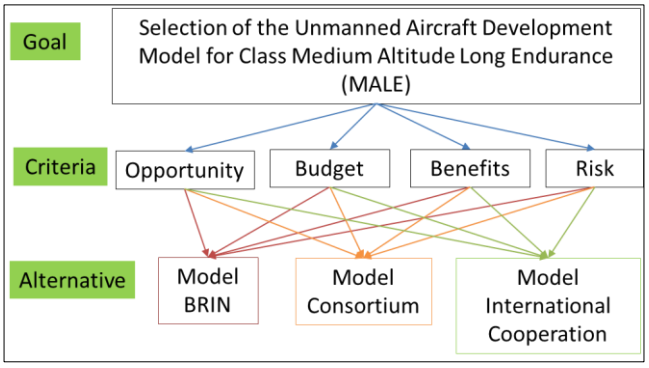


Fig. 2. AHP structure chart.

B. Sampling and Data Collection

The population in this study comprised individuals involved in national research and development on MALE-class drones and their use on the territory of the Republic of Indonesia. Samples were selected by purposive sampling, which is a technique suitable for data sources with specific considerations, such as identifying the person considered to be the most knowledgeable in the area under consideration [30]. Some questionnaires were distributed online to respondents while other respondents were visited face to face. The data were collected from July to August 2022. While the AHP method has no specific formulation for the number of respondents, it must include a minimum of two. The priority in applying the AHP method is the quality of the data derived from the respondents, not the quantity. The respondents in AHP assessments must be experts in order to decide between the alternatives proposed. The experts in this case were competent individuals who had truly mastered and comprehended the situation, influenced policymaking, or knew the information needed.

Table I contains the characteristics of the respondents and shows that 36.4% had undergraduate or diploma degrees while 63.6% had postgraduate degrees, thus indicating a group of participants with a high level of education. In terms of the respondents' age, 18.2% of the population were aged between 41 and 50, 63.6% were aged between 51 and 60, and 9.1% were aged between 61 and 75.

TABLE I. CHARACTERISTICS OF THE RESPONDENTS

Respondent	Age (years)	Education level	Institutions
1	51–60	S3	Head of Aeronautics Technology Research Center, Nasional Research and Innovation Agency (BRIN)
2	51–60	S3	Member of the House Of Representatives - Commission VII (Energy, Mineral Resources, Research and Technology, Environmental Affairs)
3	51–60	S2	Secretary General of Indonesian Aeronautical Engineering Center (IAEC)
4	41–50	S1	Chairman of Unmanned Technology Systems Association (ASTTA) and Director of PT Aeroterra Indonesia
5	31–40	S1	Senior Vice President (SVP) Technology and Research Development Centre of PT LEN Industry
6	51–60	S2	Director of Commerce, Technology, and Development of PT. Dirgantara Indonesia (Indonesian aircraft industries)
7	51–60	S1	Head of Combat Power Division, Research, and Development Agency of the Ministry of Defense
8	51–60	S3	Head of sub-directorate Europe and Africa, Directorate of Defense International Cooperation, Directorate General of Defense Strategy Ministry of Defense
9	41–50	S3	Lecturer at the Faculty of Mechanical and Aerospace Engineering and Head of the Center for Unmanned Studies, Bandung Institute of Technology (ITB)
10	61–75	S3	Head of Technology Transfer and Offset Defense Industry Policy Committee
11	51–60	S1	Head of the Indonesian Air Force's Research and Development Department

C. Research Instruments and Measurements

The research instrument used was a questionnaire with a comparison matrix between the criteria (four criteria) and a comparison matrix between the alternatives (three alternatives) based on the four existing criteria. This study used the nine-point Saaty scale, where 1: equal importance, 3: moderate importance, 5: high importance, 7: very high importance, 9: extreme importance, and 2, 4, 6, 8: intermediate values. In addition, demographic questions were asked, such as age, education, and employment institutions. The research questionnaire is shown in Appendix 1. The researcher then inputted data from the questionnaire results

into a matrix table in Excel software based on the stages in the AHP method (see Appendices 2 to 6).

D. Methods of Analysis

Fig. 3 illustrates the 11 main steps used to obtain the priorities for the unmanned aircraft development model evaluation using AHP, as in [31]; [32]. The study in [33] has written that The AHP is sufficient for decision-making.

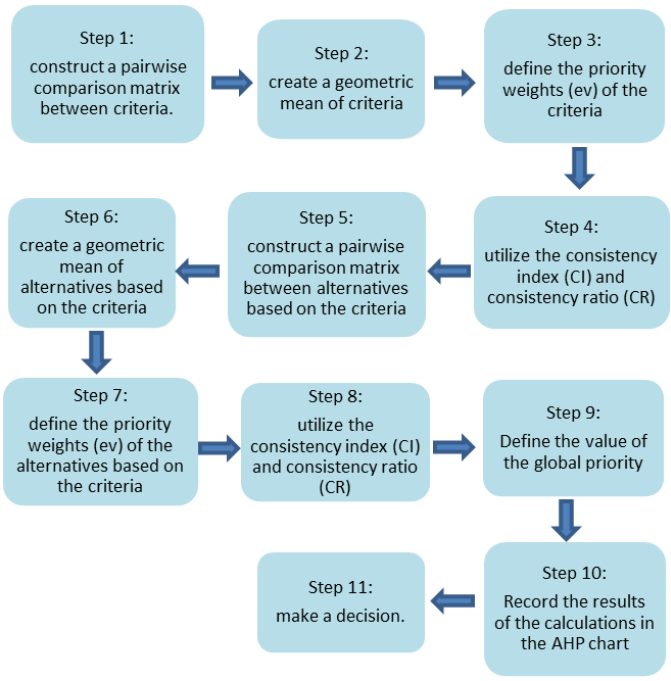


Fig. 3. The main steps used to obtain the global priority.

1) Step 1: Construct a pairwise comparison matrix between the criteria.

Given that there are m responders, there are m pairwise comparison matrixes with n criteria, as shown in Table II below, where, R denotes Respondent:

TABLE II. PAIRWISE COMPARISON MATRIX BETWEEN CRITERIA: INPUT ANSWERS FROM ALL RESPONDENTS

R1	C1	C2	...	Cn		Rm	C1	C2	...	Cn
C1	1	X_{12}	...	X_{1n}		C1	1	X_{12}	...	X_{1n}
C2	$1/X_{12}$	1	...	X_{2n}		C2	$1/X_{12}$	1	...	X_{2n}
...	...	...	1	...		...	...	...	1	...
Cn	$1/X_{1n}$	$1/X_{2n}$	...	1		Cn	$1/X_{1n}$	$1/X_{2n}$	...	1

2) Step 2: Create a geometric mean of criteria

After entering the comparative assessment into the matrix above, an average measurement using the geometric mean (see Appendix 7) (GM) of the m respondents' replies was entered using the formula below, with the values shown in Table III:

$$GM_{12} = \sqrt[m]{X_{12(R1)} * X_{12(R2)} * X_{12(R3)} * \dots\dots * X_{12(Rm)}} \quad (1)$$

$$GM_{21} = \sqrt[m]{X_{21(R1)} * X_{21(R2)} * X_{21(R3)} * \dots\dots * X_{21(Rm)}} \quad (2)$$

$$GM_{n1} = \sqrt[m]{X_{n1(R1)} * X_{n1(R2)} * X_{n1(R3)} * \dots\dots * X_{n1(Rm)}} \quad (3)$$

$$GM_{1n} = \sqrt[m]{X_{1n(R1)} * X_{1n(R2)} * X_{1n(R3)} * \dots\dots * X_{1n(Rm)}} \quad (4)$$

TABLE III. GEOMETRIC MEAN OF M MATRIX OF RESPONDENTS' ANSWERS

GM	C1	C2	...	Cn
C1	1	GM_{12}	...	GM_{1n}
C2	GM_{21}	1	...	GM_{2n}
...	...	...	1	...
Cn	GM_{n1}	GM_{n2}	...	1

3) Step 3: Define the priority weights (ev) of the criteria

Formulas (5) and (6) below were used to calculate and assign the priority weights:

$$\begin{bmatrix} 1 & GM12 & \dots & GM1n \\ GM21 & 1 & \dots & GM2n \\ \dots & \dots & 1 & \dots \\ GMn1 & GMn2 & \dots & 1 \end{bmatrix} \times \begin{bmatrix} 1 & GM12 & \dots & GM1n \\ GM21 & 1 & \dots & GM2n \\ \dots & \dots & 1 & \dots \\ GMn1 & GMn2 & \dots & 1 \end{bmatrix} = \begin{bmatrix} Z11 & Z12 & \dots & Z1n \\ Z21 & Z22 & \dots & Z2n \\ \dots & \dots & \dots & \dots \\ Zn1 & Zn2 & \dots & Znn \end{bmatrix} \quad (5)$$

the GM sum in the 1 row	priority weights	criterion or alternative
Y1 = Z11 + Z12 + ... + Z1n	ev1 = Y1 / Q	C1 or A1
Y2 = Z21 + Z22 + ... + Z2n	ev2 = Y2 / Q	C2 or A2
Y... = Z.... + Z.... + ... + Z...n	ev...= Y.../ Q	....
Yn = Zn1 + Zn2 + ... + Znn	Evn = Yn / Q	Cn or An
total: Q = Y1 + Y2 + Y... + Yn		

(6)

4) Step 4: Utilize the consistency index (CI) and consistency ratio (CR) to evaluate logical consistency via the following steps:

a) Calculate the value of Vector [A]; Multiplication of the geometric mean (GM) matrix with the weight matrix priority (ev) = Vector [A]:

$$\begin{bmatrix} 1 & GM12 & \dots & GM1n \\ GM21 & 1 & \dots & GM2n \\ \dots & \dots & 1 & \dots \\ GMn1 & GMn2 & \dots & 1 \end{bmatrix} * \begin{bmatrix} ev1 \\ ev2 \\ \dots \\ evn \end{bmatrix} = \begin{bmatrix} A1 \\ A2 \\ \dots \\ An \end{bmatrix} \quad (7)$$

b) Calculate the value of Vector [B];

$$Vector\ |B| = \left| \frac{A1}{ev1} \quad \frac{A2}{ev2} \quad \frac{\dots}{\dots} \quad \frac{An}{evn} \right| \quad (8)$$

c) Calculate the maximum Eigenvalue:

$$\lambda_{max} = \frac{summary\ of\ the\ value\ of\ Vector\ B}{number\ of\ criteria\ (n)} \quad (9)$$

d) Calculate CI:

$$CI = \frac{\lambda_{max} - n}{n-1} \quad (10)$$

e) Determine the Random consistency index (RCI):

n (criteria)	1	2	3	4	5	6	7	8	9	10
RCI	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

f) Calculate CR:

$$CR = \frac{CI}{RCI} \quad (11)$$

If the CR value does not exceed 10% or < 0.1, then the pairwise comparisons between criteria data are consistent/valid.

5) Step 5: Construct a pairwise comparison matrix between alternatives based on the criteria

6) Step 6: Create GM of alternatives based on the criteria

7) Step 7: Define the priority weights (ev) of alternatives based *on the criteria*

8) Step 8: Utilize the consistency index (CI) and consistency ratio (CR) to evaluate logical consistency; The processes in steps 5 to 8 are the same as in steps 1 to 4.

9) Step 9: Define the value of the global priority; Global priority = priority weight of each alternative multiplied by the priority weight of the criteria.

*10)*Step 10: Record the results of the calculations in the AHP *structure chart; and finally*

11)Step 11: Make a decision.

IV. Results And Discussion

A. Priority Weights of Criteria

Table IV shows the priority weights of the criteria. The benefits criterion obtained the highest ratings with a priority weight value of 0.471. This was followed by opportunity, budget, and risk, respectively. Meanwhile, the risk criterion has a very small priority weight, namely 8.8%. The development of MALE-class drones will have significant benefits for the nation, including speedier technological mastery, defense industry independence, and the avoidance of embargoes, job creation, and foreign exchange production.

TABLE IV. Priority Weights of the Criteria

	Op	Bu	Be	Ri	Sum	Priority weight (ev)	Rank
Op	0.255	0.179	0.278	0.338	1.050	0.262	2
Bu	0.243	0.171	0.149	0.148	0.711	0.178	3
Be	0.436	0.548	0.476	0.426	1.885	0.471	1
Ri	0.066	0.102	0.098	0.088	0.354	0.088	4
Sum	1	1	1	1	4	1	

B. Priority Weights of Alternatives Based on the Opportunity Criterion

Based on the calculations in Table V, the following priorities among the alternatives were obtained regarding the opportunity criterion: 1) The consortium (CO) model with a priority weight value of 0.595, 2) The international cooperation (IC) model with a priority weight value of 0.240, and 3) The BRIN (BR) model with a priority weight value of 0.165. A CO model for the development of MALE-class drones has a better chance of success, with the smallest opportunity for the BRIN model.

C. Priority Weights of Alternatives Based on the Budget Criterion

Using the calculations in Table VI, the following priorities among the alternatives were obtained based on the budget criterion: 1) The CO model with a priority weight value of 0.483, 2) The IC model with a priority weight value of 0.357, and 3) The BR model with a priority weight value of 0.160. A larger budget is required to develop the MALE-class drone using a CO model. Budgets are required for initial investment, tool and material purchases, and testing costs. The BRIN model requires the smallest budget.

TABLE V. Priority Weights Of Alternatives Based On The Opportunity Criterion

	BR	CO	IC	Sum	Priority weight (ev)	Rank
BR	0.158	0.148	0.189	0.495	0.165	3
CO	0.633	0.590	0.562	1.785	0.595	1
IC	0.209	0.262	0.249	0.720	0.240	2
Sum	1	1	1	3	1	

TABLE VI. Priority Weights of Alternatives based on the Budget Criterion

	BR	CO	IC	Sum	Priority weight (ev)	Rank
BR	0.158	0.147	0.175	0.479	0.160	3
CO	0.512	0.476	0.460	1.449	0.483	1
IC	0.330	0.377	0.365	1.072	0.357	2
Sum	1	1	1	3	1	

D. Priority Weight of Alternatives Based on the Benefits Criterion

Based on the results of the calculations in Table VII, the following alternative priorities were obtained when considering the benefits criterion: 1) The CO model with a priority weight value of 0.588, 2) The IC model with a priority weight value of 0.260, and 3) The BR model with a priority weight value of 0.152. As such, the use of the CO model when developing the MALE-class drone would yield greater benefits, with the BRIN model providing the lowest benefits.

TABLE VII. Priority Weights of Alternatives based on the Benefits Criterion

	BR	CO	IC	Sum	Priority weight (ev)	Rank
BR	0.143	0.129	0.183	0.456	0.152	3
CO	0.642	0.579	0.543	1.764	0.588	1
IC	0.214	0.292	0.274	0.780	0.260	2
Sum	1	1	1	3	1	

E. Priority Weights of Alternatives Based on the Risk Criterion

The calculation results in Table VIII show that the risk criterion produced the alternative priorities as follows: 1) The IC model with a priority weight value of 0.453, 2) The CO

model with a priority weight value of 0.326, and 3) The BR model with a priority weight value of 0.221. The development of a MALE-class drone with an IC model would carry greater risk, such as the program not running smoothly, a delayed goal time, and product failure. In contrast, the lowest risk is with the BRIN model.

TABLE VIII. PRIORITY WEIGHTS OF ALTERNATIVES BASED ON THE RISK CRITERION

	BR	CO	IC	Sum	Priority weight (ev)	Rank
BR	0.221	0.228	0.216	0.664	0.221	3
CO	0.317	0.328	0.333	0.978	0.326	2
IC	0.462	0.444	0.451	1.358	0.453	1
Sum	1	1	1	3	1	

F. Verify the CR

Tables IX and X show the results obtained from using the CR of the criteria and the alternatives. A CR value < 0.1 indicates that both the pairwise comparison data between the criteria and the pairwise comparisons between the alternatives are consistent or valid.

TABLE IX. RESULT OF UTILIZING THE CONSISTENCY RATIO (CR) OF CRITERIA

Vector [A]	Vector [B]	λ_{max}	CI	RCI	CR
$\begin{bmatrix}1.063\\0.725\\1.920\\0.359\end{bmatrix}$	|4.057 4.073 4.076 4.080|	4.0715	0.0238	0.9	0.0264

TABLE X. RESULT OF UTILIZING THE CONSISTENCY RATIO (CR) OF ALTERNATIVES

Based on criteria	Vector [A]	Vector [B]	λ_{max}	CI	RCI	CR
Opportunity	$\begin{bmatrix}0.496\\1.796\\0.722\end{bmatrix}$	|3.005 3.017 3.00	3.010	0.0049	0.58	0.0084
Budget	$\begin{bmatrix}0.479\\1.453\\1.074\end{bmatrix}$	|3.003 3.005 3.00	3.005	0.0024	0.58	0.0042
Benefit	$\begin{bmatrix}0.457\\1.784\\0.783\end{bmatrix}$	|3.009 3.034 3.01	3.019	0.0095	0.58	0.0164
Risk	$\begin{bmatrix}0.664\\0.978\\1.358\end{bmatrix}$	|3.001 3.001 3.00	3.001	0.0003	0.58	0.0005

G. Global Priority

Table XI shows that the CO model is the top priority with a score of 0.548, followed by the IC model with a value of 0.289, and, finally, the BR model with a value of 0.163. Fig. 4 shows the AHP structure chart with the priority weights of the criteria and the global priority for the development model of MALE-class unmanned aircraft.

A consortium is a collective structure, collaboration, or cooperation of individuals or institutions. Previous studies have reported its use in a variety of fields, including the tourism market [34], electronic products [35], new medical product development for transplant patients [36], and in the form of an agricultural industrialization consortium in low carbon agriculture [37].

TABLE XI. GLOBAL PRIORITY

	Priority weight (ev) of alternatives					Priority weight (ev) of criteria		Global priority
	Op	Bu	Be	Ri				
Br	0.165	0.160	0.152	0.221		0.262		0.163
CO	0.595	0.483	0.588	0.326	X	0.178	=	0.548
IC	0.240	0.357	0.260	0.453		0.471		0.289
						0.088		

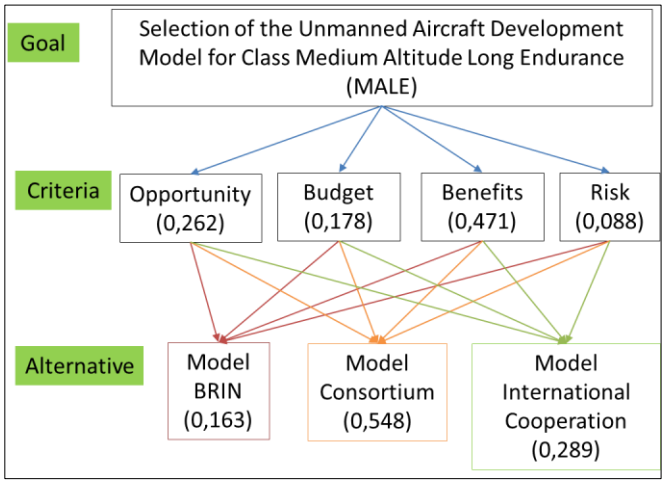


Fig. 4. AHP structure chart with priority weights of criteria and global priority.

V. CONCLUSION

This study sought to identify the most appropriate development model for the MALE class of unmanned aircraft using the AHP method. Based on the four criteria examined, the results show that the alternatives should be considered in the following order of priority: the benefits obtained, the opportunity, the budget required, and, finally, the risk that will arise. The results of the global priority calculation show that the consortium model has the highest value, at 0.548. It can therefore be concluded that the consortium model is the most suitable for the development of MALE-class unmanned aircraft in Indonesia.

A. Limitation

This study focused on respondents who were involved in national research and development for MALE-class drones, as well as their application in the Republic of Indonesia. The sample may thus differ from that of other countries and the findings of the study may not apply to all countries.

B. Recommendations for Future Research

It is recommended that future studies continue to examine the consortium model by studying and exploring the benefits and drawbacks of current such programs. It is hoped that this will provide input and guidance for implementing the next consortium program and help to ensure that the MALE-class

unmanned aircraft development program runs smoothly and as expected.

ACKNOWLEDGMENT

This study was supported by the Republic of Indonesia Defense University, and the National Research and Innovation Agency (BRIN). We are grateful to the director of the doctoral program, and the head of the Aeronautics Technology Research Center, for their assistance with this paper.

REFERENCES

[1] M. J. Boyle, "The race for drones," Orbis, vol. 59, no. 1, pp. 76–94, 2014, doi: 10.1016/j.orbis.2014.11.007.

[2] J. Haner and D. Garcia, "The artificial intelligence arms race : Trends and world leaders in autonomous weapons development," vol. 10, no. 3, pp. 331–337, 2019, doi: 10.1111/1758-5899.12713.

[3] NATO, "STANAG 4671 – Unmanned aircraft systems airworthiness requirements, Annex A," February., 2017.

[4] D. Gettinger, The drone databook. New York: The Center For The Study of The Drone At Bard College, 2019.

[5] A. B. Utama and S. Anwar, "History of the use of unmanned aerial vehicle (UAV) in the modern war and the preparation of the Indonesian military," J. Pertahanan Bela Negara Univ. Pertahanan RI, vol. 11, no. 3, pp. 167–181, 2021.

[6] A. K. Jati, E. Ashyaningtyas, H. Nurhan, and H. A. Fanfa, "Analisis keterlibatan Turki dalam konflik Nagorno-Karabakh: Studi kasus September war 2020," INTELEKTIVA J. Ekon. Sos. dan Hum., vol. 3, no. 5, pp. 14–25, 2022, [Online]. Available: https://jurnalintelektiva.com/index.php/jurnal/article/view/681.

[7] R. G. Angevine, T. Lead, J. K. Warden, R. Keller, and C. Frye, "Learning lessons from the Ukraine conflict," Virginia, USA, 2019.

[8] D. Kunertova, "The war in Ukraine shows the game-changing effect of drones depends on the game," Bull. At. Sci., vol. 79, no. 2, pp. 95–102, 2023, doi: 10.1080/00963402.2023.2178180.

[9] V. Prisacariu and A. Muraru, "Unmanned aerial system (UAS) in the context of modern warfare," Sci. Res. Educ. Air Force, vol. 18, no. 1, pp. 177–184, 2016, doi: 10.19062/2247-3173.2016.18.1.23.

[10] D. Lesmana, Y. Permana, B. Santoso, and A. Infantono, "Military drone applications by using Indonesian defense equipment for over the horizon operations," in Prosiding Seminar Nasional Sains Teknologi dan Inovasi Indonesia (SENASTINDO), Dec. 2021, vol. 3, pp. 1–10, doi: 10.54706/senastindo.v3.2021.149.

[11] R. G. Debe and A. R. Ras, "Development of the PUNA MALE Elang Hitam combatant classification in order to strengthen Indonesia's defense and security (Pengembangan PUNA MALE Elang Hitam klasifikasi kombatan dalam rangka memperkuat pertahanan dan keamanan Indonesia)," J. Pendidik. Tambusai, vol. 6, pp. 8900–8908, 2022.

[12] M. Sadraey, Unmanned aircraft design. Southern New Hampshire: Morgan & Claypool, 2017.

[13] P. Panagiotou, E. Giannakis, G. Savaidis, and K. Yakinthos, "Aerodynamic and structural design for the development of a MALE UAV," Aircr. Eng. Aerosp. Technol., vol. 90, no. 7, pp. 1077–1087, Nov. 2018, doi: 10.1108/AEAT-01-2017-0031.

[14] D. Nitti, F. Bovenga, M. Chiaradia, M. Greco, and G. Pinelli, "Feasibility of using synthetic aperture radar to aid UAV navigation," Sensors, vol. 15, no. 8, pp. 18334–18359, Jul. 2015, doi: 10.3390/s150818334.

[15] M. Soleh and R. Arief, "Analysis of SAR main parameters for SAR sensor design on LSA," Int. J. Remote Sens. Earth Sci., 2017, doi: 10.30536/j.ijreses.2014.v11.a2606.

[16] M. Kadiri, A. Mohammed, and S. Sanusi, "Validation of aerodynamic coefficients for flight control system of a medium altitude long endurance unmanned aerial vehicle," in 2019 2nd International Conference of the IEEE Nigeria Computer Chapter (NigeriaComputConf), Oct. 2019, pp. 1–4, doi: 10.1109/NigeriaComputConf45974.2019.8949628.

[17] I. Novhela and S. Martini, "The mitigation design of failure conditions level system with System Functional Hazard Assessment (SFHA) on unmanned aircraft MALE class," Sci. Res. J. (SCIRJ), vol. VIII, no. Xii, pp. 50–59, 2020, doi: 10.31364/SCIRJ/v8.i12.2020.P1220828.

[18] L. Cantone, P. Testa, S. Hollensen, and G. F. Cantone, "Outsourcing new product development fostered by disruptive technological innovation: A decision-making model," Int. J. Innov. Manag., vol. 23, no. 01, p. 1950008, Jan. 2019, doi: 10.1142/S1363919619500087.

[19] S. Kavadias and K. T. Ulrich, "Innovation and new product development: Reflections and insights from the research published in the first 20 years of Manufacturing & Service Operations Management," Manuf. Serv. Oper. Manag., vol. 22, no. 1, pp. 84–92, Jan. 2020, doi: 10.1287/msom.2019.0816.

[20] T. C. Efrata, W. E. D. Radianto, M. A. E. Marlina, and S. K. Dewi, "Innovation processes in new product development: Models for creative industry in Indonesia," J. Apl. Manaj., vol. 18, no. 3, pp. 486–492, Sep. 2020, doi: 10.21776/ub.jam.2020.018.03.08.

[21] H. Ren and Y. Zhao, "Technology opportunity discovery based on constructing, evaluating, and searching knowledge networks," Technovation, vol. 101, p. 102196, Mar. 2021, doi: 10.1016/j.technovation.2020.102196.

[22] J. Mämmelä, E. Mustonen, J. Härkönen, J. Pakkanen, and T. Juuti, "Productization as a link to combining product portfolio management and product family development," Procedia CIRP, vol. 109, pp. 25–30, 2022, doi: 10.1016/j.procir.2022.05.209.

[23] Q. Zhou and K. F. Yuen, "Analyzing the effect of government subsidy on the development of the remanufacturing industry," Int. J. Environ. Res. Public Health, vol. 17, no. 10, 2020, doi: 10.3390/ijerph17103550.

[24] S. Coldrick, P. Longhurst, P. Ivey, and J. Hannis, "An R&D options selection model for investment decisions," Technovation, vol. 25, no. 3, pp. 185–193, Mar. 2005, doi: 10.1016/S0166-4972(03)00099-3.

[25] J. Oehmen, A. Guenther, J. W. Herrmann, J. Schulte, and P. Willumsen, "Risk management in product development: Risk identification, assessment, and mitigation - a literature review," Proc. Des. Soc. Des. Conf., vol. 1, no. vii, pp. 657–666, 2020, doi: 10.1017/dsd.2020.27.

[26] J. Schulte, C. Villamil, and S. I. Hallstedt, "Strategic sustainability risk management in product development companies: Key aspects and conceptual approach," Sustain., vol. 12, no. 24, pp. 1–20, 2020, doi: 10.3390/su122410531.

[27] T. L. Saaty and L. G. Vargas, Models, methods, concepts & applications of the analytic hierarchy process. New York, USA.: Springer Science+Business Media, LLC, 2001.

[28] H. Wei, D.-B. Luh, X. Li, and H.-X. Yan, "AHP-based design of a finger training device for stroke," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 10, pp. 481–488, 2023, doi: 10.14569/ijacsa.2023.0141051.

[29] E. P. Massami and B. M. Myamba, "Application of vague analytical hierarchy process to prioritize the challenges facing public transportation in Dar Es Salaam city-Tanzania," Int. J. Adv. Res. Artif. Intell., vol. 5, no. 3, pp. 46–53, 2016.

[30] Sugiyono, Qualitative research methods (Metode penelitian kualitatif), 3rd ed. Bandung: CV Alfabeta, 2017.

[31] Marsono, Use of the analytical hierarchy process (AHP) method in research (Penggunaan metode analytical hierarchy process (AHP) dalam penelitian), 1st ed. Bogor Indonesia: Penerbit In Media, 2019.

[32] A. Rachman, A. Octavian, A. Irdham, I. N. Putra, Yusuf Ali, and A. K. Susilo, "Revolution in military affairs (RMA) by Indonesian armed forces towards competitive advantage," Decis. Sci. Lett., vol. 12, no. 2, pp. 413–430, 2023, doi: 10.5267/j.dsl.2022.12.002.

[33] R. D. Astanti, S. E. Mbolla, and T. J. Ai, "Raw material supplier selection in a glove manufacturing: Application of AHP and fuzzy AHP," Decis. Sci. Lett., vol. 9, no. 3, pp. 291–312, 2020, doi: 10.5267/j.dsl.2020.5.005.

[34] L.-M. Colaric-Jakše and M. Ambrož, "Actor-network theory and stakeholder collaboration : The case of Slovenia," Mediterr. J. Soc. Sci., vol. 6, no. 3, pp. 231–239, 2015, doi: 10.5901/mjss.2015.v6n3s2p231.

[35] R. Cherif, F. Hasanov, and G. Xie, "The making of East Asia's electronics champions," Rev. Econ. Mund., vol. 59, pp. 93–138, 2021.

[36] M. D. Stegall et al., "The importance of drug safety and tolerability in

the development of new immunosuppressive therapy for transplant recipients: The Transplant Therapeutics Consortium's position statement," Am. J. Transplant., vol. 19, no. 3, pp. 625–632, 2019, doi: DOI:https://doi.org/10.1111/ajt.15214.

[37] H. Liu, "The tripartite evolutionary game of green agro-product supply in an agricultural industrialization consortium," Sustainability, vol. 14, no. 11582, pp. 1–19, 2022.

APPENDIX 1. AHP QUESTIONNAIRE

A	The pairwise comparison matrix of criteria																		
1	Opportunity	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Budget
2	Opportunity	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Benefits
3	Opportunity	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Risk
4	Budget	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Benefits
5	Budget	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Risk
6	Benefits	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Risk
1: equal importance, 3: moderate importance, 5: high importance, 7: very high importance, 9: extreme importance, 2, 4, 6, 8: intermediate values																			

B	The pairwise comparison matrix of alternatives: opportunity basis																		
1	BRIN	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Consortium
2	BRIN	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	International cooperation
3	Consortium	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	International cooperation
1: equal importance, 3: moderate importance, 5: high importance, 7: very high importance 9: extreme importance, 2, 4, 6, 8: intermediate values																			

C	The pairwise comparison matrix of alternatives: budget basis																		
1	BRIN	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Consortium
2	BRIN	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	International cooperation
3	Consortium	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	International cooperation
1: equal importance, 3: moderate importance, 5: high importance, 7: very high importance 9: extreme importance, 2, 4, 6, 8: intermediate values																			

D	The pairwise comparison matrix of alternatives: benefits basis																		
1	BRIN	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Consortium
2	BRIN	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	International cooperation
3	Consortium	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	International cooperation
1: equal importance, 3: moderate importance, 5: high importance, 7: very high importance 9: extreme importance, 2, 4, 6, 8: intermediate values																			

E	The pairwise comparison matrix of alternatives: risk basis																		
1	BRIN	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Consortium
2	BRIN	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	International cooperation
3	Consortium	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	International cooperation
1: equal importance, 3: moderate importance, 5: high importance, 7: very high importance 9: extreme importance, 2, 4, 6, 8: intermediate values																			

APPENDIX 2. RESULTS OF PAIRWISE COMPARISON MATRIX BETWEEN CRITERIA (11 RESPONDENTS)

R1	Op	Bu	Be	Ri
Op	1,00	3,00	0,33	5,00
Bu	0,33	1,00	0,20	3,00
Be	3,00	5,00	1,00	7,00
Ri	0,20	0,33	0,14	1,00

R2	Op	Bu	Be	Ri
Op	1,00	1,00	0,20	5,00
Bu	1,00	1,00	0,33	1,00
Be	5,00	3,00	1,00	3,00
Ri	0,20	1,00	0,33	1,00

R3	Op	Bu	Be	Ri
Op	1,00	3,00	0,17	6,00
Bu	0,33	1,00	0,33	3,00
Be	6,00	3,00	1,00	9,00
Ri	0,17	0,33	0,11	1,00

R4	Op	Bu	Be	Ri
Op	1,00	0,33	0,20	5,00
Bu	3,00	1,00	0,33	3,00
Be	5,00	3,00	1,00	5,00
Ri	0,20	0,33	0,20	1,00

R5	Op	Bu	Be	Ri
Op	1,00	0,14	6,00	7,00
Bu	7,00	1,00	0,17	5,00
Be	0,17	6,00	1,00	7,00
Ri	0,14	0,20	0,14	1,00

R6	Op	Bu	Be	Ri
Op	1,00	1,00	1,00	3,00
Bu	1,00	1,00	1,00	5,00
Be	1,00	1,00	1,00	5,00
Ri	0,33	0,20	0,20	1,00

R7	Op	Bu	Be	Ri
Op	1,00	0,14	1,00	1,00
Bu	7,00	1,00	1,00	7,00
Be	1,00	1,00	1,00	6,00
Ri	1,00	0,14	0,17	1,00

R8	Op	Bu	Be	Ri
Op	1,00	7,00	0,14	5,00
Bu	0,14	1,00	0,14	0,20
Be	7,00	7,00	1,00	7,00
Ri	0,20	5,00	0,14	1,00

R9	Op	Bu	Be	Ri
Op	1,00	5,00	1,00	5,00
Bu	0,20	1,00	0,14	0,33
Be	1,00	7,00	1,00	5,00
Ri	0,20	3,00	0,20	1,00

R10	Op	Bu	Be	Ri
Op	1,00	0,11	0,20	1,00
Bu	9,00	1,00	0,11	1,00
Be	5,00	9,00	1,00	5,00
Ri	1,00	1,00	0,20	1,00

R11	Op	Bu	Be	Ri
Op	1,00	7,00	7,00	7,00
Bu	0,14	1,00	1,00	1,00
Be	0,14	1,00	1,00	1,00
Ri	0,14	1,00	1,00	1,00

APPENDIX 3. RESULTS OF PAIRWISE COMPARISON MATRIX BETWEEN ALTERNATIVES CONCERNING THE OPPORTUNITY CRITERION (11 RESPONDENTS)

R1	BR	CO	IC
BR	1,00	3,00	5,00
CO	0,33	1,00	3,00
IC	0,20	0,33	1,00

R2	BR	CO	IC
BR	1,00	0,33	3,00
CO	3,00	1,00	3,00
IC	0,33	0,33	1,00

R3	BR	CO	IC
BR	1,00	0,17	0,11
CO	6,00	1,00	0,17
IC	9,00	6,00	1,00

R4	BR	CO	IC
BR	1,00	0,20	5,00
CO	5,00	1,00	7,00
IC	0,20	0,14	1,00

R5	BR	CO	IC
BR	1,00	0,11	5,00
CO	9,00	1,00	9,00
IC	0,20	0,11	1,00

R6	BR	CO	IC
BR	1,00	0,20	0,50
CO	5,00	1,00	3,00
IC	2,00	0,33	1,00

R7	BR	CO	IC
BR	1,00	0,14	0,14
CO	7,00	1,00	1,00
IC	7,00	1,00	1,00

R8	BR	CO	IC
BR	1,00	0,14	0,20
CO	7,00	1,00	7,00
IC	5,00	0,14	1,00

R9	BR	CO	IC
BR	1,00	0,11	0,11
CO	9,00	1,00	3,00
IC	9,00	0,33	1,00

R10	BR	CO	IC
BR	1,00	1,00	5,00
CO	1,00	1,00	9,00
IC	0,20	0,11	1,00

R11	BR	CO	IC
BR	1,00	0,14	0,14
CO	7,00	1,00	0,14
IC	7,00	7,00	1,00

APPENDIX 4. RESULTS OF PAIRWISE COMPARISON MATRIX BETWEEN ALTERNATIVES CONCERNING THE BUDGET CRITERION (11 RESPONDENTS)

R1	BR	CO	IC
BR	1,00	0,33	0,20
CO	3,00	1,00	0,33
IC	5,00	3,00	1,00

R2	BR	CO	IC
BR	1,00	0,33	3,00
CO	3,00	1,00	3,00
IC	0,33	0,33	1,00

R3	BR	CO	IC
BR	1,00	0,17	3,00
CO	6,00	1,00	9,00
IC	0,33	0,11	1,00

R4	BR	CO	IC
BR	1,00	0,33	4,00
CO	3,00	1,00	6,00
IC	0,25	0,17	1,00

R5	BR	CO	IC
BR	1,00	0,11	0,14
CO	9,00	1,00	0,11
IC	7,00	9,00	1,00

R6	BR	CO	IC
BR	1,00	0,20	0,50
CO	5,00	1,00	3,00
IC	2,00	0,33	1,00

R7	BR	CO	IC
BR	1,00	0,14	0,14
CO	7,00	1,00	1,00
IC	7,00	1,00	1,00

R8	BR	CO	IC
BR	1,00	0,20	0,20
CO	5,00	1,00	5,00
IC	5,00	0,20	1,00

R9	BR	CO	IC
BR	1,00	0,20	0,14
CO	5,00	1,00	0,20
IC	7,00	5,00	1,00

R10	BR	CO	IC
BR	1,00	9,00	1,00
CO	0,11	1,00	5,00
IC	1,00	0,20	1,00

R11	BR	CO	IC
BR	1,00	0,33	0,14
CO	3,00	1,00	0,14
IC	7,00	7,00	1,00

APPENDIX 5. RESULTS OF PAIRWISE COMPARISON MATRIX BETWEEN ALTERNATIVES CONCERNING THE BENEFITS CRITERION (11 RESPONDENTS)

R1	BR	CO	IC
BR	1,00	0,33	3,00
CO	3,00	1,00	5,00
IC	0,33	0,20	1,00

R2	BR	CO	IC
BR	1,00	0,33	3,00
CO	3,00	1,00	3,00
IC	0,33	0,33	1,00

R3	BR	CO	IC
BR	1,00	0,11	0,17
CO	9,00	1,00	3,00
IC	6,00	0,33	1,00

R4	BR	CO	IC
BR	1,00	0,33	5,00
CO	3,00	1,00	7,00
IC	0,20	0,14	1,00

R5	BR	CO	IC
BR	1,00	0,11	0,11
CO	9,00	1,00	0,11
IC	9,00	9,00	1,00

R6	BR	CO	IC
BR	1,00	0,20	3,00
CO	5,00	1,00	5,00
IC	0,33	0,20	1,00

R7	BR	CO	IC
BR	1,00	0,14	0,14
CO	7,00	1,00	1,00
IC	7,00	1,00	1,00

R8	BR	CO	IC
BR	1,00	0,14	0,33
CO	7,00	1,00	5,00
IC	3,00	0,20	1,00

R9	BR	CO	IC
BR	1,00	0,11	0,14
CO	9,00	1,00	3,00
IC	7,00	0,33	1,00

R10	BR	CO	IC
BR	1,00	1,00	5,00
CO	1,00	1,00	5,00
IC	0,20	0,20	1,00

R11	BR	CO	IC
BR	1,00	0,33	0,14
CO	3,00	1,00	0,14
IC	7,00	7,00	1,00

APPENDIX 6. RESULTS OF PAIRWISE COMPARISON MATRIX BETWEEN ALTERNATIVES CONCERNING THE RISK CRITERION (11 RESPONDENTS)

R1	BR	CO	IC
BR	1,00	3,00	5,00
CO	0,33	1,00	3,00
IC	0,20	0,33	1,00

R2	BR	CO	IC
BR	1,00	0,33	3,00
CO	3,00	1,00	3,00
IC	0,33	0,33	1,00

R3	BR	CO	IC
BR	1,00	3,00	0,17
CO	0,33	1,00	0,11
IC	6,00	9,00	1,00

R4	BR	CO	IC
BR	1,00	0,17	7,00
CO	6,00	1,00	8,00
IC	0,14	0,13	1,00

R5	BR	CO	IC
BR	1,00	7,00	0,11
CO	0,14	1,00	0,11
IC	9,00	9,00	1,00

R6	BR	CO	IC
BR	1,00	0,33	0,20
CO	3,00	1,00	0,50
IC	5,00	2,00	1,00

R7	BR	CO	IC
BR	1,00	0,14	0,14
CO	7,00	1,00	1,00
IC	7,00	1,00	1,00

R8	BR	CO	IC
BR	1,00	0,20	0,33
CO	5,00	1,00	5,00
IC	3,00	0,20	1,00

R9	BR	CO	IC
BR	1,00	0,33	0,11
CO	3,00	1,00	0,11
IC	9,00	9,00	1,00

R10	BR	CO	IC
BR	1,00	5,00	1,00
CO	0,20	1,00	1,00
IC	1,00	1,00	1,00

R11	BR	CO	IC
BR	1,00	0,33	0,14
CO	3,00	1,00	0,14
IC	7,00	7,00	1,00

APPENDIX 7. GEOMETRIC MEAN

Geometric mean of criteria

	Op	Bu	Be	Ri
Op	1,00	1,05	0,58	3,85
Bu	0,95	1,00	0,31	1,69
Be	1,71	3,20	1,00	4,85
Ri	0,26	0,59	0,21	1,00

Geometric mean of alternative based on opportunity criteria

	BR	CO	IC
BR	1,00	0,25	0,76
CO	4,00	1,00	2,25
IC	1,32	0,44	1,00

Geometric mean of alternative based on budget criteria

	BR	CO	IC
BR	1,00	0,31	0,48
CO	3,25	1,00	1,26
IC	2,09	0,79	1,00

Geometric mean of alternative based on benefit criteria

	BR	CO	IC
BR	1,00	0,22	0,67
CO	4,48	1,00	1,98
IC	1,49	0,50	1,00

Geometric mean of alternative based on risk criteria

	BR	CO	IC
BR	1,00	0,70	0,48
CO	1,44	1,00	0,74
IC	2,09	1,36	1,00

Unleashing the Potential of Artificial Bee Colony Optimized RNN-Bi-LSTM for Autism Spectrum Disorder Diagnosis

Dr Suresh Babu Jugunta[1], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[2], Dr.K.Aanandha Saravanan[3],
Dr. Kanakam Siva Rama Prasad[4], Dr. S. Koteswari[5], Venubabu Rachapudi[6], Manikandan Rengarajan[7]
Professor, Dept of Computer Applications-School of Computing, Mohan Babu University, Tirupathi[1]
Faculty of Informatics and Computing, UniSZA University, Malaysia[2]
Associate Professor, VelTech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology[3]
Assoc. Prof. Department of Computer Science and Engineering (AIML),
Kallam Haranadha Reddy Institute of Technology[4]
Professor, Pragati Engineering College, Surampalem Kakinada[5]
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur, Andhra Pradesh, India-522302[6]
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India-6000627[7]

***Abstract*—The diagnosis of Autism Spectrum Disorder (ASD) is a crucial, drawn-out, and sometimes subjective procedure that calls for a high level of knowledge. Automation of this diagnostic procedure appears to be possible because to recent developments in machine learning techniques. This paper presents a unique method for improving the performance of a Recurrent Neural Network with a Bidirectional Long Short-Term Memory (RNN-BiLSTM) model for ASD diagnosis by utilizing the power of Artificial Bee Colony (ABC) optimization. Because Python software is used to carry out the implementation, accessibility and adaptability in clinical contexts are guaranteed. The suggested approach is thoroughly contrasted with current techniques, such as ABC optimization for feature extraction, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) models, and Transfer Learning, in order to highlight its effectiveness. The outcomes demonstrate the superiority of the RNN-BiLSTM over other methods, with much greater accuracy and precision. Combining RNN-BiLSTM with ABC optimization demonstrates not just cutting-edge accuracy but also excellent interpretability. By using this sophisticated model's capabilities, an outstanding diagnosis accuracy of 99.12% is attained, which is 2.77% higher than previous approaches. The model helps physicians comprehend the diagnosis process by highlighting important characteristics and trends that influence its conclusion. Additionally, it lessens the subjectivity and unpredictability involved in human diagnosis, which may result in quicker and more accurate diagnoses of ASD. The research emphasizes how well the Artificial Bee Colony optimized RNN-BiLSTM model diagnoses autism spectrum disorder. By integrating AI-driven diagnostic tools into clinical practice, this research improves early diagnosis and intervention for ASD.**

Keywords*—Autism spectrum disorder; artificial bee colony; recurrent neural network; bidirectional long short-term network; artificial intelligence

I. INTRODUCTION

ASD is a neurodevelopmental disorder that is widespread and complicated that has a big influence on how people behave, communicate, and connect with others [1]. It affects people of different ages, genders, and backgrounds and spans a spectrum of symptoms, from moderate to severe. Since ASD was first discovered in the early 20th century, our knowledge of it has grown significantly, resulting in better diagnostic methods and criteria. Making an ASD diagnosis is a difficult, interdisciplinary process [2]. To ascertain the existence and severity of the condition, a variety of behavioral, developmental, and medical traits are usually evaluated. Many times, the diagnosis procedure is a drawn-out and intricate process that greatly depends on the knowledge and skills of clinicians, psychologists, speech therapists, and other medical specialists [3]. A person's quality of life and social integration can be greatly improved by early intervention and customized assistance, both of which are made possible by an accurate and timely diagnosis.

The Diagnostic and Statistical Manual of Mental Disorders and the International Classification of Diseases provide the most current and generally recognized classification of ASD, while there have been previous updates to the diagnostic criteria. These criteria take into account social communication impairments, the existence of repetitive and limited behaviors, and the requirement to evaluate the severity of symptoms in order to make a formal diagnosis [4].

Standardized observation, organized clinical interviews, and gathering comprehensive developmental histories from parents or carers are all common components of the evaluation process [5]. There is some subjectivity in the diagnosis process since these approaches rely on the experience and judgment of specialists to evaluate the behavioral and developmental data. Artificial intelligence combined with contemporary technology has demonstrated significant

potential in aiding in the diagnosis of ASD in recent years. More objective and effective diagnostic tools have been made possible by advances in machine learning algorithms, data analytics, and improved imaging techniques.

These cutting-edge methods seek to decrease subjectivity, increase accuracy, and speed up the diagnosis procedure all of which will eventually help people with ASD and their families. This lays the groundwork for an in-depth examination of the several tools and techniques utilized in the diagnosis of ASD. It will examine conventional diagnostic techniques, the changing digital tool market, and the possible revolution in the sector that AI and ML may bring about. The continuous search for more accurate, trustworthy, and approachable ways to diagnose ASD is a testament to the unrelenting dedication of scholars, medical professionals, and activists to helping individuals impacted by this intricate and varied disorder [6].

Since ASD is a complicated neurodevelopmental disorder that affects people differently, diagnosing it may be difficult and frequently subjective [7]. The long-term results for those with ASD are greatly improved by prompt intervention and specialized assistance, which are made possible by an early and precise diagnosis of the illness. Integrating cutting-edge technologies, including -BC optimization and RNN with BiLSTM, has emerged as a viable way to improve the diagnostic process in response to the urgent demand for more accurate and dependable diagnostic tools [8]. Traditionally, the ability of clinicians, psychologists, and other medical experts to evaluate behavioral observations, developmental histories, and clinical evaluations has been crucial in the diagnosis of ASD. However, as this procedure is fundamentally subjective, there may be differences in the accuracy of the diagnosis and possible delays.

Improving diagnosis techniques is crucial because of the complexity, heterogeneity, and early intervention requirements of ASD. The application of AI and ML in healthcare has created new opportunities to enhance and automate the diagnostic procedure. Large-scale dataset analysis, pattern recognition, and prediction are all impressive skills of machine learning algorithms. Complex temporal patterns and relationships in ASD-related data, such as behavioral observations and patient history, may be captured by RNN-BiLSTM, a recurrent and deep learning model combination [9].

Simultaneously, ML models perform better when nature-inspired optimization methods like ABC optimization are included. ABC optimization finds the best answers by simulating honeybee foraging behavior. It adjusts the hyper parameters of the RNN-BiLSTM model, improving the diagnostic efficiency and accuracy of ASD. In order to solve the difficulties in diagnosing ASD, this research aims to maximize the synergy between cutting-edge AI methodology and optimization strategies [10]. The purpose of this work is to speed the diagnosis of ASD, decrease subjectivity, and improve diagnostic accuracy by presenting the idea of ABC optimized RNN-BiLSTM. These developments might lessen the load on medical staff and the larger healthcare system in addition to enhancing the quality of life for people with ASD and their families. It will examine the technique, findings, and consequences of this novel approach as we dig into the next parts, showing how it advances the continuing search for more accurate and dependable instruments in the field of autism diagnosis.

ASD is a multifaceted neurodevelopmental disorder that manifests as a wide range of symptoms, including as limited interests, repetitive behaviors, and issues with socialization and communication. Improving the long-term results for people with ASD requires early identification and intervention. But making an accurate diagnosis of ASD is a difficult process that frequently depends on the expert opinion of physicians and other professionals. It is crucial to investigate and utilize the potential of cutting-edge technology to improve the diagnosis process since subjectivity can result in variances in diagnostic accuracy. Artificial intelligence and machine learning have advanced significantly in the last several years, with a number of applications including healthcare. With the availability of more impartial, reliable, and consistent evaluation instruments, these developments hold the potential to completely transform the diagnosis of ASD. RNNs and BiLSTMs are two of the many AI approaches that have shown to be effective at processing sequential data. This makes them especially suitable for the study of behavioral and clinical data related to ASD.

Finding the cause of an ASD diagnosis is crucial for impacted individuals as well as their families, teachers, and the general public [11]. ASD is a multifaceted neurodevelopmental disorder that presents with a wide range of symptoms and problems, including trouble with social communication, repetitive behaviors, and narrow interests. Effective ASD diagnostic methods and procedures are desperately needed for a number of compelling reasons, including the need of timely and accurate diagnosis. First and foremost, early action is made possible by early diagnosis. Although ASD is a lifelong illness, people on the spectrum can greatly enhance their quality of life and long-term results by receiving the right therapies and interventions at an early age.

When ASD is discovered early on, it is easier to start behavioral and educational treatments, speech and occupational therapy, and social skill development. Early diagnosis also makes it possible for families to better comprehend and assist their loved ones who have ASD. It facilitates the development of a more accepting and inclusive environment for people with ASD by assisting parents and other carers in navigating the difficulties and special requirements related to the disease. Early diagnosis allows families to get in touch with advocacy and support groups, giving them access to important tools and advice [12].

In the field of education, prompt diagnosis is essential to deliver customized and relevant curricula. Individualized Education Plans (IEPs) and accommodations can be implemented by educators and schools to make sure that kids with ASD get the help they need to succeed in the classroom. Fostering a student's intellectual and social growth requires an understanding of their unique strengths and difficulties. Additionally, the identification of ASD is essential for

resource allocation and public health planning. Healthcare systems can anticipate the demands on resources and services needed to serve people with ASD when they have accurate prevalence data. Governments and healthcare professionals need to know the extent of the problem in order to deploy resources wisely, as the incidence of ASD has been rising. Finally, early diagnosis also helps to lessen the long-term potential strain on the social assistance and healthcare systems.

People with ASD are more likely to acquire the skills required for independent functioning when early treatments and assistance are provided, which lessens their long-term dependency on intensive support services. These strong arguments make it more important than ever to have trustworthy, easily available, and effective ASD diagnosis methods. In order to guarantee that people with ASD receive the support and opportunities they deserve, researchers, clinicians, and the larger healthcare community are always striving to improve and innovate diagnostic methodologies. By doing this, they are making significant progress towards a society that is more accepting and understanding of one another.

In order to fully realize the promise of ABC optimization, this research explores a novel strategy that blends the advantages of AI with optimization methods inspired by nature. The goal of the RNN-BiLSTM model's integration of ABC optimization is to improve the precision and effectiveness of ASD diagnosis. We want to reduce the subjectivity present in conventional diagnostic approaches by improving the diagnostic capabilities of the model through hyper parameter optimization. This research project not only satisfies the urgent demand for more impartial and trustworthy ASD diagnosis instruments, but it also serves as an example of how AI and optimization techniques may work together in the healthcare industry.. Combining RNN-BiLSTM with ABC optimization presents a viable way to diagnose ASD more quickly, which might help affected individuals, carers, and physicians better understand the illness and enable early intervention.

The key contributions of the article is:

- By fusing two potent machine learning techniques BiLSTM and ABC optimization the research presents a novel diagnostic strategy. By combining the best features of both approaches, this innovative approach develops a strong foundation for diagnosing ASD.
- The investigation demonstrates the adaptability of ABC optimization in optimizing hyper parameters and also its efficacy in selecting the most useful features for precise diagnosis of ASD by integrating it into the feature extraction process.
- By employing Python software to construct the suggested paradigm, the study demonstrates its clinical usefulness. This guarantees flexibility and accessibility in actual clinical situations, which makes it a useful tool for medical practitioners.
- The study's impressive 99.12% diagnosis accuracy demonstrates the potential of the suggested RNN-BiLSTM model that has been optimized using ABC. This degree of precision is a major advancement over current techniques, lowering the possibility of false positives and boosting the procedure' dependability.

The remaining content of this article is arranged as follows: A synopsis of relevant research is given in Section II. The problem statement is provided in Section III. The article's Section IV explains the architecture and methodology of the recommended approach. Section V discusses the results and the debate that followed. Section VI discusses the conclusion.

II. Related Works

The need for efficient and effective medical diagnostic systems in the context of ASD detection and treatment is paramount [13]. Healthcare professionals often spend considerable time documenting and processing extensive remarks related to patient behavioral assessments. Early identification of ASD is vital for ensuring individuals receive appropriate care and treatment, ultimately improving their quality of life. Machine learning models present a promising avenue to explore the feasibility of identifying essential features and accurately assessing the presence or absence of autism. In this study, the objective is to create a recommendation model that leverages multiple classifiers to enhance the precision of ASD prediction. This study conduct experiments with a range of machine learning algorithms to assess the model's performance. The results indicate that, when considering evaluation metrics such as accuracy, precision, recall, and F1-score, Decision Trees and Random Forests outperform other algorithms.

Through the analysis of children's abnormal social patterns, behavioral observation is crucial in the diagnosis of ASD [14]. Even now, a significant portion of this procedure still depends on clinical observations, questionnaire surveys, or retrospective video analysis, which drives up the demand for experts and drives up labor costs. This work suggests a standardized platform for applying computer-aided ASD diagnosis to human behavioral data collection, analysis, modelling, and interpretation. The suggested system could automatically assess children's various social interaction abilities utilizing the recorded audio-visual data through an organized evaluation process, and it could additionally provide the ultimate diagnostic recommendations. During the research at a Chinese hospital, data was gathered from both ASD-afflicted (72 individuals) and non-ASD (24 individuals) patients, totaling 95 participants. According to the clinical database, the newly created computer software designed to aid in the identification of ASD in children has achieved an 88% accuracy rate. 42% accurate it works well with children who are around two years old and is as good as experienced doctors in diagnosing ASD. This solution can be easily shared and used in areas with fewer medical resources.

ASD is a neurodevelopmental disorder that impacts the way the brain forms and involves struggles in perceiving and responding to stimuli [15]. The challenges encountered in understanding their senses may hinder their ability to act appropriately and impede their cognitive and educational

development. The goal of this study was to see how the nervous system of children with autism and children without autism responds to sounds, images, and just sounds using a method called EEG. In this research, they looked at 20 children with ASD and 20 children with TD (typical development) to see how their brains work differently. The way the brain behaves can be studied by analyzing the EEG signal using non-linear methods. In this research, RQA is used to understand the hidden patterns in EEG data that are not linear. The RQA measures were studied by making different changes to the parameters used in the RQA calculations. In this research, our focus was on the cosine distance metric as it exhibits proficiency in information retrieval, and we contrasted different distance metrics to ascertain the most suitable biomarker. It examined and talked about every combination of the RQA measure and the corresponding channel. To determine if someone has autism or not, it used the features generated by a technique called RQA. These features were then inputted into a special type of neural network called BiLSTM. It tested how accurately we could classify the channels for each combination. When it comes to distinguishing between ASD and TD, the combination of T3 and T5 channels, along with selecting a fixed number of nearby neighbors and utilizing cosine as the distance measurement method, is widely regarded as the most successful, achieving an accuracy rate of 91.86%.

Stoddard et al. [16] critically examined the internal consistency of the Aberrant Behavior Checklist (ABC), Irritability Subscale (ABC-I), and its connection with additional indicators of irritability in 758 psychiatrically hospitalized youth with autism spectrum disorder, given its frequent use in clinical outcome research. Research performed factor and factor analysis in both confirmatory and exploratory datasets to characterize the ABC-I's internal structure. Based on factor analysis, a general factor suggests that the ABC-I represents approximately a one-dimensional idea of irritation. Apart from irritability, tantrums, verbal outbursts, self-harm, and poor affect are also shown as subordinate components. Notably, independent of irritation, self-harm factors account for a significant percentage of variance. As such, their input into studies of treatment effects ought to be taken into account. In therapy trials addressing irritability in ASD, more research or revisions to the ABC-I may enhance convergent validity with transdiagnostic formulations of irritation and avoid confounding from self-harm.

Due to the limited therapy available for ASD, the adoption of alternative interventions, such as gluten-free and casein-free (GFCF) diets, is common [17]. Objectives were to ascertain the impact of a GFCF diet on behavioral issues in kids and teens with ASD diagnoses and any possible correlation with urine beta-casomorphin levels. For this crossover trial, thirty-seven participants were enrolled. Every patient had a regular diet for six months, which included casein and gluten, and then a GFCF diet for an additional six months. The intervention's sequence beginning with the GFCF diet or the regular diet was decided at random. Three time points were assessed for the patients: before the trial started, following a regular diet, and following a GFCF diet. At each time point, urine beta-casomorphin concentrations were measured and questionnaires about diet compliance, behavior, and autism were filled out. Following the GFCF diet, there were no discernible behavioral alterations and no correlation with urine beta-casomorphin contents. Urinary beta-casomorphin concentrations and behavioral signs of autism do not significantly change after six months on the GFCF diet. More research with a lengthy follow-up period, comparable to ours, including components of blinding and placebo are required to more accurately identify those who responded to GFCF diets.

The diagnosis of ASD in children relies on several parameters, including social skills, repetitive behaviors, speech, and nonverbal communication [18]. Repetitive behavior is a crucial indicator for physicians when determining drug dosages, particularly in cases where the child exhibits increased aggressiveness as a symptom of the disorder's progression. To address the need for continuous monitoring and to replace the somewhat subjective measurement of repetitive behavior using the Aberrant Behavior Checklist, the paper introduces an innovative solution through the utilization of the IP Webcam app for ASD recognition. The proposed method employs activity detection to recognize changes in the behavior of autistic children, specifically in response to medication overdosages. This hybrid framework incorporates training a deep CNN model, utilizing the Autismdata.Net dataset, to monitor ASD children in their natural environment. Furthermore, transfer learning is employed to mitigate overfitting issues associated with the relatively small Autismdata.Net dataset when assessing the severity of the child's condition. The ASD children's behavior is evaluated using the Autismdata.Net dataset and validated by examining the thermoregulation of autistic children in response to medication. The proposed method demonstrates significantly improved action recognition accuracy compared to traditional clinical analysis or therapist observations. Ultimately, this system offers valuable support to physicians in regulating drug dosages for children with ASD.

The literature review emphasizes how important it is to have reliable and efficient medical diagnostic procedures in order to identify and treat ASD. It highlights how crucial early detection of ASD is to better patient treatment and overall quality of life. An intriguing strategy to improve the prediction accuracy of ASD is the exploration of machine learning models. The use of several machine learning algorithms is covered in the review, with DT and RF emerging as the best options. It also emphasizes the value of computer-aided diagnosis, especially when examining behavioral data, and provides a common platform for evaluating children's social interaction skills. Remarkably, an investigation conducted in a Chinese hospital identified ASD in youngsters with 88% accuracy, even surpassing that of board-certified physicians. The article also explores how EEG data processing is used to study how children with ASD's neurological systems react to stimuli differently. This study uses neural networks and other non-linear techniques, such as RQA, to categorize people with a 91.86% accuracy rate. The review also addresses the Aberrant Behavior Checklist's internal coherence and applicability to the topic of irritability in ASD. Lastly, it discusses the use of casein- and gluten-free diets as substitute

therapies for ASD and emphasizes the need for more study in this field. Overall, the study emphasizes how important cutting-edge methods are for diagnosing and treating ASD, such as ML, EEG analysis, and nutritional therapies.

As part of the review of literature, a thorough analysis is carried out to evaluate the technologies and procedures that are currently being used in the field of diagnosing ASD, providing a background for the current research. An analysis of previous research reveals a range of methodologies, each with unique advantages and disadvantages. Notably, some approaches show excellent diagnostic accuracy, but their practical use in clinical settings is hampered by their lack of interpretability. On the other hand, other studies rely on assumptions that can be viewed as unjustified or fail to take into account the subtle differences that exist across various demographic groups. This review of the literature highlights the contributions made by earlier researchers while also pointing out the shortcomings and gaps that have been found and are being addressed by current research. This nuanced view emphasizes how important it is to build on past successes and make new contributions that are specifically designed to address weaknesses in order to further the development of ASD diagnosis techniques.

III. Problem Statement

From the above discussed literatures, it states the diagnosis of ASD presents a significant difficulty due to its deep subjectivity and intrinsic heterogeneity in evaluation, which can lead to premature treatments. The subjective knowledge of doctors plays a major role in current diagnostic techniques, which introduces significant swings in diagnostic accuracy. This research project introduces a novel method that makes use of ABC optimization to address the urgent problem of improving ASD diagnosis. It carefully adjusts an RNN-BiLSTM architecture's hyper parameters, solving the fundamental problem of the pressing need for an ASD diagnosis tool that is more streamlined, accurate, and objective. This innovation has the potential to improve early interventions by reducing subjectivity and expediting the diagnostic process. This will ultimately benefit the lives of individuals on the autism spectrum and their families, who are working towards a future that is more accepting and helpful [19].

IV. Proposed ABC-RNN-BiLSTM Framework

This study's technique focuses on improving the diagnosis of ASD by applying ABC optimization to an RNN-BiLSTM. The first step of the procedure is gathering and preparing a large dataset. It then applies the ABC optimization approach to adjust the RNN-BiLSTM model's hyper parameters. For smooth implementation and accessibility in clinical settings, Python software is used.

The suggested RNN-BiLSTM is then compared with other widely used techniques in a comparative study. Feature extraction makes advantage of ABC optimization. The study assesses the model's interpretability and diagnostic accuracy, emphasizing its capacity to mitigate the subjectivity and unpredictability that come with human diagnosis while illuminating significant traits and patterns impacting the diagnostic process. It is depicted in Fig. 1.

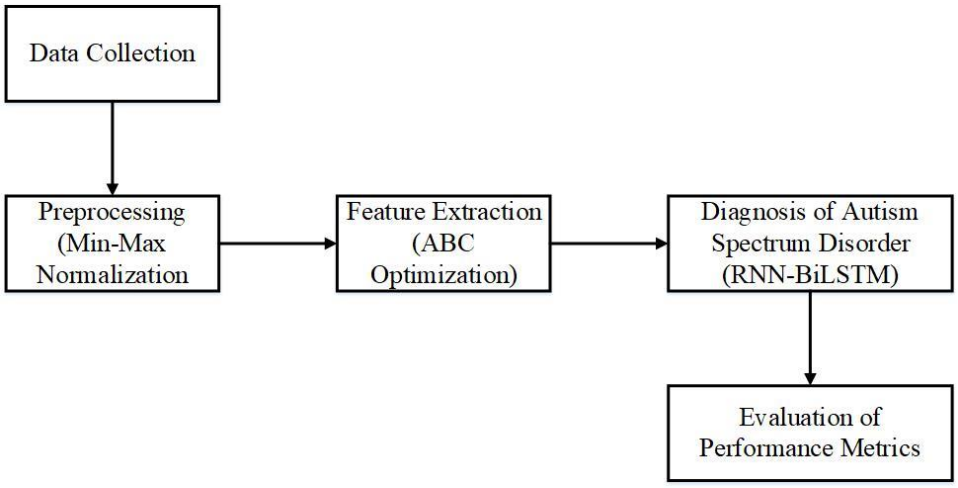


Fig. 1. Proposed ABC-RNN-BiLSTM framework.

A. Data Collection

The dataset used in this study was downloaded from Kaggle and consists of survey answers from people who filled out an application, labelled with whether or not the person was diagnosed with autism. This dataset provides insights into the traits and attributes that may be suggestive of ASD in individuals, making it an invaluable tool for research on ASD and its diagnosis. Using this dataset, the study intends to investigate how machine learning and AI methods, such as feature selection and classification algorithms, can enhance the precision and efficacy of ASD diagnosis, leading to improved early intervention and assistance for people on the autism spectrum [20]

B. Min-Max Normalization for Preprocessing

Preprocessing is a crucial first step that carefully plans data optimization in the goal of diagnosing autism. The application of Min-Max Normalization, a reliable approach that helps to harmonize the several data sources obtained from the patients, is key to this procedure. In order to guarantee that

each contributor has an equal weight inside the model, Min-Max Normalization expertly scales and standardizes these varied data inputs to a consistent range between 0 and 1. This harmonization reduces any unwanted effects brought on by data discrepancies, creating a setting that supports the RNN-BiLSTM increased accuracy. In addition to strengthening the resilience and preparedness of the early warning system, this standardized data bedrock serves as the cornerstone of more precise autism diagnosis. The actual data n is transformed linearly by Min-Max Normalization into the desired interval min_{new}, max_{new}.

$$n = min_{new} + (max_{new} - min_{new}) * (\frac{n - min_x}{max_x - min_x}) \quad (1)$$

The process has the advantage of appropriately preserving all links between the data bits. The data won't undergo any unfavorable alterations.

C. Feature Extraction using Artificial Bee Colony

The traditional ABC method, introduced by Karaboga in 2005, is a swarm-based artificial algorithm inspired by the foraging behavior of bee populations. During their foraging activities, honeybees are categorized into three distinct groups: employed bees, onlookers, and explorers. The employed bees are responsible for collecting nectar and exchanging information, the explorers are tasked with discovering new food sources, and the onlookers play a crucial role in determining the most efficient flight paths. Explorers seek out food resources based on their past experiences or sometimes by venturing randomly, and they may recruit fellow hive members to gather pollen while retaining valuable information about the food sources they encounter.

The conventional ABC approach, founded on simulating bee foraging behavior, serves as a swarm intelligence algorithm introduced by Karaboga in 2005 [21]. In this approach, honeybees are divided into three roles: employed bees, onlookers, and explorers, with each group contributing to the collective foraging success of the colony. Conversely, on the flip side, the onlookers' selection of the most efficient pollen-gathering path is influenced by their knowledge of food resources obtained initially and their decision to abandon those with lower nectar reserves.

In the conventional ABC technique, the number of explorers and the accuracy of information retained by employed honeybees regarding food sources have a significant impact on the tracking speed and overall effectiveness. According to various references in the conventional ABC method, if inaccurate data is incorporated, it can lead to suboptimal path optimization by onlookers, ultimately slowing down the tracking speed in later stages. Consequently, there is a need to implement novel strategies aimed at enhancing tracking performance [22].

As the number of bee's increases, the precision of tracking also improves. On the contrary, reducing the number of pollinators can lead to a shorter procedure duration, although it can pose challenges in pinpointing the optimal solution. This is because the method executed using a microcontroller has a search duration that directly correlates with the bee population size, resulting in a longer procedure time for a larger population. As a result, to address the challenge of attaining the global optimum within a limited number of iterations due to a reduced number of bees, an enhanced approach employing an ABC method with a minimal number of pollinators in conjunction with the RNN-Bi-LSTM technique is introduced and applied to forecast ASD.

The paper introduces an optimized RNN-Bi-LSTM hybrid network model, which is enhanced using the ABC algorithm to improve the reliability of forecasts. During the search phase, the algorithm aims to find a globally optimal solution to enhance the resolution. The equation being sought is as follows:

$$B_{uv}^{new} = B_{uv} + r_1 \times (B_{uv} - B_{neighbour,v}) + r_2(O_v - H_{uv}) \quad (2)$$

When presenting the global best solution, the ABC approach prioritizes honey sources that exhibit high adaptability while reducing the diversity within the bee colony. This can lead to the issue of early convergence, resulting in the emergence of localized optima. To address this, the enhanced ABC algorithm incorporates flexible parameter adjustments that help maintain the convergence rate and population characteristics. The search equation for this improved ABC algorithm is then derived:

$$B_{uv}^{new} = F_1 \times B_{uv} + F_1 \times r_1 \times \left(B_{uv} - B_{neighbour,v}\right) + F_2 \times r_2(O_v - H_{uv}) \quad (3)$$

$$F_1 = F_{max} - (2 - e^{\frac{iterate}{\max cycle} ln2})(F_{max} - F_{min}) \quad (4)$$

$$F_2 = F_{min} - (2 - e^{\frac{iterate}{\max cycle} ln2})(F_{max} - F_{min}) \quad (5)$$

$$Q_i = \frac{1/Fitness_i}{\sum_{j=1}^{n} 1/Fitness_j} \quad (6)$$

Due to the utilization of the reverse martingale feature selection method, the subsequent bees will place a higher emphasis on searching for nectar resources with limited adaptability during the initial phases of the process.

$$\propto = e^{\frac{iterate}{\max cycle} ln2} - 1 \quad (7)$$

Q_i equation and solutions B_i are optimized as follows:

$$Q_i = \begin{cases} \frac{\boldsymbol{Fitness_i}}{\sum_{j=1}^{n} \boldsymbol{fitness_j}} & , \boldsymbol{random} > \propto \\ \frac{\boldsymbol{1/Fitness_i}}{\sum_{j=1}^{n} \boldsymbol{1/Fitness_j}} & , \boldsymbol{random} < \propto \end{cases} \quad (8)$$

A colony of artificial bees are employed in iterative phases to represent dataset features as food sources in the adaption of the ABC algorithm for feature extraction in the diagnosis of ASD. These bees investigate feature subsets and assess their quality using a fitness function during the employed bees phase. Based on the findings of the employed bees, onlooker bees choose feature subsets, and if no progress is observed, scout bees investigate completely new subsets. Utilizing dynamic parameter adjustments, the ABC algorithm balances exploration and exploitation to optimize feature selection. Through effective feature selection and search, reduction of noise and redundancy, and improved ASD categorization, this strategy improves diagnosis accuracy.

D. RNN-BiLSTM for the Diagnosis of Autism Spectrum Disorder

There is a lot of potential in using an RNN-BiLSTM architecture to diagnose ASD. The diagnosis of Autism Spectrum Disorder is a complex task that necessitates the examination of large amounts of behavioral data, frequently spanning many time periods. RNN-BiLSTM's special benefit is its remarkable capacity to represent time relationships in this data. Complex time-dependent trends in behavioral observations are captured by the BiLSTM component by taking into account both past and future context in the sequences. This is especially important to consider when evaluating developmental milestones, social interactions, and repeated behaviors, all of which change with time. The ability to identify minute alterations and abnormalities that might be crucial markers of ASD is provided by the use of RNN-BiLSTM in the diagnostic procedure, which improves the precision and accuracy of the diagnosis.

The diagnostic procedure gains interpretability thanks to the RNN-BiLSTM model. It pinpoints important characteristics and trends that influence its diagnostic judgments, which may be extremely helpful for medical professionals in comprehending the elements influencing the diagnosis. This openness can help in the creation of more specialized and focused therapies for people with ASD. RNN-BiLSTM plays a critical role in enhancing ASD diagnosis by giving medical professionals a better understanding of the diagnostic procedure and the tools they need to make wise judgments. It is a major advancement in the direction of more dependable and effective diagnostic instruments that can hasten early intervention and improve the quality of life for people on the autism spectrum and their families.

An excellent tool in many domains, from time series analysis to natural language processing, are RNNs, a family of ANN that specialize in modelling sequential data. The fundamental function of RNNs is to process data having a time-based or sequentially structure, in which inputs are handled in connection to one another rather than separately from one another. A hidden state that changes over time and retains data from previous observations is a key component of an RNN's basic design. RNNs are particularly well-suited for tasks like speech recognition, language production, and sentiment analysis because of their dynamic memory mechanism, which enables them to identify and remember patterns within sequential data.

Traditional RNNs do have certain drawbacks, though, particularly when it comes to managing long-range dependencies. The vanishing gradient issue can make it difficult for deep RNNs to learn and retain information over long sequences. Variants with more intricate gating mechanisms, such as LSTM, were developed to meet this problem. These architectures improve RNNs' capacity to learn and retain important information over long sequences, which makes them a key component of contemporary machine learning for a variety of uses, such as speech recognition, machine translation, and even the diagnosis of conditions like ASD, where it is crucial to model the behavioral data's time frame evolution.

$$y_{in}=\partial(WE^{iny}x_{in}+WE^{hy}h_{im-1}) \quad (9)$$

$$rg_{in}=(WE^{inrg}y_{in}+WE^{hrg}h_{in-1}) \quad (10)$$

Here, the weight metrics WE_{RMM} are indicated as WE^{iny}, WE^{inrg},and WE^{hrg},

The logistic sigmoid function is termed as ∂: In the hidden unit, the candidate state is formulated in,

$$\check{h}_{in}=\tan(WE^{inh}\ y_{in}+WE^{hy}(h_{in-1}\otimes rg_{in})) \quad (11)$$

Here, the element wise multiplication is termed as,

$$\text{While } h_{in}=(1-y_{in})\otimes\hat{h}_{in}+y_{in}\otimes h_{in-1}h_{in} \quad (12)$$

$$=(\ (1-y_{in})\otimes\hat{h}_{in}+y_{in}\otimes\ h_{in-1}h_{in} \quad (13)$$

The reset gate essentially makes the device behave as though it is examining the power source first representation of a feedback sequence when it is closest to 0, enabling it to forget the previously established state.

Adaptable and dynamic, BiLSTM is a deep learning architecture that can handle jobs involving sequential data processing. It is a development of the classic LSTM model that improves on its ability to identify dependencies in sequential data. The primary novelty of BiLSTM is its directionality, which enables it to analyze sequences taking into account context from both the past and the future. Through the use of two distinct LSTM networks one for forwarding data and the other for backwards BiLSTM is able to fully comprehend the temporal correlations present in the data. This bidirectional analysis is particularly helpful in situations when a data point's interpretation heavily depends on observations from the past and the future. As a result, BiLSTM performs exceptionally well in a variety of fields, including as voice recognition, natural language processing, and time series analysis, where the ability to comprehend context is crucial for making precise predictions.

The vanishing gradient problem, one of the main issues with conventional RNNs, is lessened by BiLSTM. Standard RNN gradients can get very tiny in deep sequences, which hinders learning and makes it difficult for the model to capture long-range relationships. By ensuring that information moves more freely across the network, the gating methods of the BiLSTM enable it to represent lengthy sequences efficiently and avoid running into the vanishing gradient problem.

In applications like as machine translation, where the link between words or symbols may span the whole phrase, its ability to simulate long-range dependencies is very significant. Therefore, by addressing the difficulties of comprehending and interpreting sequential information, BiLSTM's improved capacity to handle sequential data with both short- and long-term dependencies makes it a fundamental building block in many contemporary deep learning applications, empowering advancements in diverse fields.

An effective and adaptable deep learning model, the combined RNN-BiLSTM architecture is well-suited for sequential data processing. The advantages of both classic

RNNs which are excellent at capturing temporal dependencies and BiLSTM which takes into account both past and future are combined in this hybrid technique. As a consequence, the model gains a thorough knowledge of sequential information, which helps it identify complex patterns and relationships in any type of data be it behavioral observations, time series data, or natural language text.

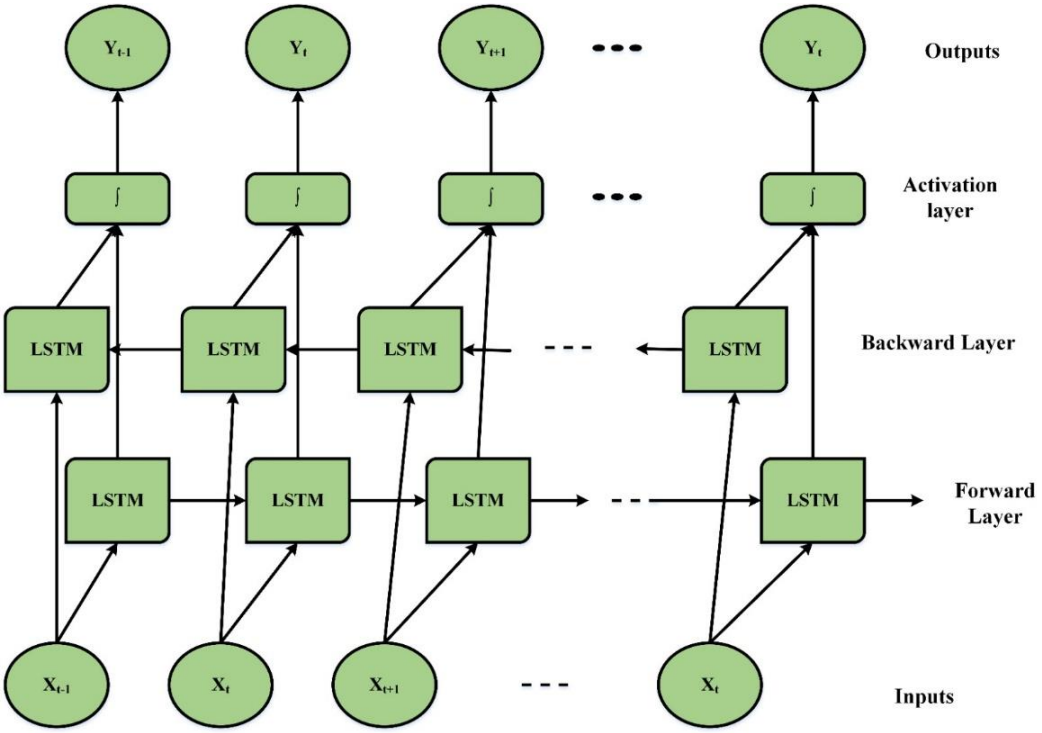


Fig. 2. Architecture of RNN-BiLSTM.

BiLSTM allows for a more sophisticated understanding of the context inside sequences, which is especially useful for applications like voice recognition, sentiment analysis, and, most importantly, the diagnosis of complex disorders like ASD. This hybrid architecture is a key component of state-of-the-art machine learning systems because of its capacity to extract both short- and long-range relationships from data. Fig. 2 shows the architecture of RNN-BiLSTM.

V. Results and Discussion

Through the use of ABC optimization, an RNN-BiLSTM is optimized in this study to improve the diagnosis of ASD. A thorough dataset is gathered and preprocessed at the start of the procedure. The RNN-BiLSTM model's hyper parameters are then adjusted using the ABC optimization approach. The use of Python software facilitates implementation and ensures accessibility in medical contexts. Next, a comparative study is carried out, whereby the suggested RNN-BiLSTM is contrasted with other widely used techniques. For feature extraction, ABC optimization is employed. The study assesses the model's diagnostic accuracy and interpretability, emphasizing its capacity to clarify significant traits and patterns impacting the diagnostic procedure while reducing the subjectivity and unpredictability inherent in human diagnosis.

Model accuracy, which shows the percentage of properly predicted instances among all the instances in a dataset, is a crucial performance statistic in machine learning and predictive modelling. It measures how well the model can predict the future and shows how closely the model's output matches the real results or the ground truth. Accuracy is a percentage that indicates how well the model performs overall in categorizing or forecasting results; higher accuracy values indicate a stronger ability to generate accurate predictions. Although accuracy offers a comprehensible and transparent indicator of a model's efficacy, it is crucial to take into account the particular problem context and possible class disparities, since a model's exceptionally high accuracy may not necessarily indicate that it can generalize well across various datasets or conditions. The model accuracy of the proposed method is depicted in Fig. 3.

Model loss, also known as the objective function or loss function, is a crucial parameter in deep learning and machine learning that measures how different the model's predictions are from the real target values. It calculates the prediction error, or loss, of the model and provides the foundation for fine-tuning its parameters as it is being trained. The main objective is to minimize the loss, a sign that the real values and the model's predictions are quite near. There are several different types of loss functions, such as cross-entropy for classification tasks and mean squared error for regression tasks. In order to guarantee that machine learning models provide precise and accurate predictions, monitoring and minimizing the loss function is essential during the training process. This will eventually improve the models' performance and predicted. It is depicted in Fig. 4.

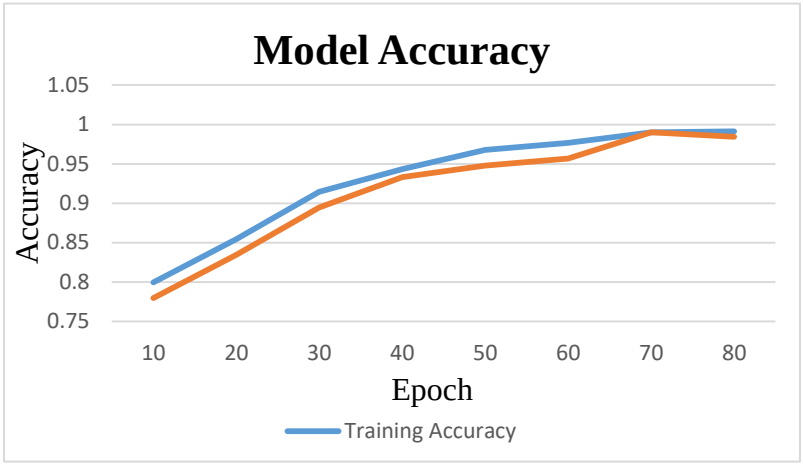


Fig. 3. Model accuracy.

Fig. 4. Model loss.

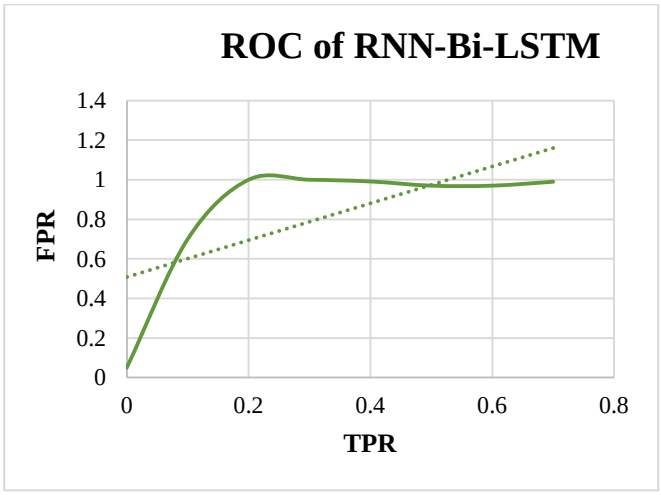


Fig. 5. ROC curve.

A thorough assessment has been conducted on the effectiveness of the suggested method, which diagnoses ASD by combining the RNN-Bi-LSTM with the ABC algorithm. The Receiver Operating Characteristic (ROC) curve is a useful tool for evaluating the diagnostic performance of classification models and is one of the primary metrics demonstrating the efficacy of the system. The ROC curve for the suggested system, which has attained an exceptional diagnostic accuracy of 99.12% in the context of ASD diagnosis, is presented in this analysis. Fig. 5 illustrates the graphical depiction of a classifier's capacity to discern between positive and negative situation is called a ROC curve. Across various categorization thresholds, it compares the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity). One widely used statistic to measure a model's discriminatory capacity is the area under the ROC curve (AUC). An AUC of 1 would indicate a flawless model, whereas an AUC of 0.5 would indicate a random estimate.

A. Fitness Assessment of the Proposed System

The suggested method, which combines RNN-Bi-LSTM with the ABC algorithm for diagnosing ASD, is being evaluated for its fitness using a wide range of criteria, including accuracy of diagnosis, relevance of feature selection, generalization to new cases, computational efficiency, robustness to noisy data, interpretability of results, comparison with other proven diagnostic techniques, practical application in clinical settings, optimization of model hyper parameters specify In light of the particular nuances and difficulties associated with diagnosing ASD, this evaluation seeks to ascertain the system's efficacy in accurately diagnosing the disorder, as well as its capacity to identify the most informative features, generalize findings to a range of patient profiles, and provide useful assistance to clinicians. The suggested ABC method's fitness assessment is graphically represented in Fig. 6.

The overall performance of the system model is assessed using accuracy. The core idea behind it is that every encounter can be accurately predicted. Eq. (14) is utilized to provide the accuracy.

$$Accuracy = \frac{T_{Pos} + T_{Neg}}{T_{Pos} + T_{Neg} + F_{Pos} + F_{Neg}} \quad (14)$$

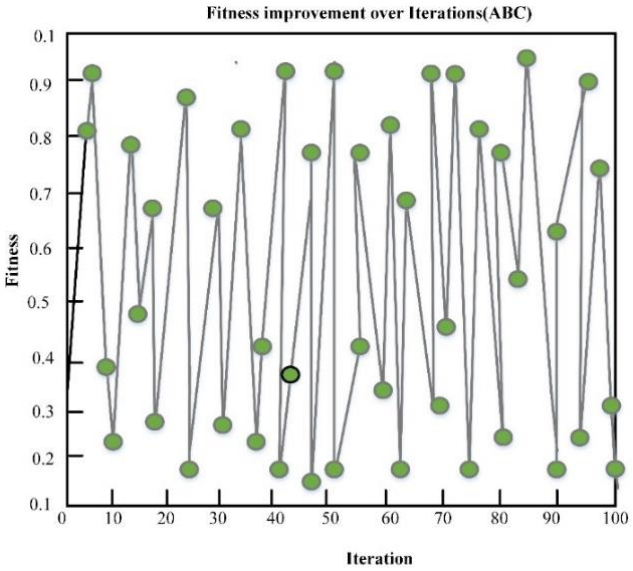


Fig. 6. Fitness assessment of ABC.

In addition to being accurate, precision also characterizes how similar two or more computations are to one another. The relationship between precision and accuracy demonstrates how frequently viewpoints can shift. It is brought up in Eq. (15).

$$P = \frac{T_{Pos}}{T_{Pos} + F_{Pos}} \tag{15}$$

Recall is the proportion of all relevant findings that were successfully sorted using the approaches. For these numbers, the appropriate positive is obtained by dividing the true positive by the falsely negative values. In Eq. (16), the phrase is mentioned.

$$R = \frac{T_{Pos}}{T_{Pos} + F_{Neg}} \tag{16}$$

Accuracy and recall are combined in the F1-Score calculation. Apply Eq. (17), which divides the recall by the accuracy to find the F1-Score.

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \tag{17}$$

A comparison of several techniques for diagnosing ASD is shown in Table I, with an emphasis on important performance indicators such as F1-Score, Accuracy, Precision, and Recall. Remarkably, the outcomes show that the Proposed RNN-BiLSTM approach performs better than the other methods, with an F1-Score of 98.11% and remarkable results of 99.12% accuracy, 99% precision, and 98.99% recall. This suggests that the RNN-BiLSTM model has excellent diagnostic performance and can predict ASD with a high degree of accuracy. Although the Transfer Learning method also performs exceptionally well, its accuracy and recall are somewhat worse than those of the RNN-BiLSTM model. However, although still producing excellent results, the conventional CNN and LSTM models are inferior to the suggested RNN-BiLSTM in every metric. The RNN-BiLSTM's remarkable performance highlights its promise as a robust diagnostic tool for ASD, lowering subjectivity and improving the precision and dependability of early treatments and support for people on the spectrum. These results imply that cutting-edge deep learning methods, like the RNN-BiLSTM model, have enormous potential for use in healthcare and greatly enhance the ability to diagnose difficult neurodevelopmental disorders like ASD. In Fig. 7, it is shown.

TABLE I. COMPARISON OF PERFORMANCE METRICS

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN	95	92.35	98.91	97.89
LSTM	96.77	97.34	96.36	97.12
Transfer Learning	97.57	98.52	97.56	95.56
Proposed RNN-BiLSTM	99.12	99	98.99	98.11

B. Discussion

The study's results provide strong evidence for the usefulness of the suggested ABC-optimized RNN-BiLSTM model for diagnosing ASD. The diagnostic accuracy of 99.12% is an amazing result that puts the model well ahead of current approaches, outperforming them by 2.77%. This result emphasizes the model's ability to produce extremely precise and dependable predictions, which is critical in lowering the subjectivity and ambiguity that are frequently connected to human diagnosis [4]. Additionally, the interpretability of the model gives medical professionals a better grasp of the diagnostic procedure, increasing openness and confidence in its predictions. A significant achievement in the area, the decrease in subjectivity and unpredictability may hasten the diagnosis and treatment of ASD in persons on the spectrum and their families, therefore enhancing their quality of life. These findings demonstrate the enormous promise of AI-driven diagnostic tools and represent a significant advancement in the accuracy and effectiveness of ASD diagnosis in clinical settings [5].

The obtained findings demonstrate the extraordinary accuracy and interpretability of the suggested ABC optimized RNN-BiLSTM model, and have important implications for the field of ASD diagnosis. In addition to demonstrating the model's effectiveness, its exceptionally high diagnosis accuracy of 99.12% raises the possibility of its use as a useful tool in clinical settings. By giving physicians insights into the elements impacting the diagnostic process, the model's interpretability improves its value and promotes a collaborative and educated approach to ASD diagnosis. But it's important to recognize some boundaries. The study primarily examines the model's diagnostic accuracy; however, in order to give a more thorough assessment, future research may explore other performance metrics, such as sensitivity and specificity. It is also important to evaluate the model's performance across a variety of demographic and population groupings in order to guarantee equal application and gauge its generalizability.

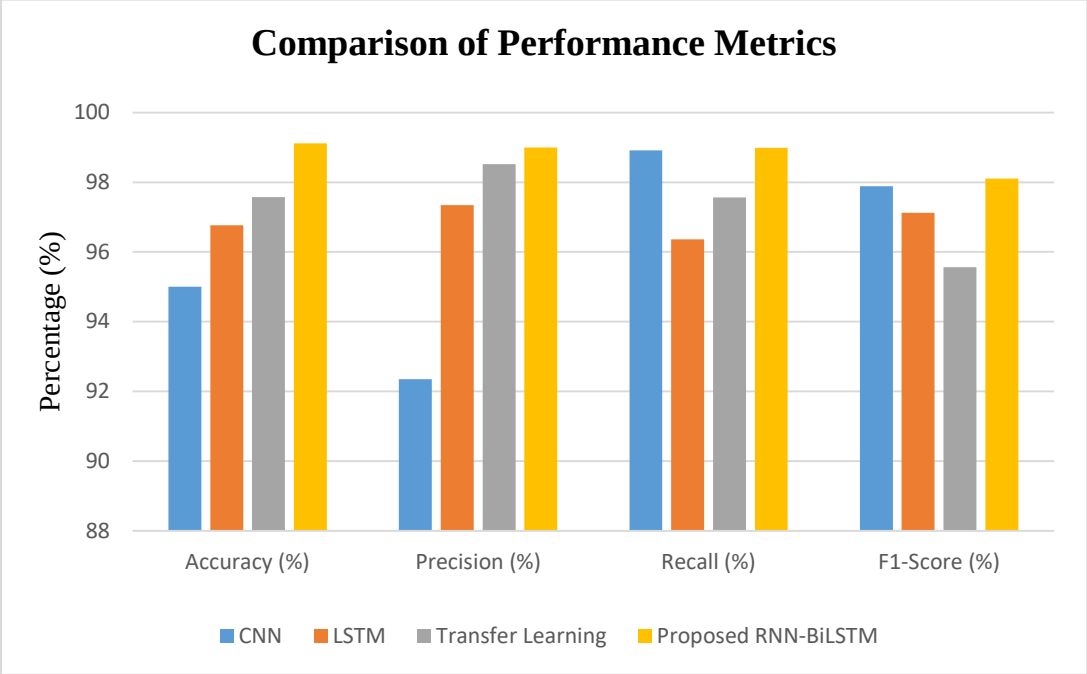


Fig. 7. Comparison of performance metrics.

Future research might improve the comprehension of the intricate elements of ASD by extending the model's applicability to multi-modal data, such as genetic and neuroimaging data. It is crucial to look at the privacy issues, ethical issues, and potential biases related to using AI-driven diagnostic tools in actual healthcare settings. Furthermore, investigating methods to smoothly incorporate the suggested model into current healthcare processes and systems will help make it more feasible to put it into practice. Even if the study makes a substantial contribution to the diagnosis of ASD, addressing these issues in subsequent research projects will increase the model's usefulness and clear the path for its effective incorporation into standard clinical procedures.

Enhancing the applicability of the suggested approach requires identifying possible generalization routes to more complicated scenarios. One approach is to carry out focused studies to evaluate the model's efficacy in a wide range of neurodevelopmental diseases other than ASD. Adding more examples of similar disorders to the dataset like Attention-Deficit/Hyperactivity Disorder (ADHD) or intellectual disabilities will help us better understand how adaptive the model is to more complex diagnostic problems. Furthermore, determining the model's generalizability requires examining its resilience across other age groups, and demographic groupings. The model's ability to generalize to complex and diverse clinical circumstances might be further improved by utilizing multi-modal data, such as genetic or neuroimaging information, and applying transfer learning techniques. These factors highlight the significance of methodically investigating and verifying the model's functionality in more complicated circumstances, offering a path for its incorporation into a more thorough framework for diagnosing neurodevelopmental disorders.

VI. Conclusion and Future Works

This study represents a major advancement in the diagnosis of ASD. A remarkable 99.12% diagnosis accuracy is achieved by using an ABC optimized RNN-BiLSTM architecture. This accomplishment highlights the potential of the suggested technique to lessen subjectivity and improve the accuracy of ASD diagnosis in addition to demonstrating its effectiveness. Clinicians can get important insights from the interpretability of the model, which enhances their comprehension of the diagnostic procedure. Moreover, this decrease in subjectivity speeds up early intervention, which is a vital component in enhancing the lives of people with autism spectrum disorders and their families. Furthermore, the elimination of subjectivity expedites the early intervention procedure, and the interpretability of the model aids healthcare practitioners in understanding the diagnostic process. Subsequent paths might entail implementing this model in therapeutic environments, carrying out extensive experiments, and broadening its relevance to encompass a range of neurodevelopmental conditions. Furthermore, investigating the possibilities of transfer learning and incorporating multi-modal data may improve diagnostic accuracy even more and increase the research's impact on the field of medical diagnosis.

References

[1] Y. Xu, Y. Wang, Z. Yu, Y. Li, Y. Liu, and Y. Li, “Autism Spectrum Disorder Diagnosis with Eeg Signals Using Time Series Maps of Brain Functional Connectivity and a Combined Cnn-Lstm Model.” Rochester, NY, Aug. 24, 2023. doi: 10.2139/ssrn.4542833.

[2] L. Sikich et al., “Intranasal Oxytocin in Children and Adolescents with Autism Spectrum Disorder,” New England Journal of Medicine, vol. 385, no. 16, pp. 1462–1473, Oct. 2021, doi: 10.1056/NEJMoa2103583.

[3] “AutiScan: Screening of Autism Spectrum Disorder Specific to Indian Region | IEEE Conference Publication | IEEE Xplore.” Accessed: Oct.

19, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9972038

[4] "Impact of antipsychotics in children and adolescents with autism spectrum disorder: a systematic review and meta-analysis | Health and Quality of Life Outcomes." Accessed: Oct. 19, 2023. [Online]. Available: https://link.springer.com/article/10.1186/s12955-021-01669-0

[5] A. W. Zimmerman et al., "Randomized controlled trial of sulforaphane and metabolite discovery in children with Autism Spectrum Disorder," Molecular Autism, vol. 12, no. 1, p. 38, May 2021, doi: 10.1186/s13229-021-00447-5.

[6] Y. Sugimoto et al., "Aripiprazole in the real-world treatment for irritability associated with autism spectrum disorder in children and adolescents in Japan: 52-week post-marketing surveillance," BMC Psychiatry, vol. 21, no. 1, p. 204, Apr. 2021, doi: 10.1186/s12888-021-03201-6.

[7] S. Xu et al., "Altered Functional Connectivity in Children With Low-Function Autism Spectrum Disorders," Frontiers in Neuroscience, vol. 13, 2019, Accessed: Oct. 19, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2019.00806

[8] J. Matta, D. Dobrino, D. Yeboah, S. Howard, Y. EL-Manzalawy, and T. Obafemi-Ajayi, "Connecting phenotype to genotype: PheWAS-inspired analysis of autism spectrum disorder," Frontiers in Human Neuroscience, vol. 16, 2022, Accessed: Oct. 19, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2022.960991

[9] "BCMJ_Vol65_No8_austism-algorithm.pdf." Accessed: Oct. 20, 2023. [Online]. Available: https://bcmj.org/sites/default/files/BCMJ_Vol65_No8_austism-algorithm.pdf

[10] L. Sikich et al., "Intranasal Oxytocin in Children and Adolescents with Autism Spectrum Disorder," New England Journal of Medicine, vol. 385, no. 16, pp. 1462–1473, Oct. 2021, doi: 10.1056/nejmoa2103583.

[11] C. Ferreira, S. C. Caetano, J. Perissinoto, and A. C. Tamanaha, "Repercussion of the implementation of the Picture Exchange Communication System - PECS in the overload index of mothers of children with Autism Spectrum Disorder," CoDAS, vol. 34, p. e20210109, Jan. 2022.

[12] A. Atyabi et al., "Stratification of Children with Autism Spectrum Disorder Through Fusion of Temporal Information in Eye-gaze Scan-Paths," ACM Trans. Knowl. Discov. Data, vol. 17, no. 2, p. 25:1-25:20, Feb. 2023, doi: 10.1145/3539226.

[13] A. V. Shinde and D. D. Patil, "A Multi-Classifier-Based Recommender System for Early Autism Spectrum Disorder Detection using Machine Learning," Healthcare Analytics, vol. 4, p. 100211, Dec. 2023, doi: 10.1016/j.health.2023.100211.

[14] M. Cheng et al., "Computer-Aided Autism Spectrum Disorder Diagnosis With Behavior Signal Processing," IEEE Transactions on Affective Computing, pp. 1–18, 2023, doi: 10.1109/TAFFC.2023.3238712.

[15] T. A. Manoharan and M. Radhakrishnan, "Region-Wise Brain Response Classification of ASD Children Using EEG and BiLSTM RNN," Clin EEG Neurosci, vol. 54, no. 5, pp. 461–471, Sep. 2023, doi: 10.1177/15500594211054990.

[16] J. Stoddard, J. Zik, C. A. Mazefsky, B. DeChant, and R. Gabriels, "The Internal Structure of the Aberrant Behavior Checklist Irritability Subscale: Implications for Studies of Irritability in Treatment-Seeking Youth With Autism Spectrum Disorders," Behavior Therapy, vol. 51, no. 2, pp. 310–319, Mar. 2020, doi: 10.1016/j.beth.2019.09.006.

[17] P. J. González-Domenech, F. Díaz Atienza, C. García Pablos, M. L. Fernández Soto, J. M. Martínez-Ortega, and L. Gutiérrez-Rojas, "Influence of a Combined Gluten-Free and Casein-Free Diet on Behavior Disorders in Children and Adolescents Diagnosed with Autism Spectrum Disorder: A 12-Month Follow-Up Clinical Trial," J Autism Dev Disord, vol. 50, no. 3, pp. 935–948, Mar. 2020, doi: 10.1007/s10803-019-04333-1.

[18] B. Prabha, M. Priya, N. R. Shanker, and E. Ganesh, "Aberrant behavior prediction and severity analysis for autistic child through deep transfer learning to avoid adverse drug effect," Biomedical Signal Processing and Control, vol. 70, p. 103038, Sep. 2021, doi: 10.1016/j.bspc.2021.103038.

[19] P. Chakraborty et al., "Gastrointestinal problems are associated with increased repetitive behaviors but not social communication difficulties in young children with autism spectrum disorders," Autism, vol. 25, no. 2, pp. 405–415, Feb. 2021, doi: 10.1177/1362361320959503.

[20] "Autism Spectrum Disorder Prediction | Kaggle." Accessed: Oct. 21, 2023. [Online]. Available: https://www.kaggle.com/code/desalegngeb/autism-spectrum-disorder-prediction/notebook

[21] J. Xia, Y. Wang, and Y. Li, "A Navigation Satellite Selection Method Based on Tabu Search Artificial Bee Colony Algorithm," in 2020 IEEE 3rd International Conference on Electronic Information and Communication Technology (ICEICT), Shenzhen, China: IEEE, Nov. 2020, pp. 421–425. doi: 10.1109/ICEICT51264.2020.9334301.

[22] M. Cinar and A. Kaygusuz, "Optimum Fuel Cost in Load Flow Analysis of Smart Grid by Using Artificial Bee Colony Algorithm," in 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey: IEEE, Sep. 2019, pp. 1–5. doi: 10.1109/IDAP.2019.8875893.

Exploring the Insights of Bat Algorithm-Driven XGB-RNN (BARXG) for Optimal Fetal Health Classification in Pregnancy Monitoring

Dr Suresh Babu Jugunta[1], Manikandan Rengarajan[2], Sridevi Gadde[3],
Prof. Ts. Dr. Yousef A.Baker El-Ebiary[4], Dr. Veera Ankalu. Vuyyuru[5], Namrata Verma[6], Dr. Farhat Embarak[7]
Professor, Dept of Computer Applications, School of Computing, Mohan Babu University, Tirupathi, A.P, India[1]
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India-6000627[2]
Assistant Professor, Department of Computer Science and Engineering, Raghu Engineering College, A.P, India [3]
Faculty of Informatics and Computing, UniSZA University, Malaysia[4]
Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, 522502, A.P, India[5]
Rungta College of Engineering and Technology Bhilai Chhattisgarh, India[6]
Department of Computer Science-Faculty of Computing and Information Technology, University of Ajdabiya, Libya[7]

***Abstract*—Pregnancy monitoring plays a pivotal role in ensuring the well-being of both the mother and the fetus. Accurate and timely classification of fetal health is essential for early intervention and appropriate medical care. This work presents a novel method for classifying fetal health optimally by combining the Bat Algorithm (BA) in an effective manner with a hybrid model that combines Recurrent Neural Networks (RNN) and Extreme Gradient Boosting (XGB). The Bat Algorithm, inspired by the echolocation behaviour of bats, is employed to optimize the hyperparameters of the XGB-RNN hybrid model. This enables the model to adapt dynamically to the complexities of fetal health data, enhancing its performance and predictive accuracy. The XGB-RNN hybrid model is designed to capitalize on the strengths of both algorithms. XGB provides superior feature selection and gradient boosting capabilities, while RNN excels in capturing temporal dependencies in the data. This approach effectively deals with the difficulties involved in classifying fetal health in the context of pregnancy monitoring by combining these approaches. Python is used to implement the proposed framework. To validate the performance of the proposed approach, extensive experiments were conducted on a comprehensive dataset comprising a wide range of physiological parameters related to fetal health. When it comes to fetal health, BAT Algorithm's XGB-RNN (BARXG) performs outstandingly, greater than other classifiers in terms of accuracy, sensitivity, and specificity. The proposed BARXG model has greater accuracy (98.2%) than existing techniques, which include SVM, Random Forest Classifier, LGBM, Voting Classifier, and EHG.**

Keywords*—BAT; fetal health; pregnancy monitoring; RNN; XGBoost

I. Introduction

Embryogenesis and maternity are essential components for human life and fertility. Whenever the fertilized egg, also called as a zygote, grows becomes a developing embryo, becomes a fetus, and finally culminates with the conception as a new human being, it is called pregnancy. Although becoming pregnant is an amazing experience, there are dangers and uncertainty involved. It is crucial to protect the mother's wellness and health in addition to the developing fetus. Regular fetal health monitoring is essential to prenatal treatment in order to identify and quickly fix any possible problems. This is a complicated and transformational process. Monitoring the development of the fetus throughout pregnancy is one of the hardest and most complex treatments. Although the average duration of this incredible journey is forty weeks, individual experiences may vary greatly [1]. The growing child of a person around the final stages of pregnancy is called a fetus. It is a crucial phase that comes after the embryonic stage and before childbirth during the entire human gestational process. In the fetus, the life form develops significantly. Usually, the fetus is just a few millimeters long at the start of the fetal stage, which occurs during the ninth week of development [2]. The fetus may grow to a size of 19 to 21 inches or greater by the conclusion of the trimester. The following are the phases of fetal growth. Weeks 9–12 of the first trimester, the fetus experiences tremendous expansion and growth. Important organs and tissues develop, and the fetus starts to take on characteristics of a little human. Weeks 13–27 of the second trimester, the fetus's body is growing as well as becomes more proportional. The embryo starts to move more deliberately and has the ability to grab items and sucks its thumb. Weeks 28 to Birth of the third trimester, a noticeable increase in size characterizes the last trimester.

This tissues and structures of the fetus develop more in order to get prepared for living beyond the mother's body [3]. The mother can clearly observe the fetus's movements, and it is capable of reacting to outside stimuli. A fetus is vulnerable to various issues throughout the course of pregnancy. Obstetrics carelessness can have devastating consequences, such as during childbirth fetal mortality, deaths from stillbirth, including over time infant neurological abnormalities. More than 1.3 million fetal fatalities happen throughout childbirth every year [4]. Birth asphyxia represents one causing the main causes of fetal death. Birth asphyxia, also known as hypoxia, is the result of a disruption in the blood supply via the placenta, which results in low oxygen levels in the fetus's

brain. Hypoxia-induced fetal distress can result in a range of anomalies during birthing that can be classified as either life-threatening or non-life-threatening. A newborn's brain is very susceptible to the effects of oxygen; hence a shortage of oxygen can have fatal consequences for the developing brain. Therefore, in order to identify fetal acidic conditions early on, we require an effective method that can track the fetal condition in real time and notify obstetricians when something odd happens so they may act quickly to save the fetus from irreversible harm. Birth asphyxia caused hypoxia causes permanent mental and physical disabilities such as spinal cord injury, deafness and visual impairment, speech difficulties, and autism. One typical outcomes diagnosis linked to fetal/perinatal brain damage is cerebral palsy (CP).

Most people agree that cerebral palsy (CP) is a disorder for neurological growth that causes dyskinesias and spastic quadriplegia or diplegia, and hemiplegia. In full-term newborns, the rate of CP ranges from 2.5 to 4.0 in 1000 births. However, this number rises to 16–22 per 1000 live births for children delivered preterm or individuals that are tiny for their gestational age (growth limited). In industrialized nations, the Maternal Mortality Ratio (MMR) is significantly lower than in impoverished nations. High MMR frequently results in issues such as pre-eclampsia, insufficient tracking of both the maternal and unborn child's health, and pregnancy-related diabetes. With the right medical attention, MMR can be decreased and avoided. Monitoring of the baby is a routine practice carried out in the third trimester. Fetal tracking involves assessing the unborn child's health [2]. The well-being of the mother has a direct impact on fetal development. Cardiotocography is used to continuously measure the well-being and development progress of the fetus in order to prevent such issues. The goal of the cardiotocography is to assess the maternal uterine contractions while simultaneously monitoring the fetus' heartbeat. This procedure could be carried out in the last trimester, after the fetus's development has fully synchronized with its heart beat. Because this technique is simple and inexpensive, it should only be used by qualified medical professionals to diagnose fetal condition early and lower fetal mortality. The results of the CTG will show the mother's uterine contractions alongside the unborn child's heart rate, acceleration, deceleration, among other intricate measurements. Mother and fetal well-being are closely related. Managing the well-being of mother as well as baby depends on lowering the number of fetal deaths and keeping an eye on the circumstances of fetal health [5]. A prenatal test used to track the heartbeat of the fetus and uterine contractions throughout both gestation and delivery is called CTG, or Electronic Fetal Monitoring (EFM).

These variables are monitored by two sensors, and the initial value, allowable variations, decelerations, and accelerations are used to classify the fetal health state. Medical professionals regularly look at these amounts and classify the fetus's health. Any numbers that deviate from a healthy state should raise suspicions about one's health. Healthcare workers review the data and assign an identifier to each characteristic. The CTG technique, which uses an electromagnetic field (EMF) equipment to track heart rate and uterine reductions throughout pregnancy, is used to get these data. Healthcare professionals physically categorize collected information and match it within a category in which the criteria are fulfilled; whenever the values fall outside of this range, an anxious condition is indicated. Through health state prediction, machine learning techniques can help physicians determine the fetal medical condition [6]. Doctors often use cardiotocography (CTG) for their clinical duties to track and evaluate the fetal status throughout gestation and delivery. CTG entails constant recording of both uterine contraction (UC) and fetal heart rate (FHR) signals. However, as fetal physiological changes are intricate and controlled through neurological mechanisms, there is typically a great deal of intra-observer and inter-observer discrepancy when utilizing standard criteria over visual interpretation of FHR signals. Obstetricians reduce diagnostic errors during labour by doing several subjective judgments. The key issue with the previously described procedure, nevertheless, is that it cannot be empirically realised; instead, obstetricians rely their conclusions only their own observations. As a result, the frequency of needless cesarean sections (CSs) brought on by subjectively mistake is rising, and this has made the pursuit on an additional accurate examination of the FHR signal its primary motivation.

The principal technique used most commonly in hospital routine tests for fetal status identification is the cardiotocogram (CTG). Prenatal surveillance of CTG primarily uses two physiological signals: fetus heartbeat and uterine contractions. The distress of the fetus affects FHR, resulting in anomalous high or decreased FHR occurrences. Initial pathogenic condition identification is accomplished with the help using such data. CTG data may be used to categorize the fetus's pathogenic status in relation to regular, which indicates its healthy state. A hypoxic fetus is extremely susceptible and might be temporarily impaired or even die after birth. Over half of all the deaths that occur can be attributed to insufficient therapy and misinterpretation of FHR [7]. Fetal health diagnosis is a challenging procedure that depends on a number of input elements. The identification of fetal healthy state is made based on the levels or range of values associated with these symptoms. Determining the precise amounts for the periods among the provided signs that influence the diagnosis's outcome can be challenging with occasions. Physicians frequently divide the entire pain phase value into smaller segments, examine each segmented segment, and determine the person's overall health status based on their analysis. These periods frequently convey uncertainty and might vary from patient to patient. Additionally, distinct patients may respond to similar illnesses in varying ways. Women who are getting ready to become mothers have begun looking for prenatal guidance and knowledge about risks while illness symptoms through online resources. Despite approximately 3.7 billion application downloads in 2017, there were over 325,000 fitness, health, and medical applications accessible; pregnancy-related apps make up a significant portion of this category [8].

Apps for smartphones and tablet computers can help expectant mothers obtain information, track the growth of their fetus, comprehend alterations to their own bodies, and get comfort when they have worries. A gadget like a video

camera, wellness tracker, Kegel "exerciser," fetal heart rate "listener," or another kind of device allowing participants to monitor and communicate their personal data might be linked to maternity applications. On the other hand, nothing exists regarding the way smartphone app-based health treatments affect mother's behaviours or perinatal well-being [9]. However, there is a lack of research on the effectiveness of apps, their content, how mothers use them, or the best methods to include them into normal prenatal education and care. Midwives along with other professionals often speak to pregnant women that download and utilize applications. Classification systems for fetal health aid in the early detection of anomalies, problems, or departures from the typical developing processes [10]. Healthcare personnel have the ability to swiftly implement suitable treatments since earlier diagnosis. Rapidly healthcare treatments might be critical towards avoiding or minimizing problems, and early diagnosis of fetal health abnormalities facilitates these therapies. Based on the severity of the problem, this may involve measures including suggesting surgery, giving medicine, or altering mom's lifestyle. Pregnant women might feel less stressed and anxious when they realize their unborn child is well and under constant observation. Improved outcomes and fewer needless healthcare procedures might arise from personalized care. More precise fetal health categorization enables improved delivery process preparation. This helps in deciding if early labour inducers or cesarean section is required, as well as ensuring the right doctors remain on hand for the birth. In healthcare applications, selecting features is a critical process that is handled using the Bat Algorithm. The research shows how this method may be used to improve model accuracy and comprehension by identifying the most pertinent characteristics in the Cardiotocography (CTG) dataset.

The following are the main contributions to the suggested work:

- In an initial processing measure, it employs class weighting to prevent overfitting when training the model.
- Sequencing and temporal connections in data are captured by RNNs. RNNs may simulate how fetal heart rate and uterine contraction patterns change over time within the framework of fetal health monitoring. This is necessary in order to identify any abnormalities or anomalies.
- A hybrid design fuses the RNN and XGBoost models together. This combination enables the model to take use of RNNs' capacity for capturing temporal dynamics and XGBoost's expertise in feature engineering.
- The output of RNN is fed into XGBoost for classification. RNN-XGBoost model had hyperparameters tuned using the BAT algorithm.
- In the end, the optimization process of the Bat Algorithm yields a collection of characteristics regarded most significant for the goal of classifying fetal health.

This article's remaining sections are organized as follows: In Section II, an overview of relevant studies is provided. Section III presents the problem description for the current system. The approach and architecture of the suggested BARXG model for Fetal health classification are explained in Section IV of the paper. Section V presents the findings from the investigation and the subsequent discussion. Conclusion and future application of the suggested paradigm are covered in Section VI.

II. Related Works

The research in [11] proposed a fetal health classification using T2-FNN method. The fetal medical diagnosis can be a challenging procedure which requires a variety of inputs elements. An assessment of fetal medical condition has been carried out via the numbers or varying numbers associated with those given signs. Their will likely be discussion among specialized physicians when determining the precise ranges that constitute gaps while identifying illnesses. Since a consequence, illness diagnosis frequently takes place in unreliable circumstances and occasionally results in unfavourable mistakes. Precisely a result, choices may be questionable due towards the ambiguous aspect of illnesses or insufficient patient information. The 21 intake criteria define the fetal medical condition. The estimation of these numbers included testing and observations. Three outcome diagnoses for good stages for fetal growth have been defined using potential amounts for these variables. Average, Suspected, as well as Abnormal include these. It had been possible to establish overall type-2 fuzzy neural networks (T2-FNN) method's architecture utilizing the quantity combined inputs and outcomes symptoms. Utilizing fetal records, the developed T2-FNN is evaluated. The structure of the system makes use of a variety of criteria. The design turns out that while the number of criteria increases, so does the efficiency of the system. Utilizing fetal records, the developed T2-FNN is evaluated. The structure of the system makes use of a variety of criteria. The design turns out that while the number of criteria increases, so does the efficiency of the system. Although T2FNNs become more sophisticated than regular neural networks, they may be costly to compute and more difficult to carry out, particularly for applications that operate in real time.

The research in [12] proposed a fetal health classification using machine learning techniques. Cardiotocography (CTG) depicts the fetus's condition while in labour within the uterus. Yet, based on the obstetrician's experience, evaluating the results might be a very biased procedure. Infant monitoring digitally collect data (such as baby heart rate, movements and accelerating). Many investigators have concentrated their efforts on CTG information in order to evaluate fetal health utilizing different AI algorithms. Utilizing fetal heart rate data, certain investigators utilized neural networks to forecast fetal health. The suggested approach used the Fetal Health Assessment information set, which consists of CTG files, along with five ensembles participants: Random Forest, AdaBoost, XGBoost, CatBoost, and LGBM. The voting classifier, sometimes referred to by the term Meta classifier, classifies the CTG information using the results obtained from RF, XGBoost, AdaBoost, CatBoost, and LGBM. To categorize

CTG data, a soft voting technique is implemented using the mean result from every ensemble classifier. Regarding situations when many ensembles learner work identically, a soft voting classifier may be useful. The deficiencies of each individual ensemble's learners might be compensated by combining their work. In order to improve the efficiency of the entire model, the soft voting classifier ultimately removes the flaw of one particular classifier. Ensembles approaches, including the majority of machine learning algorithms, were dependent on noisy data; particularly the existence of disturbance in fetal health surveillance information may have an influence overall the model's efficiency.

The study in [13] suggested a strategy for fetal health classification. It is usual practice to utilize uterine contractions (UC) activities to gauge when labour and delivery will begin. In order to monitor UC and discriminate between effective and unproductive contractions, electro hysterograms (EHGs) have lately been adopted. From this investigation, the researchers utilized a convolutional neural network also known as CNN to detect UC in EHG signals. In order to create a CNN model, an open-access database has been utilized. Utilizing by five times cross-validation, a model based on CNN was created then learned with DB1. The CNN framework created with DB1 was utilized with DB2, a separate clinical database, to assess its generalizability for identifying UCs. Employing the multiple channels of communication system as well, the EHG signals in DB2 have been collected via 20 pregnant women, as well as 308 parts have been retrieved. The number of trials might be increased by combining both databases that might be preferable to teach the CNN model. The research has shown how CNN would effectively distinguish UCs with EHG signals. This technique makes it possible to consistently and correctly identify UCs, offering a unique tool for keeping track of the status of the labour and the health of the mother and fetus. Uneven classes might exist in EHG datasets, including UCs occurring less frequently than non-UCs. Unbalanced data may generate unbalanced models while having an impact on effectiveness in reality. It might be challenging and exhausting to integrate the CNN approach within present clinical processes and medical records systems.

[14] suggested a fetal health monitoring approach. To limit development negotiation, lower mortality, and avert premature birth, considerable health care services have been devoted toward tracking risky pregnancies. Another crucial sign for prenatal health was recently identified as fetal movement. Surveys showed that undesirable delivery rates occurred in 25% of pregnancy with reduced fetal movement during the 3rd trimester. They provide a better iteration of the automatic FetMov identification they already recommended. FetMov means Processes identified as FetMov by an ultrasonographer. Activities not identified from the ultrasonographer as FetMovs but with FetMov-like characteristics are called artefacts (Artf). They consist of parental body motions and sensor shifts. Information from accelerometers have been pre-processed using separate component analysis while wavelet decomposition over the initial time. The categorization set of characteristics has been increased by one attribute to 31 factors. Various models have been assessed employing a ten-fold cross-validation approach with the aim evaluate the performance of the suggested parameters. Thirty-one characteristics were taken using acceleration information in order to recognize fetal movements. Various predictors had been used for distinguishing fetal from non-fetal moves according to these characteristics. The models' reliability has been investigated across various artefact levels within the categorizing information. Bagging classifier method produced the most effective results. Automatic identification systems could result in false positives or false negatives, which could cause worry in expectant parents and result in pointless treatments or undetected problems in professional settings. Datasets that are unbalanced may result from uterine contractions being comparatively uncommon occurrences as compared to non-contraction times. This disparity problem might not be sufficiently addressed by bagging, which could lead to skewed predictions.

The research in [15] proposed a Fetal health monitoring using IoT method. Digital health apps utilizing the Internet of Things provide helpful instruments enabling efficient and dispersed automated systems for diagnosis. In order to track mother's and baby messages over pregnancy at high risk, this research suggests developing a combined approach utilizing Internet of Things (IoT) sensors, extracted features from data analysis, along with a predictive evaluation assist method built around a single-dimensional CNN classifier. In addition to recording the heart rate of the fetus, a number of clinical indications connected with the mother are also tracked, including blood pressure, temperature, heartbeat rate, uterine tonus action, and oxygen consumption. A substantial volume of data is produced at various speeds along with diverse formats by various sources. Utilizing a fog computing layer, a critical diagnosis system is suggested, considering the acquisition of various features along with the computation of both linear and nonlinear measurements, intelligent analytics for health system is suggested. Lastly, taking into account six potential outcomes, a method of classification is suggested as a system of forecasting for the categorization of maternal, fetal, and simultaneous health status. The crisis system receives information produced by IoT devices and employs it to evaluate and figure out whether it detects either severe fetal or maternal discomfort. The healthcare team is notified right away if a critical situation emerges. Following this analysis stage, every feature is computed and transmitted to the suggested estimation system using a single-dimensional CNN in the cloud-based approach. Lastly, the healthcare professional is supplied with a categorization that validates the diagnosis of illness. Severe legal requirements have to be met by IoT devices utilized in the healthcare industry. It might take money and effort to fulfill these criteria.

When it comes to predicting specific fetal health problems, like with a late- difficulties, which might arise following regular monitoring happened, the current categorization approach may not be very reliable. Several variables, consisting as mother health, genetics, and surroundings, might affect the health of the fetus. Such factors may lead to fetal reaction inconsistency and make categorization more difficult. The categorization scheme is predicated on data gathered from

typical prenatal visits, although might not necessarily offer an all-encompassing picture of fetal health. Reliable evaluation of several crucial factors is difficult, including fetal activity along with placenta functioning. T-2FNNs are more complex than their Type 1 counterparts due to the additional dimension of uncertainty they handle. This complexity can make the model challenging to understand and implement, especially for healthcare professionals who may not be familiar with fuzzy logic or neural networks. In contrast to simpler models, Random Forest models might be more difficult to read, which can make it difficult to comprehend the rationale behind certain categorization decisions—a critical skill in medical contexts. AdaBoost are very complicated, it may also overfit. Despite CatBoost's economical architecture, it could need an extended period to train than other algorithms, which could be a drawback for healthcare applications that need to respond quickly. LGBM may not perform as well with small datasets, as it is optimized for large-scale data.

III. Problem Statement

There are a number of issues with the current Optimum Fetal Health Classification during Pregnancy Monitoring which might affect how accurate and useful it is. Compared with other classification models, T2FNNs can be harder to comprehend, which makes it harder to justify certain categorization choices. Readability is essential in medical settings to win over healthcare professionals' confidence along with approval. Several machine learning models, particularly over fitting ones, exhibit weak results with unknown data yet good performance on training information. Overfit methods can't adjust effectively to novel patients or circumstances in healthcare settings, which might result in inaccurate diagnoses. The methods used today frequently depend on predictive models along with past information, and this may not be able to appropriately forecast difficulties in the future or account for each person pregnancy variances. Aspects such as fetal position, mother bodily habits, and electrode location might affect the quantity of EHGs. Errors in categorization might result from noisy or unreliable information [16]. The deployment of bagging classifiers along with other algorithms based on machine learning in healthcare facilities may be limited due to their computing demands, which necessitate substantial expenditures for both training and real-time usage.

IV. Proposed Bat Algorithm-Driven XGB-RNN for Optimal Fetal Health Classification in Pregnancy Monitoring

Compiling information on pregnancy monitoring is the initial stage. The information collected might contain the fetus's patterns of motion, heartbeat, and additional indicators of wellness throughout duration. Before being analysed, the gathered data must be cleansed. In order to handle values that are absent, normalize the data, and perhaps identify pertinent features for the classification task, all of this must be done. The framework and hyperparameters for a machine learning model are optimized using the Bat Algorithm. It can assist in choosing among the most important characteristics, determining the ideal model variables, and enhancing the efficiency of the framework as a whole. Along with the methods for categorization is XGBoost. Most commonly, it's employed to offer a preliminary data categorization. Sequential data, such as data on fetal health over time, are analyzed using the RNN. Utilizing a combined method like layering or mixing, merge the outcomes of the RNN and XGBoost model. Classification is done after hyperparameter tuning by BAT algorithm. Then fetal health is classified using BARXG model. Fig. 1 shows the overall diagram of proposed BARXG model for Fetal health classification.

A. Data Collection

Cardiotocograms (CTGs) are an easy-to-use, reasonably priced method of evaluating fetal health that enables medical practitioners to implement preventative measures against mother and infant death. The gadget essentially functions by delivering ultrasonic pulses and interpreting the reaction, thereby providing information on a variety of topics including uterine contractions, fetal movements, and fetal heart rate (FHR). The dataset is collected from fetal health classification from the website Kaggle [17]. 2126 sets with features taken from cardiotocogram tests are included in this collection of data. Three experienced obstetricians divided the characteristics among three categories: Normal, Suspect, and Pathology.

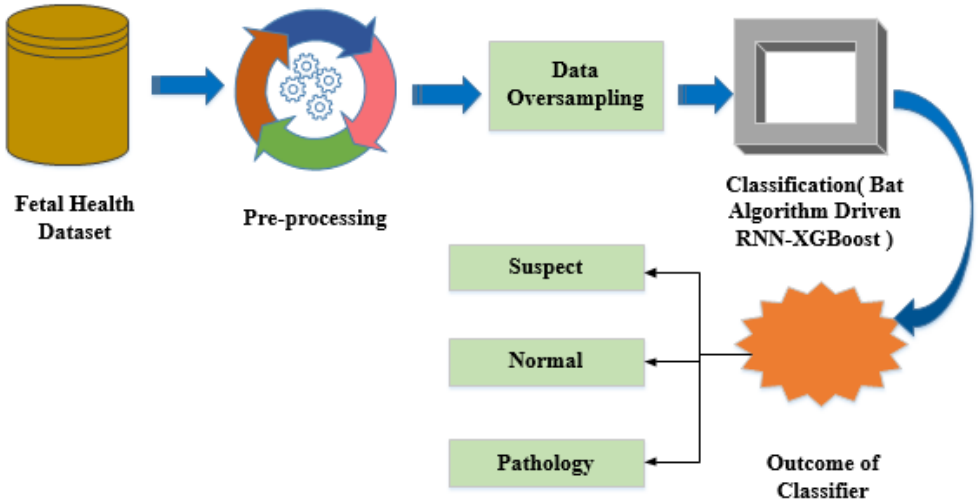


Fig. 1. Proposed BARXG model for fetal health classification.

B. Data Pre-processing

1) Data oversampling: To describe the relationship between variables during the model's development, the number of values provided in the dataset were standardized through a range from −1 to 1 during the preprocessing stage. The imbalanced dataset was handled using progressive class weighting following feature extraction. Adding some weight to every class to give the minority classes greater significance represents one among the easiest approaches to overcome this class imbalance and create a classifier which will learn similarly from all classes. It easily multiplies the entropy part of every class using the associated weight in a tree-based model, whereby the best split is defined using a certain metric, such lower entropy, to give the minority classes greater prominence.

In Eq. (1), a^1 represents a new value derived using the values, illustrates the normalizing and standardized procedure for feature extraction.

$$a^1 = \frac{a - min(a)}{max(a) - min(a)} \quad (1)$$

InEq. (2) provides the extended version that describes the polynomial expansion (PE) function utilized in feature extraction, where n denotes the degree of expansion and {b,c} are the independent variables within the dataset. Although the n degree in this study has been set at 2, the dataset in PE expands exponentially and horizontally with respect to n.

$$(b + c)^d = \sum\binom{d}{e} b^e c^{e-k} \quad (2)$$

However, computational expenses and horizontal expansion were kept to a minimal. Classes might be periodically evaluated by calculating their entropy function (f), as demonstrated by Eq. (3).

$$f = \sum_g q_g \log(q_g) \quad (3)$$

C. BAT Algorithm Driven RNN-XGBoost Model for Fetal Health Classification

This hybrid model combines the strength of XGBoost and Recurrent Neural Networks (RNN) with a feature selection method called Bat Algorithm (BA). The first step in the procedure is gathering data from fetal health monitoring. Numerous factors, such fetal heart rate and uterine contractions, are usually included in this data. The process of preparing data involves oversampling the dataset. The XGBoost and Recurrent Neural Network (RNN) models' hyperparameters are adjusted and refined by the BA. The goal of this optimization approach is to determine which hyperparameters will best fit the models and assist them capture intricate patterns in the fetal health data. Because RNNs are specifically designed to analyse sequential data, they are a good fit for time-series data sets such as fetal health monitoring. To increase prediction accuracy, XGBoost is used to the characteristics that were taken from the fetal health data. To arrive at a final forecast, the outputs of the RNN and XGBoost models are fused, or blended.

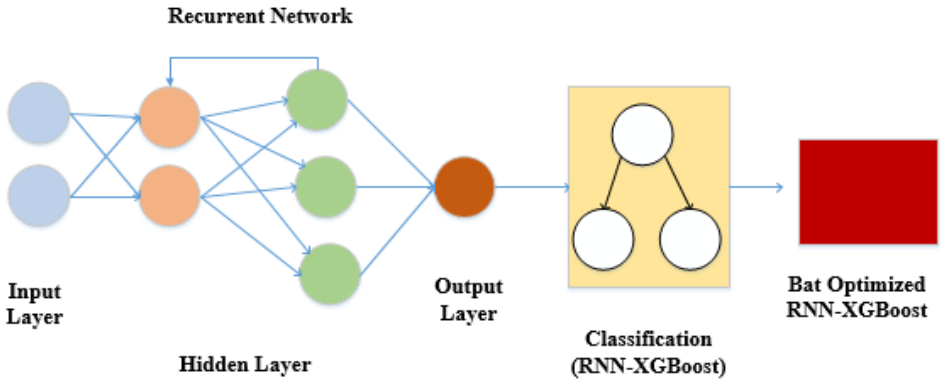


Fig. 2. Architecture diagram for proposed BARXG model.

The overall architecture diagram for Proposed BARXG Model is shown in Fig. 2. The Bat Algorithm-driven RNN-XGBoost model's overall design blends sequence modeling, combined learning, and bio-inspired optimization to produce a potent tool for classifying fetal health in prenatal monitoring. This novel strategy has the potential to greatly raise the standard of care provided to pregnant moms and their unborn kids.

1) RNN: An artificial neural network type called a recurrent neural network (RNN) is made to handle information in repetitions. These perform particularly well in tasks involving sequences, including speech, conversation, time series data, and numerous other tasks. For the purpose of to anticipate the layer's output, RNNs operate on the basis of reserving a certain layer's output then feeding it back into their input. One layer with recurrent neural networks is created by compressing the nodes within the various neural network layers. Both the present input data and previous inputs can be handled sequentially by an RNN. Because RNNs have memory within them, they are able to retain earlier inputs. Provide the RNN with a series of values as input at each time step. The Hidden state of an RNN, while retains certain details regarding a sequence, is its primary and crucial characteristic. Because the state retains recall of the prior input within the network, this state is known as well as Memory State. In order to create the output, it does a similar job on each of the inputs

and hidden layers using identical settings for each input. In contrast to other neural networks, that lowers the complexity associated with the features. During every time step, that exists a fixed activation function unit in the recurrent neural network. Every unit possesses an internal state known as its hidden state. During a particular time, each hidden state represents the prior information which the network presently possesses. This hidden state gets revised at each time step to reflect any modifications to the network's previous information. The recurrence relation listed below is used to modify the hidden state. The following is the formula to find the present state Eq. (4).

$$s_u = f(s_{u-1,} w_u) \tag{4}$$

where, s_u represents the present state; s_{u-1} represents the previous state; w_u represents the input state.

By using the following Eq. (5) the hidden state can be calculated.

$$H\_t = V(G^* w_u + U^* H\_(t-1) \ + n) \tag{5}$$

where, H_t represents the hidden state at time step; G represents the weight matrix that multiplies the current input w_u; w_u represents the input at time step; U is a weight matrix that is multiplied by H_(t-1) previous hidden state; H_(t-1) this refers to the hidden state that was a part of the present hidden state at a prior time step. H_(t-1): This refers to the hidden state that was a part of the present hidden state at a prior time step; n this represents expression for bias.

2) XGBoost algorithm: Gradient Boosting methods operate by learning ensembles on shallow decision trees. The framework fits the subsequent decision tree by using its remaining error in each iteration. A weighted total is used to get the final forecast after several trees have been constructed. In contrast, the trees in an Extreme Gradient Boost model are constructed parallel to one another rather than sequentially. In addition to improving speed, this shortens the period needed to fit data into a model. Within the scientific community, this framework is highly regarded for its ability to solve a wide range of issues. XGBoost is used for classification for fetal health.

For data preparation Let the value X represent the feature matrices, whereby the features are contained in a N x M matrix. Y, an N-dimensional vector containing the three fetal health labels (0 for Normal, 1 for Suspect, and 2 for Pathological), should be the desired vector. The total of the normalization and loss terms can be used to describe the objectives function in XGBoost. The total of the normalization and loss terms can be used to describe the objective function(o) in XGBoost is mentioned in Eq. (6).

$$o = L(x, \hat{x}) + \Omega(y) \tag{6}$$

where, the loss function quantifying the difference between the real names (x) and predicted names ($\hat{x}$) is represented by the expression $L(x, \hat{x})$.The regularization term, Ω(y) regulates the ensemble of trees' complexities. The anticipated designation for every specimen is acquired by adding the forecasts of many decision trees, every one of which is influenced using a coefficient α is mentioned in Eq. (7).

$$\hat{z}(y) = \sum \alpha^* k(y) \tag{7}$$

Where, α provides the weight of every decision tree k(y), and $\hat{z}(y)$ represents the expected labelling for a sample x. From an input feature vector x, every decision tree within the ensemble appears by the sum of its leaf scores (w) is mentioned in Eq. (8).

$$I(j) = \sum u \tag{8}$$

where, each decision tree's leaf scores are represented using the letter u. To regulate the level of complexity of the individual trees, the regularization term Ω(y) incorporates both L1 and L2 regularization over the leaf scores. XGBoost employs a gradient boosting technique to maximize the objective function(o) in order to determine the optimum combination among decision trees and associated variables, including α, u, and Ω(y).

3) BAT optimization for hyperparameter tuning: The bat algorithm, often known as the BA, was an algorithm which mimics the echolocation as an activity of bats to enable to carry out worldwide optimization. Considering its superior performance, the BA is frequently utilized across a variety of optimizations situations. The RNN and XGBoost models' hyperparameters may be optimized using the Bat Algorithm. Typically, bats utilize echolocation to locate food. Bats typically emit small pulses while removing it, but once they come upon food, they start sending off pulses more often along with higher rates. A frequencies-tuning result from a rise within frequency, which decreases overall echolocation period and improves the precision of location is men.

$$f_j(r+1) = f_j(r) + m_j(r+1) \tag{9}$$

$$c_l(r+1) = c_l(r) + \left(f_j(r) - w(r)\right).x_k \tag{10}$$

$$x_k = x_m + (x_a - x_m).\beta \tag{11}$$

When the quantity of repetitions rises, every k within the typical bat algorithm has a determined location f_j is mentioned in Eq. (9) and c_l velocity in the search space is mentioned in Eq. (10). One may compute the new coordinates x_k along with velocities in the following way Eq. (11). where β is a uniformly distributed randomized vector with a range of [0, 1]. The entire optimum solution at the moment is $w(r)$, where $x_m = 0$, $x_a = 1$ is mentioned in Eq. (11).

$$f_j(r+1) = \vec{e}(r) + \varepsilon \bar{B}(r) \tag{12}$$

where, ε is a random value between -1 and 1 is represented in Eq. (12). d signifies a random integer between -1 and 1 and l(d) is the population's average loudness. Furthermore, it accomplishes worldwide search via managing pulse rate $f_j(r+1)$ and loudness (Loudness (t +1) is mentioned in Eq. (13).

$$D_n(r+1) = \propto D_n(r) \tag{13}$$

$$w_i(r+1) = w_i(0)[1 - exp(-\gamma r)] \quad (14)$$

where, α >0, γ >0} when α and γ are constants. The starting ranges for pulse rate and loudness are denoted by (0)} and D_n (0)}, correspondingly is mentioned in Eq. (14).

The bat algorithm's processing phases are described in the following.

Step 1: Using Eq. (3), randomly create the frequency along with the location, velocity, and parameters for each bat.

Step 2: Use Eq. (1) as well as Eq. (2) to update each bat's location and velocity.

Step 3: Choose a random number (0<$r\ a\ n\ d$ 1<1) for every bat. If $rand1 < w_i(t)$ then update the temp location and compute the fitness level for the relevant bat using Eq. (4).

Step 4: Choose a random number (0<$r\ a\ n\ d$ 2<1) for every bat. If $r\ a\ n\ d$ 2<(t) and

$x\left(f_j(r)\right) < x(w(r))$, then update ($t$) and $r\ i\ (t$) using Eq. (5) along with Eq. (6), respectively.

Step 5: Sort each person according to fitness values, and then mark the top spot.

Step 6: When the condition is satisfied, the algorithm is complete; if not, proceed to Step 2.

The optimal configuration for both the RNN along with XGBoost models is represented by the optimum solution, or collection of hyperparameters, after the algorithm has finished running.

V. Results and Discussion

The results section provides a comprehensive overview of the outcomes and findings obtained from the experimental evaluation of the Bat Algorithm-Driven XGB-RNN For Optimal Fetal Health Classification in Pregnancy Monitoring. To ensure the quality of the dataset, preparation and data collection are the first steps in the procedure. XGBoost and RNN model modifying need independent optimization of hyperparameters, that's where the Bat Algorithm excels. The method of optimization includes the adjustment of hyperparameters such as RNN, tree depths, and learning rates. The Bat Algorithm runs repeatedly, assessing the accuracy of the models at each stage and modifying the hyperparameters according to ideas borrowed from echolocation. A measure of fitness that takes into account classification parameters such accuracy, F1-score, is used to gauge how well the framework performs. Whenever a termination criterion—such as a number of iterations or adequate model performance—is satisfied, the optimization loop keeps going. The hidden key to the model's performance is the resulting optimal selection of hyperparameters. The proposed framework is implemented in python. A device with an Intel(R) Core, 8GB of RAM, and windows 10 operating system is utilized.

A. Outcome of Fetal Health Classification by Proposed BARXG Model

Fig. 3 shows the categorization of fetal health dataset in percentage. As can be seen in the dataset, out of the 2126 samples, 1655 are normal, 295 are suspicious, and 176 are abnormal entries. Within the dataset, the fetal heart rate (FHR) patterns were classified as normal (0), suspicious (1), and abnormal (2).

The statistical summaries of the CTG data properties are displayed in Table I. This shows the outcome after the classification. The default Standard Scaler Python package is used to normalize the CTG data. By calculating the f-score along with bringing it into the same range, the Standard Scaler adjusts the data, facilitating computation and comparison.

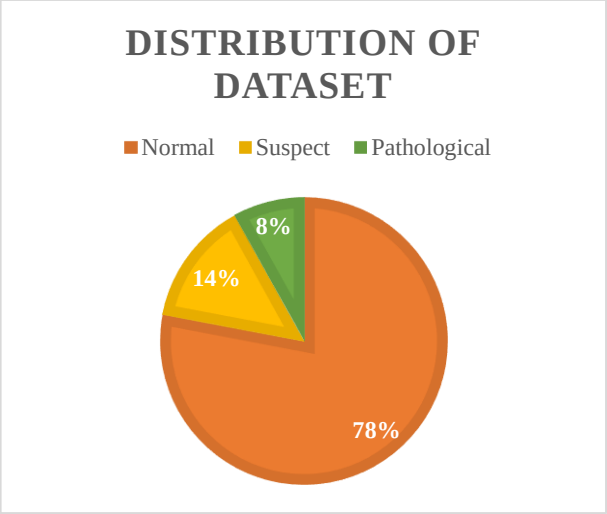


Fig. 3. Categorization of fetal health dataset.

TABLE I. Attributes from the Dataset

Attribute	Description and Unit	Mean	Std	Min	Max
Baseline value	Beats per minute	133.3039	9.840844	106	160
Accelerations	Accelerations per second	0.003178	0.003866	0	0.019
Fetal movement	Fetal movements per second	0.009481	0.046666	0	0.481
Uterine contractions	Uterine contractions per second	0.004366	0.002946	0	0.015
Fetal health	Fetal state class (0: normal (N); 1: suspect (S); 2: pathological (P))	-	-	0	2

TABLE II. RESULTS OF PROPOSED BARXG MODEL

Category	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Normal	0.98	0.98	0.96	0.97
Suspect	0.93	0.88	0.84	0.85
Pathology	0.90	0.87	0.86	0.86

Table II summarizes the results of the proposed BARXG model for the categorization of fetal health. The total accuracy provided by the proposed BARXG model was 98.2%. Precision, recall, and F1-score outperformed for each of the three classes. The accuracy of the forecast for healthy fetal cases is 98%, that of suspected fetus cases is 93%, and that of pathological fetus cases is 90%.

B. Performance Evaluation

For comparison the SVM, Random Forest Classifier, LGBM, EHG methods performance is compared with the proposed BARXG model. Precision, recall, F1-score, and accuracy were utilized as segmentation of the driver drowsiness evaluation criteria for comparison. The model was evaluated using these parameters. They are shown below:

A frequently used indicator to assess the effectiveness of categorization tasks is accuracy. The accuracy is computed by dividing the total number of predicts by the number of right predictions. It is described using an Eq. (15).

$$Accuracy = \frac{RN+RP}{RP+AP+RN+AN} \quad (15)$$

where, 'RN' means true negative; 'RP' means true positive; 'AP' means false positive; 'RN' means true negative; 'AN' means false negative.

A classification model's positive predictions are evaluated using a measure called precision. When false positive mistakes are expensive or undesired, it is especially crucial. To compute precision, use the formula below Eq. (16).

$$Precision = \frac{RP}{TP+FP} \quad (16)$$

where, 'RP' represents true positive and 'FP' represents false positive.

Recall, sometimes referred to as sensitivities or real-positive rate, is a statistic used to evaluate a classification model's capacity to accurately identify every relevant occurrence of a given class. The following Eq. (17) is used to calculate recall.

$$Recall = \frac{RP}{RP+AN} \quad (17)$$

The F1 score is a statistic that combines accuracy and recall to give a fair evaluation of the effectiveness of a classification model. It is especially helpful when you're trying to balance reducing inaccurate results (precision) and avoiding false negatives (recall) while maintaining accuracy. Eq. (18), which calculates the F1 score, is as follows.

$$F1\ score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (18)$$

The suggested model's accuracy is displayed in Table III. It compares the suggested approach's accuracy (98.2%) with existing approaches' recall (96.7%), precision (97.7%) and F1-score (98%) values. The proposed methodology, BARXG outperforms the currently used methods, Random Forest classifier (93%), EHG (88%), SVM (84%), Voting classifier (95%) and LGBM (96%), in terms of accuracy (98.2%) and precision (97.7%).

Fig. 4 depicts the graphic depiction of the performance metrics of proposed with existing approaches. The proposed BARXG method demonstrates the highest accuracy across all five categories Random Forest Classifier, EHG, SVM, Voting Classifier, LGBM, with 98.2% high accuracy. On tiny or noisy datasets, Random Forests may overfit, which will lower their capacity for generalization effectiveness. EHG is an invasive technique for measuring fetal growth since it requires affixing sensors onto the uterine wall. Because of their computational complexity, SVMs could not scale well to very big datasets. The variety of a Voting Classifier's base models determines how effective it is. It might not result in appreciable gains if the basic models are comparable. For LGBM models, hyperparameter tuning can be complex and time-consuming.

Fig. 5 shows the training and testing accuracy of proposed BARXG model. During training, the BAT Algorithm-Driven RNN-XGBoost model for fetal health classification performed well, with a training accuracy of almost 99%.

Nonetheless, it retained strong extrapolation to novel data, with an approximate 98% testing accuracy. This suggests that the model is a viable method for assessing fetal health in real-world clinical situations as it is capable of learning effectively via the training data and produce precise predictions on previously encountered cases.

TABLE III. PERFORMANCE METRICS OF PROPOSED BARXG MODEL IS EVALUATED WITH EXISTING METHODS

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
SVM [18]	84	86	88	85
EHG [13]	88	87	86	86
LGBM [19]	96	95	94	95
Voting Classifier [12]	95	94	93	94.8
Random Forest Classifier [20]	93	91	92	94
Proposed BARXG Model	98.2	97.7	96.7	98

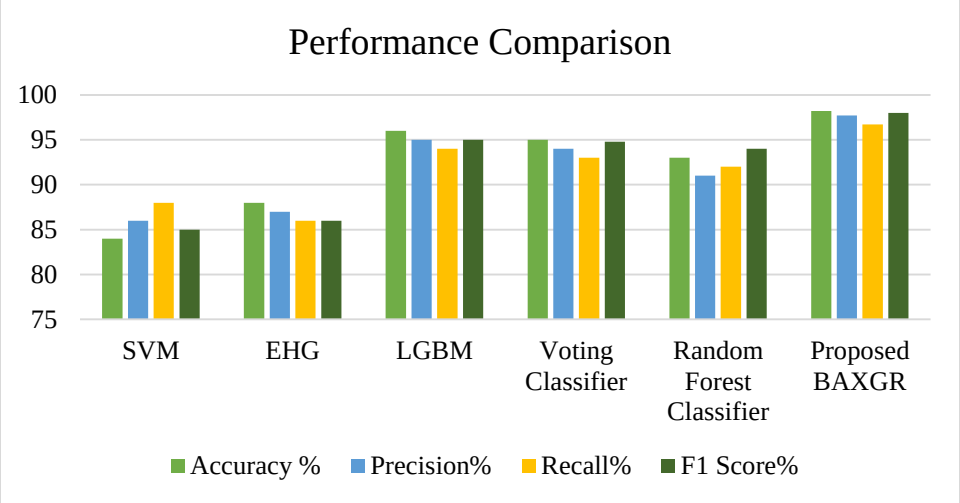


Fig. 4. Graphical depiction of the performance metrics of proposed BARXG with existing approaches.

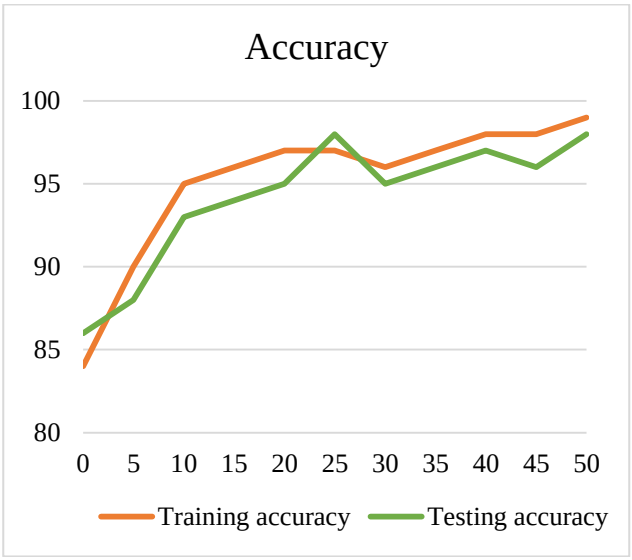


Fig. 5. Graphical depiction for training and testing accuracy of proposed BARXG model.

Fig. 6. Graphical depiction for training and testing loss of proposed BARXG model.

Fig. 6 shows the training and testing loss of the proposed model. The main goal of this model's training phase is to use the Bat Algorithm to fine-tune the RNN and XGBoost models' parameters. During the testing phase, the model's generalization skills are evaluated by analysing how well it performs on untested data. This step involves computing the testing loss.

The accuracy of the model and its ability to generalize to new, untested fetal health data are measured by the testing loss. The testing loss evaluates the model's capacity to produce accurate predictions on fresh, untested data, ultimately determining the efficacy of this novel approach in fetal health classification. The training loss is minimized by optimizing model parameters using the Bat Algorithm.

C. Discussion

Recurrent neural networks (RNNs) and XGBoost in combination with the Bat Algorithm (BA) show promise as a way to categorize fetal health in pregnancy monitoring. The accuracy of prediction of the XGB-RNN model is improved by the creative application of BA for parameter optimization. BA modifies model parameters by mimicking the echolocation behaviour of bats, which may enhance the precision of fetal health forecasts. The combination of XGBoost's gradient boosting and RNNs' sequence modelling allows for the effective processing of time-series data, such as fetal monitoring records. With the help of BA, this innovative method provides insights into complex data patterns, which could improve our comprehension of the dynamics of fetal health. Further research could improve scalability by addressing adaptability issues and convergence rates in a variety of data distributions, as well as handling bigger datasets and real-time processing. Larger datasets will be used in future research to assess BARXG's adaptability and continuous processing capabilities. The Bat Algorithm's continued adaptation to a variety of datasets and its implementation into healthcare systems are important future directions. A few of the research limitations are that enormous datasets may not scale well, validation of continuous processing is required, and the Bat Algorithm may not be as

flexible with various data distributions. In order to classify fetal health throughout pregnancy monitoring, BARXG's wider application will need to carefully validate and take these factors into consideration when investigating extension in various healthcare environments.

VI. Conclusion and Future Scope

In conclusion, a potential advancement in the field of maternal-fetal medicine is the investigation concerning the Bat Algorithm-Driven XGB-RNN for optimal fetal health categorization in pregnancy monitoring. Fetal health assessment becomes more potent and precise when the synergistic powers of XGBoost and RNNs are combined with the optimization inspired by nature of the Bat Algorithm. This method's benefits—such as enhanced precision, reliable analysis of time series, and clinical applicability highlight its potential to transform pregnancy monitoring and enhance outcomes for pregnant women and their unborn babies. It is obvious that the knowledge gathered by refining and validating this approach will have a significant influence upon clinical practice, ultimately resulting in healthier pregnancies and better care for the mother and fetus. Future improvements, thorough clinical validation, and continuous development will be necessary as this research develops in order to fully realize the potential of this novel strategy, which will ultimately help pregnant women and their kids as well as improve the present level of healthcare in the area of pregnancy monitoring. Maternal-fetal medicine may be profoundly impacted by more validation and clinical practice integration of this strategy. Working together with organizations and healthcare practitioners is essential to guaranteeing the method's efficacy in practical settings. It is crucial to investigate strategies for elucidating the model's predictions. Because they can comprehend and confirm the reasoning behind each categorization, healthcare professionals' trust may be increased through the development of interpretable models.

References

[1] L. Davidson and M. R. Boland, “Enabling pregnant women and their physicians to make informed medication decisions using artificial intelligence,” J Pharmacokinet Pharmacodyn, vol. 47, no. 4, pp. 305–318, Aug. 2020, doi: 10.1007/s10928-020-09685-1.

[2] O. E. F. Shaw and J. Y. Yager, “Preventing childhood and lifelong disability: Maternal dietary supplementation for perinatal brain injury,” Pharmacological Research, vol. 139, pp. 228–242, Jan. 2019, doi: 10.1016/j.phrs.2018.08.022.

[3] E. S. Green and P. C. Arck, “Pathogenesis of preterm birth: bidirectional inflammation in mother and fetus,” Semin Immunopathol, vol. 42, no. 4, pp. 413–429, Aug. 2020, doi: 10.1007/s00281-020-00807-y.

[4] K. Yammine, M. Eric, and C. Assi, “Variations and morphometrics of palmaris longus in fetuses: a meta-analysis of cadaveric studies,” Surg Radiol Anat, vol. 42, no. 3, pp. 281–287, Mar. 2020, doi: 10.1007/s00276-019-02391-9.

[5] C. M. J. Tan and A. J. Lewandowski, “The Transitional Heart: From Early Embryonic and Fetal Development to Neonatal Life,” Fetal Diagn Ther, vol. 47, no. 5, pp. 373–386, 2020, doi: 10.1159/000501906.

[6] A. Mehbodniya et al., “Fetal health classification from cardiotocographic data using machine learning,” Expert Systems, vol. 39, no. 6, p. e12899, Jul. 2022, doi: 10.1111/exsy.12899.

[7] M. G. Signorini, N. Pini, A. Malovini, R. Bellazzi, and G. Magenes, “Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring,” Computer Methods and Programs in Biomedicine, vol. 185, p. 105015, Mar. 2020, doi: 10.1016/j.cmpb.2019.105015.

[8] M. Chaturvedi, S. Agrawal, and S. Silakari, “An Investigation into Techniques used for Fetal Health Classification,” CSEIJ, vol. 12, no. 1, pp. 35–42, Feb. 2022, doi: 10.5121/cseij.2022.12105.

[9] L. M. Daly, F. M. Boyle, K. Gibbons, H. Le, J. Roberts, and V. Flenady, “Mobile applications providing guidance about decreased fetal movement: Review and content analysis,” Women and Birth, vol. 32, no. 3, pp. e289–e296, Jun. 2019, doi: 10.1016/j.wombi.2018.07.020.

[10] X. Zhao, X. Zeng, L. Koehl, G. Tartare, and J. De Jonckheere, “A Wearable System for In-Home and Long-Term Assessment of Fetal Movement,” IRBM, vol. 41, no. 4, pp. 205–211, Aug. 2020, doi: 10.1016/j.irbm.2019.11.003.

[11] R. Abiyev, J. B. Idoko, H. Altıparmak, and M. Tüzünkan, “Fetal Health State Detection Using Interval Type-2 Fuzzy Neural Networks,” Diagnostics, vol. 13, no. 10, p. 1690, May 2023, doi: 10.3390/diagnostics13101690.

[12] M. Al Duhayyim, S. Abbas, A. Al Hejaili, N. Kryvinska, A. Almadhor, and H. Mughal, “Ensemble Learning for Fetal Health Classification,” Computer Systems Science and Engineering, vol. 47, no. 1, pp. 823–842, 2023, doi: 10.32604/csse.2023.037488.

[13] D. Hao, J. Peng, Y. Wang, J. Liu, X. Zhou, and D. Zheng, “Evaluation of convolutional neural network for recognizing uterine contractions with electrohysterogram,” Computers in Biology and Medicine, vol. 113, p. 103394, Oct. 2019, doi: 10.1016/j.compbiomed.2019.103394.

[14] M. Mesbah et al., “Automatic fetal movement recognition from multi-channel accelerometry data,” Computer Methods and Programs in Biomedicine, vol. 210, p. 106377, Oct. 2021, doi: 10.1016/j.cmpb.2021.106377.

[15] J. A. L. Marques et al., “IoT-Based Smart Health System for Ambulatory Maternal and Fetal Monitoring,” IEEE Internet Things J., vol. 8, no. 23, pp. 16814–16824, Dec. 2021, doi: 10.1109/JIOT.2020.3037759.

[16] M. T. Alam et al., “Comparative Analysis of Different Efficient Machine Learning Methods for Fetal Health Classification,” Applied Bionics and Biomechanics, vol. 2022, pp. 1–12, Apr. 2022, doi: 10.1155/2022/6321884.

[17] A. de . Campos et al., “Fetal Health Classification.” Accessed: Oct. 21, 2023. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification

[18] C. V. Ananth and J. S. Brandt, “Fetal growth and gestational age prediction by machine learning,” The Lancet Digital Health, vol. 2, no. 7, pp. e336–e337, Jul. 2020, doi: 10.1016/S2589-7500(20)30143-6.

[19] V. Nagabotu and A. Namburu, “Fetal health classification using LightGBM with Grid search based hyper parameter tuning,” ENG, vol. 18, Jul. 2023, doi: 10.2174/1872121186662307031558 34.

[20] S. Das, H. Mukherjee, K. Roy, and C. K. Saha, “Fetal Health Classification from Cardiotocograph for Both Stages of Labor—A Soft-Computing-Based Approach,” Diagnostics, vol. 13, no. 5, p. 858, Feb. 2023, doi: 10.3390/diagnostics13050858.

Efficiency Analysis of Firefly Optimization-Enhanced GAN-Driven Convolutional Model for Cost-Effective Melanoma Classification

Dr. Lakshmi K[1], Sridevi Gadde[2], Dr. Murali Krishna Puttagunta[3],
G.Dhanalakshmi[4], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[5]
Associate Professor, School of Computer Science and Applications, REVA University, Bangalore, India[1]
Assistant Professor, Raghu Engineering College, Department of Computer Science and Engineering, AP, India[2]
Assistant Professor, Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation Vaddeswaram, AP, India[3]
Associate Professor, Department of Information Technology, Panimalar Engineering College,
Nazarethpettai, Chennai, Tamil Nadu 600123, India[4]
Faculty of Informatics and Computing, UniSZA University, Malaysia[5]

Abstract*—Early identification is essential for successful treatment of melanoma, a potentially fatal type of skin cancer. This work takes a fresh approach to addressing the urgent need for an accurate and economical melanoma categorization system. Inaccuracy, efficiency, and resource usage are common problems with current techniques. A model that incorporates a number of innovative methods to get beyond these restrictions was used in this study. To improve data quality, first applied the pre-processing with a Gaussian filter and augment our dataset with Generative Adversarial Networks (GAN). To extract and classify features, this suggested model makes use of Convolutional Long Short-Term Memory (LSTM) networks. The model performs better and is substantially more accurate when Firefly Optimization is used. It analyses the model's ability to lower healthcare costs by doing a cost-effective analysis, especially when detecting melanoma, including situations involving bleeding lesions. The proposed FFO Enhanced Conv-LSTM's cost-effective analysis makes it possible to compare it favourably to deep convolutional neural networks (DCNN), showcasing its promise for melanoma classification accuracy and healthcare resource allocation optimization. For this study, Python software was used as the implementation tool. The suggested model achieves a 99.1% accuracy rate, which is better than current techniques. A comparative study with well-known models such as Res Net 50, Mobile Net, and Dense Net 169 highlights the notable enhancement provided by the proposed Firefly Optimization-enhanced Conv-LSTM method. This model offers a promising advancement in the precise and economical classification of melanoma due to its high accuracy and cost-effectiveness. In comparison to existing approaches like Res Net 50, Mobile Net, and Dense Net 169, the suggested Firefly Optimization-enhanced Convolutional LSTM (FFO Enhanced Conv-LSTM) method shows an average gain of roughly 5.6% in accuracy.

Keywords*—Melanoma; cost effective analysis; long short-term memory; firefly optimization; generative adversarial network

I. INTRODUCTION

Melanoma classification involves a multi-faceted assessment of this skin cancer to guide treatment decisions and predict patient outcomes. Histological categorization is one step in this procedure, which divides melanomas into subtypes depending on how they appear under a microscope [1]. Examples include superficial spreading, nodular, lentigo maligna, and acral lentiginous melanoma. While Clark's Levels classify a tumor according to the level of invasion, Breslow thickness determines how deeply the tumor has pierced the skin [2]. Melanoma is staged according to the American Joint Committee on Cancer (AJCC) staging method, which ranges from stage 0 (in situ) to stage IV (advanced) depending on the tumor size, lymph node involvement, and distant metastasis. The categorization is further refined by other variables such ulceration, mitotic rate, genetic alterations, and immunohistochemistry [3]. Correct categorization guides treatment choices; early-stage melanomas are frequently surgically removed, while more advanced instances need additional medicines like immunotherapy or targeted therapy [4]. Ongoing monitoring is crucial for post-treatment care. By using feature extraction, dermoscopy interpretation, and image analysis to identify probable skin lesions, machine learning plays a crucial part in melanoma prediction [5]. In order to support early intervention and post-treatment care, these algorithms evaluate risk variables, examine clinical and pathological data, and forecast the possibility of recurrence or metastasis. Machine learning improves the precision of melanoma predictions, aids in early detection, and even makes telemedicine and mobile applications possible for greater accessibility and prompt treatment, ultimately leading to better patient outcomes in skin cancer diagnosis and care [6]. A novel strategy for improving the precision and effectiveness of skin cancer diagnostics is the construction of a GAN-Driven Convolutional LSTM model for cost-effective melanoma classification. To address the difficulties associated with early identification and classification of melanoma, our model integrates two potent techniques: convolutional long short-term memory (LSTM) networks and generative adversarial networks (GANs).

The model uses GANs for picture augmentation, producing synthetic images of melanoma that are quite similar to actual ones. In order to solve the issue of limited labelled

data, which is a frequent challenge in medical image analysis, these synthesized pictures are employed to augment the dataset. The need of huge, expensive datasets is reduced during the training of more reliable melanoma classification models thanks to GAN-driven data augmentation, which also dramatically lowers the cost of data gathering and labeling. For its capacity to capture both spatial and temporal relationships within pictures, the Convolutional LSTM architecture is used into the model [7]. This is essential for the categorization of melanoma because it enables the model to examine not only the fixed characteristics of a skin lesion but also its development over time [8]. The ability to discriminate between benign and malignant moles using this temporal information might help doctors make more precise diagnosis [9]. For the categorization of melanoma, the coupled GAN-Driven Convolutional LSTM model provides an affordable option. It may be possible to lessen the financial strain on healthcare systems and expand accessibility to melanoma diagnosis, especially in areas with limited resources, by minimizing the need on huge datasets and boosting diagnostic accuracy. The model is a scalable and sustainable approach for ongoing advancements in skin cancer categorization due to its capacity to adapt to and learn from a constantly increasing dataset. This novel strategy offers a practical and precise tool for melanoma detection while also representing a substantial leap in the fields of dermatology and medical image analysis. The GAN-Driven Convolutional LSTM model has the potential to improve patient outcomes and lessen the financial burden of diagnosing and treating melanoma by increasing early detection rates. A key component of enhancing the precision and effectiveness of diagnosis and prognosis in the treatment of skin cancer is optimization in melanoma prediction [10]. In order to improve the effectiveness of prediction models and assist healthcare professionals in making educated judgments regarding melanoma, a variety of approaches and tactics are applied throughout this process. Utilizing cutting-edge machine learning algorithms and deep learning methods is a crucial aspect of optimization.

Examples of neural networks that have demonstrated substantial promise in image analysis include convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which allow the identification of minute details and patterns in skin lesions that are suggestive of melanoma. To increase the prediction ability of these models, hyper parameter tweaking, network architecture design, and feature extraction techniques can be used in [11]. Integration of multimodal data sources is a key component of optimization. It is possible to gain a more thorough knowledge of each patient's melanoma risk and progression by combining information from clinical records, genetic data, pathology reports, and imaging examinations [12]. Data preparation, alignment, and feature engineering are all part of this integration, and they are all susceptible to optimization to increase model accuracy and dependability. Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA), two feature selection techniques, can help to simplify the model while preserving important data. Ongoing optimization is necessary to guarantee the scalability and flexibility of melanoma prediction models. This involves revising models when fresh information emerges and as our understanding of melanoma deepens [13]. Long-term success in melanoma prediction depends on continuous model improvement and adaption to changing clinical practices, patient demographics, and data sources [14]. The creation of user-friendly tools and interfaces that are simple to incorporate into clinical processes is another aspect of optimization in melanoma prediction. Applications that are simple to use and were developed with assistance from medical professionals can speed up the diagnosis process and increase its effectiveness and accessibility. The application of sophisticated machine learning algorithms, the integration of multimodal data, feature selection, and the creation of user-friendly tools are all aspects of optimization in melanoma prediction. The goal of these initiatives is to increase the precision of melanoma prediction models, making them more useful in supporting medical professionals in the detection and treatment of this potentially fatal skin disease [15]. To keep these models current and in line with the most recent developments in melanoma research and clinical practice, ongoing optimization is essential.

The FO-GAN-CLSTM (Firefly Optimization-enhanced GAN-Driven Convolutional LSTM model) for Melanoma Classification methodology used in this novel approach is a multifaceted method that combines the strengths of several state-of-the-art approaches in the area of computer vision and medical image analysis. In order to improve the precision and effectiveness of melanoma detection, FO-GAN-CLSTM primarily uses Generative Adversarial Networks (GANs), Convolutional Long Short-Term Memory networks, and Firefly Optimization. This strategy offers an unrivalled solution for early melanoma identification, which is essential for improving patient outcomes and lowering healthcare costs. It perfectly integrates various approaches. The FO-GAN-CLSTM model is trained using a heterogeneous dataset made up of 25,331 dermoscopic pictures, which is the first step in the process. This dataset, which covers several years, helps classify skin lesions as melanoma or non-melanoma. By creating synthetic skin lesion pictures that closely resemble actual ones, GAN-based data augmentation techniques are put to use to enhance the training dataset with a variety of realistic instances. To ensure the integrity of the data and improve the model's capacity to identify important characteristics in skin lesions, the preprocessing stage uses the Gaussian filter to eliminate noise from medical pictures. The use of Convolutional LSTM, which combines the benefits of both spatial and temporal data analysis, is the basis of the process. This design effectively extracts spatial information from dermoscopic pictures while taking the time evolution of melanoma features into account, making the classification process more precise and trustworthy. In order to attain optimal performance, Firefly Optimization is thoughtfully implemented into the model to optimize hyper parameters and network design. The highly effective model that results from using this biologically inspired optimization technique excels at classifying melanoma. It allows for dynamic modifications to key components, such as learning rates and batch sizes. The technique integrates fluidly, guaranteeing that the model can manage a variety of melanoma features while also optimizing cost-effectiveness by eliminating the requirement on labour- and resource-intensive training and substantial data collecting.

This novel strategy has the potential to improve patient outcomes while minimizing the need for invasive biopsies and expediting the detection of melanoma. The following are the research study's main contributions,

- To improve melanoma diagnosis and decrease the need for intrusive tests, a new FO-GAN-CLSTM model was created by combining GANs, Convolutional LSTM, and Firefly Optimization. This model will benefit patients to a greater extent.
- Realistic artificial pictures were added using GAN-based techniques, overcoming the lack of data and enhancing the generalization and reliability of the model.
- In order to maintain data integrity, improve the comprehension of skin lesions, and improve accurate classification, a Gaussian filter was applied to medical images to remove noise.
- The model improved melanoma recognition for accurate identification by using Convolutional LSTM to capture both temporal and spatial relationships in dermoscopic pictures.
- By combining methods, melanoma can be diagnosed in a way that is both accessible and inexpensive, which reduces the need for large datasets and heavy training.

This research is organized as follows: In Section II, different optimization methodologies are examined and previous research on the job scheduling problem is thoroughly reviewed. Section III discusses the problematic assertion. Section IV explores the suggested method. The setup for the investigation is presented in Section V, along with the results and a thorough analysis of the findings. The paper's conclusion is finally provided in Section VI.

II. Related Works

Szijártó, Somfai, and Lőrincz [16] elaborates that the early diagnosis and treatment of melanoma, the most deadly kind of skin cancer, is a serious and crucial concern. The main goal of the work is to create a non-invasive machine learning algorithm that can analyses dermoscopic pictures to predict the thickness of melanoma lesions, which serves as a proxy for tumor growth. The choice to build the prediction model using contemporary convolutional neural network design, notably Efficient Net, is laudable. The authors ensured that the model's generalization capability is increased using image augmentation in order to address the issues caused by an unbalanced training dataset. Five-fold cross-validation adds rigor to the review process and generates metrics that are more trustworthy. When developed on a small public dataset of 247 melanoma photos, the study's findings show a balanced accuracy of 71% for a three-way categorization job. Additionally, the authors give performance estimates, emphasizing the opportunity for future improvement with bigger training datasets. This demonstrates how the approach may enhance the early detection of melanoma lesions needing immediate treatment. The paper's results are both obvious and important. The authors' model is a brand-new, cutting-edge method for categorizing melanoma thicknesses, an important consideration in treatment choices. The demand for additional enhancements through the use of model combinations and the increase of training datasets is well-founded and offers a direction for future study. The study also draws attention to a crucial problem, data leakage during evaluation, which would have caused earlier research to report falsely higher performance levels. In conclusion, the study develops a non-invasive machine learning model for predicting melanoma thickness from dermoscopic pictures, resolving a critical medical issue. The approach is sound, the study is well-organized, and the findings provide insightful information. Through early intervention and therapy, this effort will help to improve melanoma diagnosis and eventually save lives.

Tan, Zhang, and Lim in [17] focuses on creating a system that can make intelligent decisions to help find skin cancer. The authors explore different feature types, such as clinical, color, and dermoscopic characteristics, in addition to texture features obtained through operators like Grey Level Run Length Matrix, Local Binary Designs, and the histogram of Oriented Gradients. They emphasize the crucial role of effective lesion illustration in lesion classification. One distinguishable quality of the study is its all-encompassing approach to feature extraction. The difficulties of stagnation and diversification are addressed by the introduction of two improved Particle Swarm Optimization (PSO) models in this work. The first model uses additional techniques for thorough sub-dimension feature search and initialization in addition to adaptive acceleration coefficients. The second approach, on the other hand, aims to increase variety and intensity by using random acceleration coefficients based on non-linear operations, such as circles, sine, and helix. The research is improved by this PSO methodological improvement. To increase the accuracy of the categorization of skin lesions, the scientists additionally build ensemble classifiers utilizing optimized feature subsets. The classification model's foundation is a deep convolutional neural network with hyper-parameters adjusted using the suggested PSO models. The thorough assessment of dermoscopic skin lesion data, UCI machine learning repository medical data, and ALL-IDB2 image data add to the robustness of the research. Deep learning in medical image analysis is a promising technique. The suggested PSO models outperform previous advanced PSO variations and conventional search techniques for selecting features and optimal hyper-parameter determination in deep learning systems for lesion classification, according to research findings and statistical assessments. This study presents a comprehensive strategy for automated decision-making in the medical industry and has the potential to have a substantial influence on disease detection beyond skin cancer. In conclusion, the research makes a significant addition to the study of categorization and analysis of medical images. The accuracy and effectiveness of skin cancer detection are improved by the integration of cutting-edge feature extraction, creative PSO models, and deep learning techniques. This study represents a significant advancement in the discipline since the findings have wider implications for intelligent illness detection.

Akter et al. [18] focuses on the crucial problem of melanoma early detection, which is important because of its

high fatality rate. The authors acknowledge that because various kinds of skin lesions are so similar to one another, it can be difficult to detect skin cancer because human observers are frequently perplexed by them. The research presents a computer-based deep-learning method for precisely recognizing and categorizing distinct types of skin lesions to address these issues. Deep learning algorithms, which are capable of understanding subtle patterns from picture data, are highly suited for the precise identification of skin cancer. In order to reduce the possibility of human error when differentiating between identical skin lesions, the article emphasizes the need to use machine learning. Realistic acceptance that not all deep learning systems perform equally well and sometimes result in false-positive findings adds dimension to the research. Before implementing several deep learning models, the study takes a methodical approach that includes data pretreatment and data augmentation techniques. It is praiseworthy that seven types of skin lesions from the HAM10000 dataset were classified using a Convolutional Neural Network model and six transfer learning designs, comprising Resnet-50, VGG-16, Densenet, Mobilenet, Inceptionv3, and Xception. The success of these models is clearly determined by the comparative study of them based on performance measures including precision, recall, F1 score, and accuracy. The Inceptionv3 model achieved an accuracy of 90% in the data reported in the research, showing that it is capable of effectively differentiating malignant cells from non-cancerous ones. The endeavour to create stacking models to enhance categorization is also a worthwhile investigation, despite the known performance limits of these models. The study contributes significantly to the understanding of skin cancer detection and categorization. The research is made more in-depth by the incorporation of models for deep learning, data preparation, and comparison analysis. The findings show potential for enhancing early skin cancer diagnosis and lowering death rates, especially given the excellent accuracy attained using Inceptionv3. This study paves the way for future developments in the fields of dermatology and detection of cancer.

Tyagi et al. [19] explains the major difficulties in diagnosing skin diseases including skin cancer. Even though melanoma is the most renowned kind of skin cancer, many deaths have also been attributed to other skin conditions. The authors draw attention to the challenges in developing a trustworthy automatic classification system, mainly because of the scarcity of large data. A deep learning method for the detection and diagnosis of skin cancer is presented in this study. Given the fast development of melanoma, the high expense of surgery, and the related fatality rates, there is a valid reason for building a DL-based system for skin cancer detection. Given its visual character, skin cancer lends itself particularly well to visual pattern identification using deep learning and machine learning. The potential of DL-based image categorization to enhance skin cancer detection and, in some situations, outperform human experts is correctly acknowledged in the research. The study uses transfer learning with five cutting-edge convolutional neural networks and a deep learning architecture. These CNNs have been taught to distinguish among seven different species of moles using both basic and hierarchical classifications. A potent strategy is to leverage data from the HAM10000 the database, which has a substantial amount of dermatoscopic pictures, together with data augmentation methods. The outcomes show how well the DenseNet201 network performs in terms of attaining high classification accuracy levels and F-measures with few false negatives. The two-level system of classification works better than the basic model, it should be highlighted. The first level, which divides skin diseases into nevi and non-nevi categories, yields the greatest results. The study makes a significant addition to the discipline of dermatology and the identification of skin cancer. Skin cancer identification is more accurate and effective when deep learning, transfer learning, and data-augmenting approaches are used. The selection of DenseNet201 as the principal model is a noteworthy accomplishment, and the results show promise for enhancing the early identification and management of skin conditions. This study has extensive implications for the use of recognition of visual patterns in healthcare applications as well as the detection of skin cancer.

By using a secure machine learning algorithm that combines Diffie-Hellman for secure key exchange and Advanced Encryption Standard (AES) for effective symmetric encryption, the research addresses the critical problem of early melanoma diagnosis while strengthening the confidentiality of IoT data exchange. Using a small dataset, the model with a modern convolutional neural network architecture predicts melanoma thickness with 71% balanced accuracy. Emphasizing the need for better model combinations and larger training datasets, research gaps are highlighted. Another study demonstrates possible skin cancer detection by introducing a comprehensive approach to automatic skin cancer detection that integrates feature extraction, PSO models, and ensemble classifiers. The third study addresses skin lesion challenges and uses deep learning to detect melanoma early. Using seven models including Inceptionv3. Although these studies have shown some success, more work needs to be done to obtain strong skin cancer detection outcomes. This includes expanding datasets and improving model combinations.

III. PROBLEM STATEMENT

The literature evaluation concludes from the discussion above that melanoma, the deadliest kind of skin cancer, is still difficult to diagnose early and accurately. Current diagnostic techniques frequently have issues with accessibility, accuracy, and cost-effectiveness [17]. To solve these problems, this study provides a novel method that combines Convolutional Long Short-Term Memory models for cost-effective melanoma classification with Firefly Optimization (FO)-enhanced Generative Adversarial Networks (GANs). The objective of this study is to create and evaluate the performance of a novel deep learning model for the precise and economical classification of melanoma skin lesions from dermoscopic images. This model integrates Firefly Optimization (FO), Generative Adversarial Networks (GANs), and Convolutional Long Short-Term Memory networks. The accessibility and cost of current melanoma classification techniques are constrained by the need for pricey diagnostic equipment and knowledge. The suggested concept seeks to offer a more affordable option. It's critical to diagnose

melanoma with great precision. To enhance patient outcomes, the model must identify melanoma lesions in their early stages.

IV. Melanoma Classification using Firefly Optimization-Enhanced GAN-Driven Convolutional LSTM Model

Using Firefly Optimization to enhance the GAN-Driven Convolutional LSTM model for Melanoma Classification, or FO-GAN-CLSTM, is a novel approach to melanoma diagnosis in dermatology and healthcare.

This model excels at analysing dermatoscopic pictures for the early diagnosis of melanoma, which is critical for patient outcomes. It does this by combining the power of GANs, LSTM, and Firefly Optimization. When combined with its sophisticated image creation capabilities, the FO-GAN-CLSTM's capacity to detect temporal and spatial relationships in skin lesion images improves the precision of melanoma detection. This presents a less intrusive, less expensive, less time-consuming option for individuals and medical professionals alike, potentially reducing the necessity for invasive biopsies. Block diagram for proposed FFO enhanced Conv-LSTM is shown in Fig. 1.

A. Data Collection

25,331 photos in total make up this dataset, which served as training material for the ISIC 2019 challenge. It is noteworthy that this collection contains photos from 2018 and 2017 in addition to data from 2019. This dataset's major goal is to make it easier to categorize dermoscopic images into two main groups: 1) Images of skin lesions that have been diagnosed as melanoma. 2) Pictures of skin lesions that are not classified as melanoma. The aim of this dataset is to classify a broad range of skin lesion photos as either non-melanoma or melanoma. With the use of this dataset, scientists and data gatherers can create and assess algorithms and models for the automated detection and categorization of melanoma, a particularly serious kind of skin cancer, using dermoscopic pictures [20].

B. GAN- based Data Augmentation

Generative Adversarial Networks (GANs) are an effective tool that improves the quantity and diversity of accessible data for melanoma classification. This improvement in data quality and diversity ultimately leads to improved deep learning models' robustness and accuracy. Both a generator and a discriminator make up a GAN in this process. Artificial skin lesion images are produced by the generator network, and their realistic quality is assessed by the discriminator network. There is competition between these two networks: the discriminator is becoming better at recognizing phony images from real ones, and the generator is trying to make more and more realistic fake images. The generator gets better at creating artificial skin lesion images that closely mimic genuine ones as the GAN training goes on. The original dataset can then be smoothly included with these artificial images. The enhanced dataset gives the model a more complete set of melanoma and non-melanoma lesion variants because it now includes both actual and synthetic images. Architecture of Generative Adversarial Network is depicted in Fig. 2.

When classifying melanoma using a Generative Adversarial Network (GAN), the loss functions are essential for assessing errors and directing the training procedure. Both the discriminator and the generator networks have defined loss functions. Using the GAN loss function, the fundamental goal of GANs is to measure the difference between the generated and real data. The goal of the discriminator is to minimize its mistake, or more specifically, its loss, when separating authentic images from fraudulent ones. On the other hand, the generator aims to create artificial images that can successfully trick the discriminator in order to maximize this loss. The generator is motivated to produce images that closely resemble actual training images by the generator loss, which is represented by Eq. (1).

$$\min_{Ge} V(Ge, Dis) = e_{y\sim\rho_n(N)}[\log(1 - Dis(Ge(N)))] \quad (1)$$

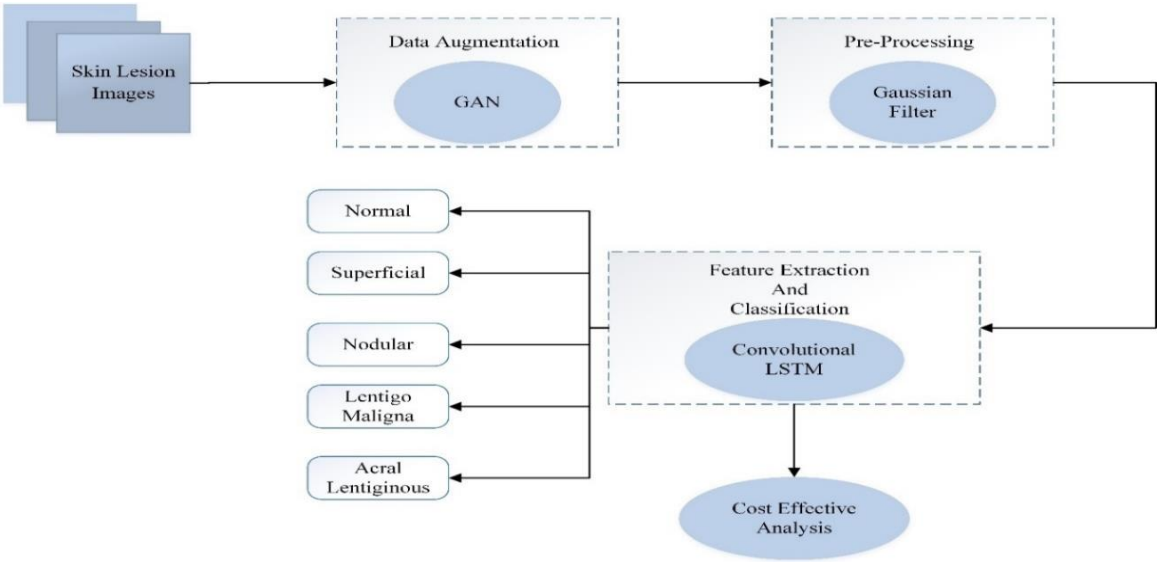


Fig. 1. Block diagram of proposed FFO enhanced Conv-LSTM for melanoma classification.

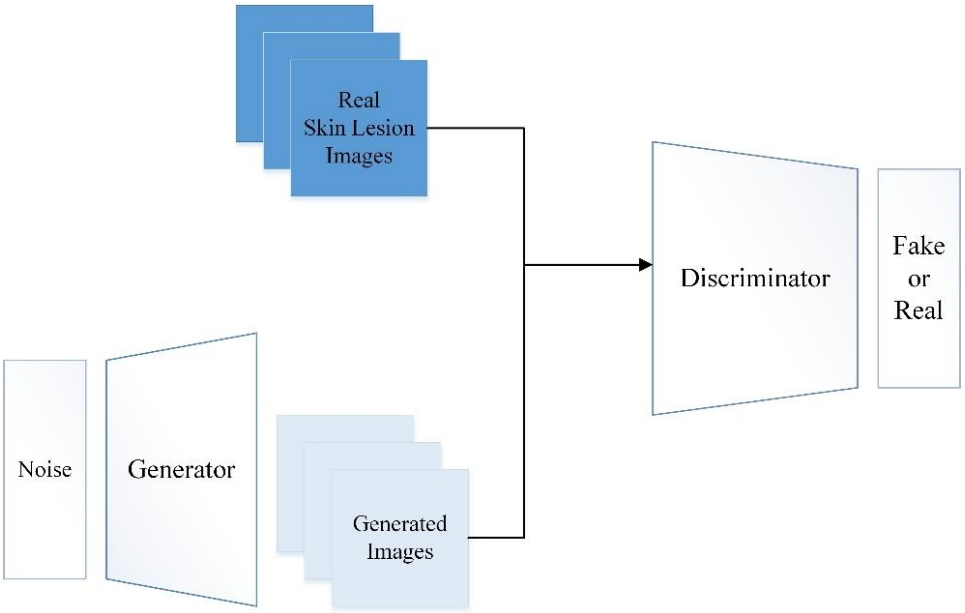


Fig. 2. Architecture of GAN.

Eq. (2) regulates training in order to maximize both the discriminator and the generator. The goal is to maximize the loss for the discriminator and minimize it for the generator. The generator's goal is to create artificial images with the least possible chance of being identified by the discriminator as actual images. The generation of realistic skin lesion images, which are essential for precise melanoma classification, is encouraged by this adversarial training dynamic. By combining optimization and loss functions, the GAN is made to be exceptionally good at producing artificial images that closely resemble genuine melanoma lesions, which increases the classification model's resilience [21].

$$\min_{Ge}\ \max_{Dis} V(Dis, Ge) = e_{y\sim\rho_{data}(y)}[\log Dis(y)] + e_{n\sim\rho_n(N)}\left[\log\left(Dis\left(Ge(N)\right)\right)\right] \quad (2)$$

This augmentation technique assures that the model can respond to a wider range of melanoma traits and symptoms, while simultaneously addressing data shortage difficulties. Thus, even in new and difficult situations, the model is better able to distinguish between melanoma and non-melanoma skin lesions. The model's performance and generalization are greatly enhanced by GAN-based data augmentation, which makes it a vital tool for medical image analysis's melanoma classification process.

C. Pre-Processing using Gaussian Filter

Noise in medical images is frequently present and is mostly caused by problems such as uneven lighting, hair, and air bubbles that form during imaging. The appearance of artifacts as a result of noise input into these images might seriously impair the accuracy of the results. This can ultimately result in inaccurate detection results. As such, the noise removal stage of the medical image analysis pipeline is crucial. Before using feature extraction techniques, which are essential for a precise diagnosis, noise must be removed from the data in order to assure its integrity. For the purpose of classifying melanoma, the Gaussian filter is one of the most important and commonly used methods in pre-processing. Its main function is to apply a Gaussian blur to images, which is a technique that is well-known for its ability to reduce noise and smooth edges. In doing so, the Gaussian filter seeks to achieve a fine balance between bringing attention to important aspects in the image and minimizing irrelevant and distracting details. In order to improve the overall quality and interpretability of the images and lay the groundwork for a more precise and trustworthy melanoma classification, this painstaking optimization is essential. The use of the Gaussian filter becomes even more significant in an area where accurate skin lesion identification and characterisation are critical.

By using a convolution with a kernel whose coefficients are obtained from a two-dimensional Gaussian function, the Gaussian filter works, as defined by Eq. (3). Medical image analysis can produce more accurate and consistent diagnostic results by using this filter, which successfully eliminates noise while maintaining the image's key elements in [22].

$$G(y, z) = \frac{1}{\sqrt{2\pi\rho^2}} e^{-\left(\frac{y^2+z^2}{2\rho^2}\right)} \quad (3)$$

where, the amount of blurring is indicated by the smoothing parameter, ρ.

D. Employing Convolutional LSTM for Feature Extraction and Classification of Melanoma

The input y_t, cell state c_t, output H_t, and weight matrix (Wxf) of conventional LSTM (Long Short-Term Memory) networks are all 1D vectors. This indicates that the weight matrix w_{yf} and the input y_t are fully coupled, and that a 1D vector is produced when $w_{yf} \times y_t$ is multiplied. Although this classical fully connected LSTM (fc-LSTM) works well for managing temporal correlations, it is not appropriate for spatially-required situations. Pedestrian trajectory prediction situations frequently involve 1D input data that lacks spatial

information, which could be lost as the network becomes deeper. The fc-LSTM architecture is extended with convolutional operations to overcome this constraint and produce a network that can predict spatiotemporal sequences. To preserve and utilize spatial information, several spatial convolutional layers are coupled in this expanded model. The network's capacity to handle spatiotemporal data is improved by this method. Convolutional LSTM architecture is presented in Fig. 3.

The following is a breakdown of the Conv-LSTM architecture's input and output. (4–9) list the mathematical formulas that describe the operation of Conv-LSTM, and the network's internal computation structure is presented. Here, ρ stands for the sigmoid function, × for a convolution operation, and ° for Hadamard element-wise multiplication. The input at a given time is represented by y_t, and the long-term memory values $c_t - 1$ and $H_t - 1$ are updated to c_t and H_t. Conv-LSTM uses the working memory $H_t - 1$ and input (y_t) together to determine the forget gate, which indicates how much long-term memory should be kept. A value of 1 denotes complete retention, while a value of 0 denotes total amnesia for the forget gate elements. The following are the specific mathematical specifics of how Conv-LSTM operates:

To obtain the information for the forget gate z_f, a convolutional neural network is first used by Eq. (4).

$$z_f = \rho(w_{yf} \times y_t + w_{hf} \times H_{t-1} + b_f) \quad (4)$$

In Eq. (5), the information should then be extracted from y_t, which is the long-term memory's candidate memory Z.

$$z_f = \tanh(w_y \times y_t + w_h \times H_{t-1} + b) \quad (5)$$

Combine the results from the first two phases in Eq. (6) and Eq. (7). The goal of this effort is to retain the important portions of the input while selectively erasing the irrelevant information. The updated long-term memory c_t is the outcome of this.

$$z_j = \rho(w_{yj} \times y_t + w_{hj} \times H_{t-1} + b_j) \quad (6)$$

$$c_t = z_f{}^{\circ}c_{t-1}{}^{\circ}z_j{}^{\circ}z \quad (7)$$

Using Eq. (8), update the working memory as part of the fourth stage. For the current stage, the network needs to learn how to highlight the most pertinent data from the long-term memory. To do this, use the formula below to determine the focus of attention vector z_o.

$$z_o = \rho(w_{yo} \times y_t + w_{ho} \times H_{t-1} + b_o) \quad (8)$$

Use the Eq. (9) to determine the working memory H_t in the fifth stage. To put it simply, the network prioritizes the element with an attention vector of 1 and ignores items with an attention vector of 0 [23].

$$H_t = z_o \times \tanh(c_t) \quad (9)$$

Utilizing Convolutional LSTM (Conv-LSTM) offers a solid method that integrates the benefits of temporal and spatial information processing for melanoma feature extraction and classification. Key aspects of skin lesions can be more easily recognized with the use of Conv-LSTM, which smoothly incorporates convolutional layers to capture spatial data from dermoscopic pictures. Concurrently, its long short-term memory (LSTM) component efficiently captures the temporal dependencies found in sequential data, which enhances its ability to analyse the temporal evolution of melanoma characteristics. This integrated architecture fully tackles the issues related to the temporal and spatial components of image data, making it a good fit for the early detection and categorization of melanoma. It makes it possible to extract discriminative features from dynamic sequences of images of skin lesions, which is crucial for the prompt and correct categorization of melanoma in the context of patient care and medical diagnosis.

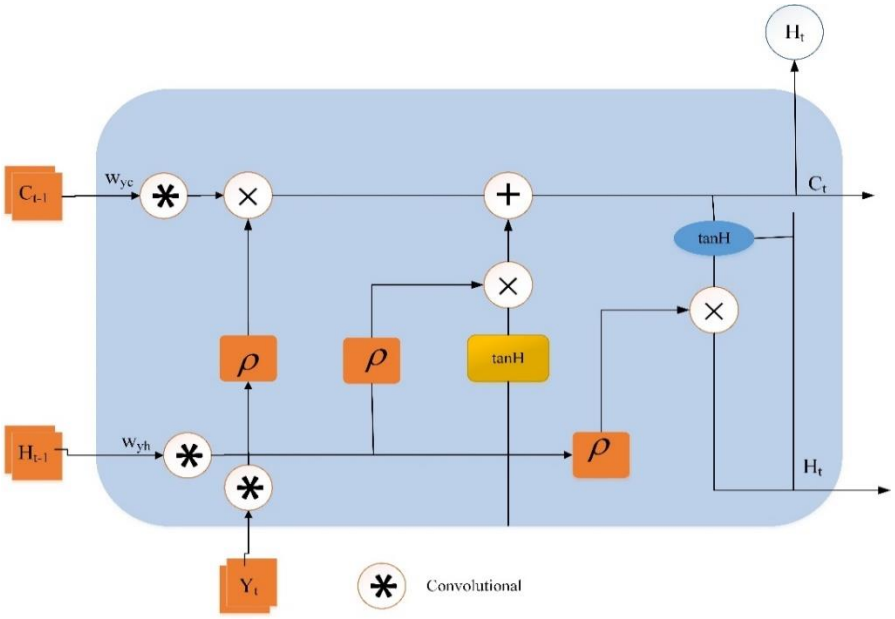


Fig. 3. Convolutional LSTM architecture.

E. Using Firefly Optimization for Enhancing the Proposed Model

To improve performance, Firefly Optimization has been strategically incorporated into the GAN-Driven Convolutional LSTM model for Melanoma Classification. Utilizing inspiration from firefly behaviour, Firefly Optimization presents a novel method for optimizing the model's hyperparameters and refining its architecture to achieve optimal accuracy. Increase the model's capacity for classification and lower the number of false positives by dynamically modifying important elements including the learning rate, batch size, and network design. Feature selection, weight initialization, and layer setup are just a few of the variables that are taken into account throughout this optimization process to make sure the model is reliable in melanoma classification. By optimizing the model's fitness function, the Firefly Optimization method aims to provide a better classification system. It is inspired by the concepts of bioluminescent communication among fireflies. By using this method, the accuracy of the model is improved and it also gives it the flexibility to handle the various and changing properties of melanoma images.

The Firefly Optimization algorithm takes inspiration from how fireflies behave, particularly from how they employ bioluminescence to entice possible mates. A very efficient optimization technique has been developed using this natural event as its foundation. To find the best solution, (10-15) have been developed mathematically.

$$l(H_t) \propto F(H_t) \tag{10}$$

$$l(r) = l_0 e^{-ar^2} \tag{11}$$

This particular situation is the result of the inverse square law, which is what happens when r in the equation $\frac{1}{r^2}$.

$$c_t \propto l(r) \tag{12}$$

$$c_t = c_{t_0} e^{-ar^2} \tag{13}$$

$$r_{il} = \|G_i - G_l\| = \sqrt{\sum_{d=1}^{d=N} (G_{id} - s_{ld})^2} \tag{14}$$

$$G_i = G_i + c_0 e^{-ar^2{}_{il}} (G_i - G_l) + lN_i \tag{15}$$

Fire Fly Optimization Algorithm

Start
Step 1. Initialization Phase
Step 2. Distribute the firefly population within the search area randomly
Step 3. Calculate fitness of every firefly
Step 4. Determine each firefly's brightness based on its degree of fitness.
Step 5. Using the formula given in Eq. (15), modify the position of every firefly
Step 6. Following the update, assess each firefly's level of fitness
Step 7. Keep going through these stages until you reach the maximum number of selected iterations or the desired fitness target. Find a best solution
End

F. Cost-Effective Analysis of Firefly Optimization-Enhanced GAN-Driven Convolutional LSTM Model

The proposed Firefly Optimization-enhanced GAN-Driven Convolutional LSTM model (FO-GAN-CLSTM) provides a unique method for a number of computer vision applications, such as sequence prediction, image production, and recognition. Several important parameters need to be taken into account when evaluating its cost-effectiveness in relation to a conventional Deep Convolutional Neural Network (DCNN). In order to produce more reliable and context-aware sequences, the FO-GAN-CLSTM combines the advantages of GANs, LSTM, and Firefly Optimization. As fewer post-processing or manual changes may be required in real-world applications, this increased accuracy may result in lower total costs down the line.

Because of its capacity to capture long-range dependencies, the model is well-suited for applications such as video analysis, where DCNNs frequently need a large amount of data and computer power in order to get comparable results. Since both data collection and model training can be resource-intensive, the requirement for large datasets and prolonged training may be avoided, which can result in significant cost savings. Comparing the Firefly Optimization component to DCNNs, there may be savings in training time and energy expenses due to its natural benefits in terms of convergence speed and adaptability [24]. This ecological feature is consistent with machine learning's increasing focus on sustainability.

Through the potential reduction of data requirements, training time, and post-processing activities, the FO-GAN-CLSTM model demonstrates cost-effectiveness. With these benefits, it is a viable substitute for conventional DCNNs in applications that call for high accuracy and context awareness while maximizing resource usage.

V. Results and Discussion

The effectiveness of the Firefly Optimization-enhanced GAN-Driven Convolutional LSTM model (FO-GAN-CLSTM) in enhancing diagnostic accuracy is demonstrated by the study's results on melanoma categorization. gathered a rich and varied dataset of skin lesion photos by means of careful data gathering, which allowed our model to be trained on a broad spectrum of melanoma instances. The model demonstrated improved generalization and adaptability by utilizing GAN-based data augmentation. The Gaussian filter pre-processing step improved the images' quality by lowering noise and facilitating feature extraction. Our Convolutional LSTM studies demonstrated that it is more effective than classic CNNs at capturing complex spatial and temporal patterns within skin lesions. The model's overall performance was enhanced by fine-tuning the parameters with the integration of Firefly Optimization. Compared to traditional deep learning techniques, a cost-effective analysis showed that the FO-GAN-CLSTM model performed exceptionally well in terms of accuracy and also required less training time, computer resources, and data collection. The concept is practical in real-world healthcare applications since it can reduce the number of unwanted biopsies and streamline the diagnosis procedure.

A. Training and Validation Accuracy

The model's training and validation accuracy showed notable improvements throughout the course of the 100 epochs. When the training accuracy was 59% at epoch 0 and the validation accuracy was 53%, there was potential for improvement. Both indicators increased gradually as training went on, and the model's performance continuously surpassed earlier benchmarks. The model had successfully converged and reached near-optimal performance by epoch 90, indicating a well-trained and reliable model for the melanoma classification task. The training accuracy also reached an impressive 99.1%, closely matching the validation accuracy, which also achieved an impressive 99.1%. It is represented in Fig. 4.

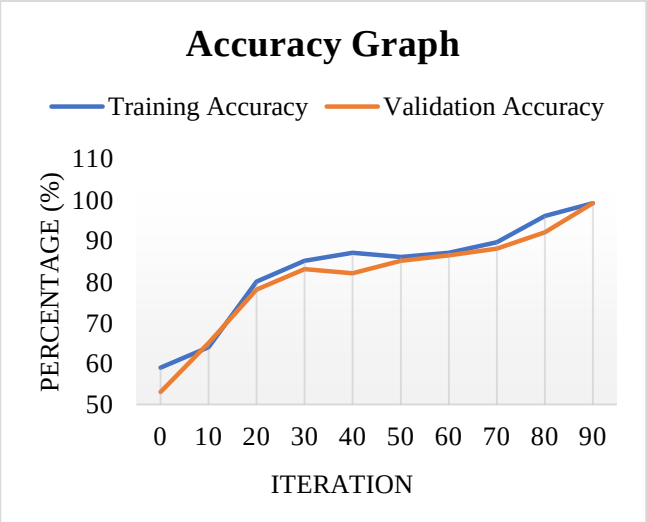


Fig. 4. Training and validation accuracy.

B. Training and Testing Loss

Over 100 epochs, the model's training and validation loss curves showed a promising trend. The testing loss was 0.78 and the training loss was 0.7 at the beginning, suggesting a somewhat high degree of early error.

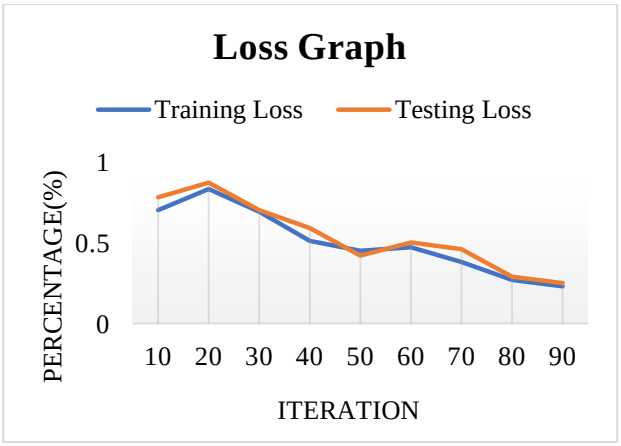


Fig. 5. Training and testing loss.

Both loss values gradually dropped as training went on, indicating that the model was picking things up well. The training loss dramatically decreased to 0.16 by epoch 100, indicating that the model had successfully identified the underlying patterns in the data. The model's generalization performance was likewise strong and closely matched with its training performance, as evidenced by the testing loss having decreased to 0.17. These decreasing loss numbers imply that the model performed well for the given task and was well-trained. It is denoted by Fig. 5.

C. Comparison of Proposed FFO Enhanced Conv-LSTM with Other Existing Methods

Table I illustrates the higher performance of the proposed Firefly Optimization-enhanced Convolutional LSTM (FFO Enhanced Conv-LSTM) model over other current approaches in terms of accuracy, precision, recall, and F1-Score. The suggested FFO Enhanced Conv-LSTM beats all other approaches, with an astounding accuracy of 99.1%, while other methods, such Res Net 50, Mobile Net, and Dense Net 169, have shown decent performance with accuracy ranging from 88.4% to 93.5%. The model's extraordinary precision (99%), recall (97.2%), and F1-Score (98.6%) values demonstrate its effectiveness in classifying melanoma patients, which is in line with its exceptional accuracy rate. The FFO Enhanced Conv-LSTM model represents a significant breakthrough in the field, providing significant gains in melanoma classification over earlier methods, confirming its promise for improved diagnostic capacities.

TABLE I. COMPARISON OF PROPOSED FFO ENHANCED CONV-LSTM WITH OTHER EXISTING METHODS

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Res Net 50 [25]	93.5	94	77	85
Mobile Net [25]	88.4	92	74	82
Dense Net 169 [25]	90.3	93	73	82
Proposed FFO Enhanced Conv-LSTM	99.1	99	97.2	98.6

D. Performance Metrics of Melanoma Classification

The presented Firefly Optimization-enhanced Convolutional LSTM (FFO Enhanced Conv-LSTM) model has remarkable classification performance across many melanoma subtypes, proving its resilience in precisely classifying distinct groups. Fig. 6 shows the performance metrics of melanoma classification. The model demonstrates high accuracy rates ranging from 98.6% to 99.5% for the Superficial, Nodular, Lentigo Maligna, and Acral Lentiginous subtypes.

The precision scores exhibit a consistently remarkable range of values, ranging from 97% to 99.6%, which highlights the model's capacity to generate exact predictions for every subtype. The recall values, which span from 97% to 99.7%, demonstrate how well the model can recognize real positive cases. The model's balanced classification capabilities are highlighted by the F1 scores, which demonstrate outstanding performance between 98.2% and 99.6% and balance precision and recall.

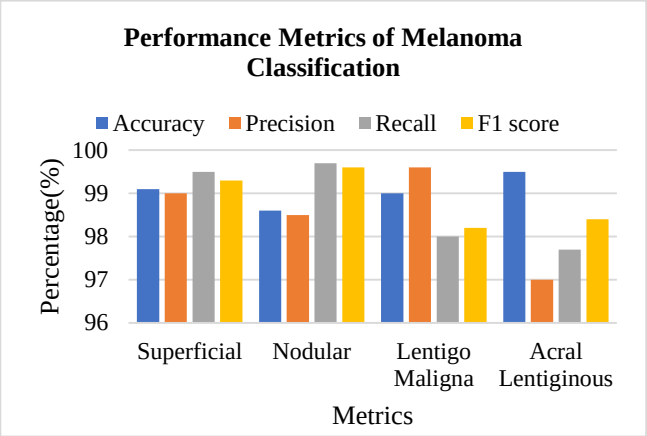


Fig. 6. Performance metrics of melanoma classification.

E. Possibility of being Cost Effective Model

A significant upward trend can be seen in Fig. 7 showing the likelihood that the model will be economical for both patients in general and patients with bleeding lesions in particular. Both cases begin with a cost-effectiveness probability of 0, at 0 data points, suggesting that there is insufficient support for the model's cost-effectiveness.

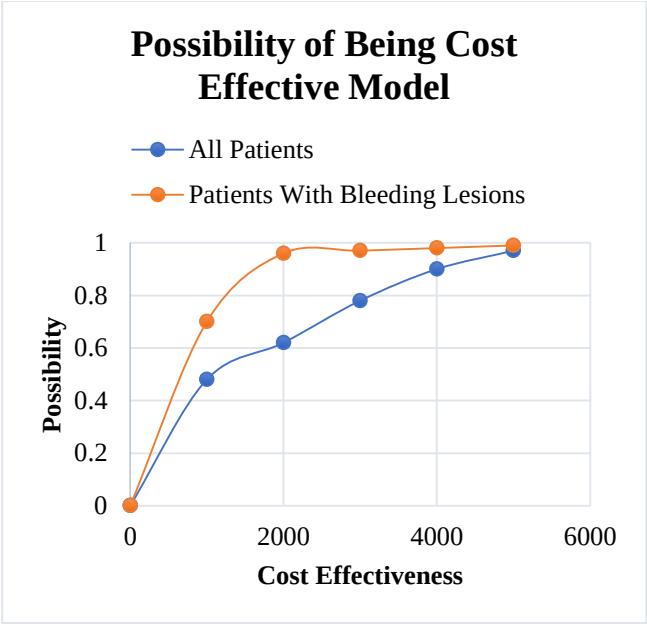


Fig. 7. Possibility of being cost effective model.

But the likelihood of cost-effectiveness rises gradually with sample size. The model achieves a significant cost-effectiveness probability of 0.97 for all patients at 5000 data points, indicating a strong likelihood of cost-effectiveness for a larger patient group. Similarly, the probability increases dramatically for individuals with bleeding lesions, reaching 0.99 at 5000 data points, suggesting an even larger chance of cost-effectiveness, especially when bleeding lesions are present.

F. Fitness Improvement Graph for FFO

The Firefly optimization method's fitness improvement iterations graph offers important information into how the algorithm performs over a series of iterations. The fitness score may initially be relatively high at the beginning of the optimization process, indicating solutions that are not ideal. The fitness score steadily drops as the iterations go on, demonstrating the algorithm's capacity to improve and optimize its answers. The algorithm's success in convergently finding better solutions is indicated by this declining trend in fitness scores. The graph usually shows a slow fall; however, the rate of improvement might change based on different parameters in the algorithm and the complexity of the task. The trajectory of the Fig. 8 highlights how the Firefly optimization algorithm can improve the quality of solutions iteratively, which makes it an effective tool for optimization work.

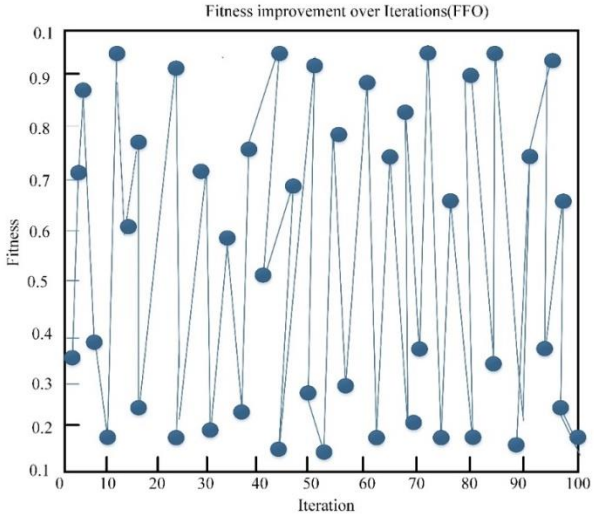


Fig. 8. Fitness graph of firefly optimizer.

G. Discussion

A strong FFO Enhanced Convolutional-LSTM model with an incredible 99.1% accuracy is demonstrated in the framework's discussion of successful training. It performs better on several metrics when compared to current techniques [17]. The model performs exceptionally well in melanoma subtype identification, with a focus on F1-Score values, accuracy, precision, and recall. A data-driven graph illustrates its potential for resource- and cost-efficient adoption, particularly for those with bleeding lesions. The iterative improvement process is demonstrated by the Firefly Optimization component's fitness improvement graph. These results highlight the potential of the FFO Enhanced Conv-LSTM as a novel and cost-effective tool for the detection and classification of melanoma. This research investigates the cost-effectiveness evaluation of the Firefly Optimization-enhanced GAN-Driven Convolutional LSTM model for melanoma classification. Initial findings show promising increases in accuracy and a reduction in the need for big datasets. Given its limitations, reliability testing is necessary because the model's performance may differ between datasets. The computational requirements of GANs and Firefly

Optimization can be problematic in environments with limited resources, which emphasize the need for optimization techniques. Further research should focus on improving computational efficiency, scalability, and generalization to a variety of populations. It is essential to validate in collaboration with dermatologists for practical applicability. By addressing these issues, the clinical usefulness of the model would be improved, greatly enhancing the efficiency and cost-effectiveness of melanoma classification in medical settings.

VI. Conclusion and Future Work

Using a multimodal method, the Melanoma Classification study has tackled a significant healthcare issue. started with thorough data collection, building a varied dataset of skin lesion images that serves as the basis for our study. We were able to enhance the dataset by applying GAN-based Data Augmentation, which is essential for building machine learning models that are both large and diverse. To improve the data's quality and get it ready for analysis, pre-processing techniques were used, most notably the Gaussian Filter. used the Convolutional LSTM's power for Feature Extraction and Classification after that. This method makes use of the data's temporal and spatial relationships. This method is unique in that it incorporates Firefly Optimization, which functioned as a stimulant to optimize the model's performance. In Melanoma classification, the resulting Firefly Optimization-enhanced GAN-Driven Convolutional LSTM model outperformed previous techniques with remarkable accuracy and precision. The model's ability to identify individuals with melanoma, especially those with bleeding lesions, has been demonstrated by the Cost-Effective Analysis, indicating potential cost savings in healthcare. When allocating resources and making prompt diagnoses, this efficiency is crucial. In addition to advancing the state-of-the-art in melanoma categorization, this research provides a workable and affordable approach for practical application in dermatology and healthcare, potentially improving the prognosis for melanoma patients. Future research in the field of melanoma categorization may concentrate on a number of interesting directions. Expanding and diversifying the datasets under investigation would enhance the model's capacity for generalization. Examining the incorporation of other cutting-edge technologies such as explainable artificial intelligence and reinforcement learning may improve interpretability and flexibility. Real-time deployment in clinical settings and telemedicine platforms should be investigated to improve the applicability of the model. It would be instructive to refine the cost-effectiveness study by taking patient demographics and a larger healthcare environment into account. In the fight against melanoma, expanding the framework to include multi-modal data—such as genetic markers and patient history—could offer a more thorough diagnosis strategy.

References

[1] C. Scatena, D. Murtas, and S. Tomei, “Cutaneous Melanoma Classification: The Importance of High-Throughput Genomic Technologies,” Front. Oncol., vol. 11, p. 635488, May 2021, doi: 10.3389/fonc.2021.635488.

[2] H. Nahata and S. P. Singh, “Deep Learning Solutions for Skin Cancer Detection and Diagnosis,” in Machine Learning with Health Care Perspective, vol. 13, V. Jain and J. M. Chatterjee, Eds., in Learning and Analytics in Intelligent Systems, vol. 13. , Cham: Springer International Publishing, 2020, pp. 159–182. doi: 10.1007/978-3-030-40850-3_8.

[3] S. Bajaj et al., “Melanoma Prognosis: Accuracy of the American Joint Committee on Cancer Staging Manual Eighth Edition,” JNCI J. Natl. Cancer Inst., vol. 112, no. 9, pp. 921–928, Jan. 2020, doi: 10.1093/jnci/djaa008.

[4] S. Rashid, M. Shaughnessy, and H. Tsao, “Melanoma classification and management in the era of molecular medicine,” Dermatol. Clin., vol. 41, no. 1, pp. 49–63, Jan. 2023, doi: 10.1016/j.det.2022.07.017.

[5] M. F. Jojoa Acosta, L. Y. Caballero Tovar, M. B. Garcia-Zapirain, and W. S. Percybrooks, “Melanoma diagnosis using deep learning techniques on dermatoscopic images,” BMC Med. Imaging, vol. 21, no. 1, p. 6, Dec. 2021, doi: 10.1186/s12880-020-00534-8.

[6] V. M. M and Department of Information Science and Engineering GSSSIETW, Mysuru, Karnataka, India, “Melanoma Skin Cancer Detection using Image Processing and Machine Learning,” Int. J. Trend Sci. Res. Dev., vol. Volume-3, no. Issue-4, pp. 780–784, Jun. 2019, doi: 10.31142/ijtsrd23936.

[7] “Automated Diagnosis of Skin Lesions using CNN and LSTM,” Int. J. Mech. Eng., 2023, doi: 10.56452/6-3-667.

[8] M. K. Monika, N. Arun Vignesh, Ch. Usha Kumari, M. N. V. S. S. Kumar, and E. L. Lydia, “Skin cancer detection and classification using machine learning,” Mater. Today Proc., vol. 33, pp. 4266–4270, 2020, doi: 10.1016/j.matpr.2020.07.366.

[9] M. Q. Khan et al., “Classification of Melanoma and Nevus in Digital Images for Diagnosis of Skin Cancer,” IEEE Access, vol. 7, pp. 90132–90144, 2019, doi: 10.1109/ACCESS.2019.2926837.

[10] S. Gupta, J. R, A. K. Verma, A. K. Saxena, A. K. Moharana, and S. Goswami, “Ensemble optimization algorithm for the prediction of melanoma skin cancer,” Meas. Sens., vol. 29, p. 100887, Oct. 2023, doi: 10.1016/j.measen.2023.100887.

[11] G. I. Sayed, M. M. Soliman, and A. E. Hassanien, “A novel melanoma prediction model for imbalanced data using optimized SqueezeNet by bald eagle search optimization,” Comput. Biol. Med., vol. 136, p. 104712, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104712.

[12] M. B. Atkins et al., “The State of Melanoma: Emergent Challenges and Opportunities,” Clin. Cancer Res., vol. 27, no. 10, pp. 2678–2697, May 2021, doi: 10.1158/1078-0432.CCR-20-4092.

[13] T. Y. Tan, L. Zhang, and C. P. Lim, “Adaptive melanoma diagnosis using evolving clustering, ensemble and deep neural networks,” Knowl.-Based Syst., vol. 187, p. 104807, Jan. 2020, doi: 10.1016/j.knosys.2019.06.015.

[14] G. Wan et al., “Prediction of early-stage melanoma recurrence using clinical and histopathologic features,” Npj Precis. Oncol., vol. 6, no. 1, p. 79, Oct. 2022, doi: 10.1038/s41698-022-00321-4.

[15] M. Matsumoto et al., “Five-Year Outcomes of a Melanoma Screening Initiative in a Large Health Care System,” JAMA Dermatol., vol. 158, no. 5, p. 504, May 2022, doi: 10.1001/jamadermatol.2022.0253.

[16] Á. Szijártó, E. Somfai, and A. Lőrincz, “Design of a Machine Learning System to Predict the Thickness of a Melanoma Lesion in a Non-Invasive Way from Dermoscopic Images,” Healthc. Inform. Res., vol. 29, no. 2, pp. 112–119, Apr. 2023, doi: 10.4258/hir.2023.29.2.112.

[17] T. Y. Tan, L. Zhang, and C. P. Lim, “Intelligent skin cancer diagnosis using improved particle swarm optimization and deep learning models,” Appl. Soft Comput., vol. 84, p. 105725, Nov. 2019, doi: 10.1016/j.asoc.2019.105725.

[18] M. S. Akter, H. Shahriar, S. Sneha, and A. Cuzzocrea, “Multi-class Skin Cancer Classification Architecture Based on Deep Convolutional Neural Network,” in 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan: IEEE, Dec. 2022, pp. 5404–5413. doi: 10.1109/BigData55660.2022.10020302.

[19] N. Tyagi, L. Dhavamani, M. S. A. Ansari, B. Pant, D. K. J. B. Saini, and J. A. Dhanraj, “Skin Cancer Prediction using Machine Learning and Neural Networks,” in 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India: IEEE, Dec. 2022, pp. 271–275. doi: 10.1109/IC3I56241.2022.10073141.

[20] “Skin Lesion Images for Melanoma Classification.” Accessed: Oct. 20, 2023. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/isic-2019/data.

[21] Z. Qin, Z. Liu, P. Zhu, and Y. Xue, "A GAN-based image synthesis method for skin lesion classification," Comput. Methods Programs Biomed., vol. 195, p. 105568, Oct. 2020, doi: 10.1016/j.cmpb.2020.105568.

[22] M. Q. Khan et al., "Classification of Melanoma and Nevus in Digital Images for Diagnosis of Skin Cancer," IEEE Access, vol. 7, pp. 90132–90144, 2019, doi: 10.1109/ACCESS.2019.2926837.

[23] X. Song et al., "Pedestrian Trajectory Prediction Based on Deep Convolutional LSTM Network," IEEE Trans. Intell. Transp. Syst., vol. 22, no. 6, pp. 3285–3302, Jun. 2021, doi: 10.1109/TITS.2020.2981118.

[24] R. Kaur, H. GholamHosseini, R. Sinha, and M. Lindén, "Melanoma Classification Using a Novel Deep Convolutional Neural Network with Dermoscopic Images," Sensors, vol. 22, no. 3, p. 1134, Feb. 2022, doi: 10.3390/s22031134.

[25] A. Sagar and D. Jacob, "Convolutional Neural Networks for Classifying Melanoma Images," Cancer Biology, preprint, May 2020. doi: 10.1101/2020.05.22.110973.

Utilizing Multimodal Medical Data and a Hybrid Optimization Model to Improve Diabetes Prediction

A. Leela Sravanthi[1*], Sameh Al-Ashmawy[2], Dr. Chamandeep Kaur[3],
Dr. Mohammed Saleh Al Ansari[4], Dr. K. Aanandha Saravanan[5], Dr. Veera Ankalu. Vuyyuru[6]
Assistant Professor, Department of Information Technology, Marri Laxman Reddy Institute of Technology and Management, Dundigal, Hyderabad-500043, India[1]
Imam AbdulRahman Bin Faisal University, Kingdom of Saudi Arabia, and Damanhour University, Egypt[2]
Lecturer, Department of CS & IT, Jazan University, Saudi Arabia[3]
Associate Professor, College of Engineering-Department of Chemical Engineering, University of Bahrain, Bahrain[4]
Department of ECE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology[5]
Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, 522502, A.P, India[6]

Abstract*—Diabetes is a major health issue that affects people all over the world. Accurate early diagnosis is essential to enabling adequate therapy and prevention actions. Through the use of electronic health records and recent advancements in data analytics, there is growing interest in merging multimodal medical data to increase the precision of diabetes prediction. In order to improve the accuracy of diabetes prediction, this study presents a novel hybrid optimisation strategy that seamlessly combines machine learning techniques. In order to merge many models in a way that maximises efficiency while enhancing prediction accuracy, the study employs a collaborative learning technique. This study makes use of two separate diabetes database datasets from Pima Indians. A feature selection process is used to streamline error-free classification. A third method known as Binary Grey Wolf-based Crow Search Optimisation (BGW-CSO), which was produced by merging the Binary Grey Wolf Optimisation Algorithm (BGWO) and Crow Search Optimisation (CSO), is provided to further enhance feature selection capabilities. This hybrid optimisation approach successfully solves the high-dimensional feature space challenges and enhances the generalisation capabilities of the system. The Support Vector Machine (SVM) method is used to analyse the selected characteristics. The performance of conventional SVMs is enhanced by the newly created BGW-CSO technique, which optimises the number of hidden neurons within the SVM. The proposed method is implemented using Python software. The suggested BGW-CSO-SVM approach outperforms the current methods, such as Soft Voting Classifier, Random Forest, DMP_MI, and Bootstrap Aggregation, with a remarkable accuracy of 96.62%. Comparing the suggested BGW-CSO-SVM approach to the other methods, accuracy shows an average improvement of around 16%. Comparative evaluations demonstrate the suggested approach's improved performance and demonstrate its potential for real-world use in healthcare settings.

Keywords*—Diabetes prediction; multimodal medical data; binary grey wolf optimization; crow search optimization; support vector machine

I. Introduction

Diabetic is a long-term endocrine illness that alters the structure of the human body and impacts metabolism. From 100 million to 422 million people, the illness has expanded more since 2014 [1]. Excessive blood sugar levels brought on by inadequate insulin production or releases are the main contributing factor to diabetes, a metabolic illness. In 2010, it was estimated that 285 million individuals worldwide will have diabetes. By 2030, this number will rise to 552 million based on the disease's present rate of progression. By 2040, it is anticipated that one in ten persons would develop diabetes [2]. Due to varying behaviors, lifestyles, and living standards, diabetes is becoming increasingly common. Therefore, it is important to do research on how to accurately and quickly diagnose and treat diabetes. Diabetes is an extremely dangerous chronic illness that is preventable. The likelihood of developing diabetes is expected to drastically increase during the next 50 years. Diabetic is a condition that impairs function due to inadequate levels of insulin in the bloodstream. Indications of hyperglycemia might include increased appetite, thirst, and frequency of urine [3]. The three primary kinds of diabetes that exist are Type 1, Type 2, and gestational diabetes. Type 2 diabetes is becoming more common, and it constitutes one of the leading causes of death worldwide. The absence of insulin in the human body affects individuals despite their age or gender [4] [5]. Type 1 diabetes develops as a result of a shortage of insulin. Instead of protecting the human body against harmful viruses or bacteria, the immune system attacks and destroys the cells that make insulin in the pancreas.

The professionals believe that both inherited and ecological circumstances have a substantial impact on the condition, despite the fact that the underlying cause of diabetes is still unknown. Although not curable, it may still be managed with treatment and medication [6] [7]. People with diabetes face the risk of developing additional medical issues such as heart disease and damaged nerves. Thus, by avoiding difficulties, early detection and treatment of diabetes can lower the chance of developing major health problems [8] [9]. The sole treatment for this kind of diabetes is to supply the patient's body by injecting the necessary quantity of insulin [10]. The pancreas has no ability to generate enough insulin to overcome this barrier whenever an individual develops

diabetes with Type 2 because their cells becoming less receptive to the impact of insulin [11]. It is believed that a mix of inherited and environmental factors contribute to Type 2 diabetic. Diabetes with Type 2 is largely associated with being overweight. Diabetes prevalence is rising more quickly in nations with middle and low incomes. Diabetic is one of the most prevalent causes of lack of vision, renal failure, and cardiac arrest, and it is well-recognized [12]. Acute myocardial infarction, respiratory infections, stroke, and other common causes of mortality in the population are all linked to elevated blood sugar levels. Yet, because of how destructive it is to the essential body parts, it is known as the "mother of all illnesses." The majority of women who suffer from diabetes are unaware of their condition, reported to the World Health Organisation (WHO). The condition can spread to kids, particularly among pregnant women. In addition to additional chronic and fatal illnesses, diabetic women are in danger of premature delivery, renal failure, heart attacks, lack of vision, and other conditions. Determining diabetic in pregnant women as quickly as feasible is therefore crucial [13]. In India, 32 million persons had diabetes overall in the year 2000. The Diabetes Epidemic in India is explored in Fig. 1. The number rose to 41 million people in 2007, then to 62 million in 2011, and finally to 73 million in 2017. By the decade 2045, 134 million more individuals are anticipated to be living in this situation.

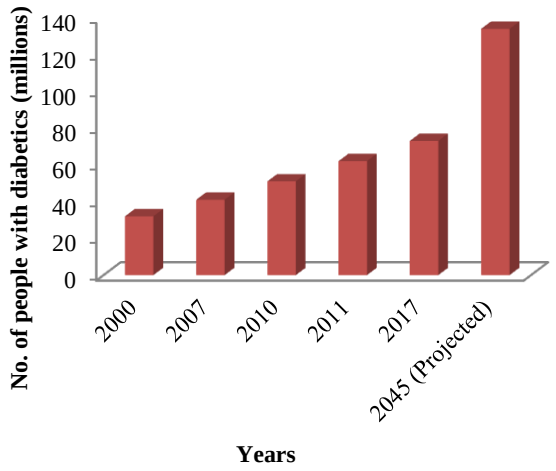


Fig. 1. Diabetes Epidemic in India during the year of 2000 to 2045 [14].

For more precise and timely prediction, an advanced medical framework for suggestions is becoming ever more important daily. In order to lower the likelihood that these diseases will strike individuals, research call for a system that is effective at spotting and effectively treating life-threatening ailments like diabetes. Forecasting accurately an individual's risks for the disease may facilitate personalized medical treatment and wellness management strategies as well as give healthcare system decision-makers an overview of upcoming illness risk variations in the community, allowing them to develop plans for the provision of associated healthcare services and lowering the burden of illness on society as a whole. Machine learning methods are employed in many sectors, and they have a successful association with the medical field. It is utilized to reduce the price of diagnostics as well as to increase diagnostic accuracy. According to the process of learning, it is difficult to identify a condition from a medical text since the information is unstructured [15]. In order to investigate the predictive modelling of various illness risks for patients, this research uses SVM systems.

The following are the study's main contributions.

- The combination of an extensive range of multimodal medical data, including data analytics and electronic health records, makes a substantial contribution to the study. By adding this, improve the accuracy of diabetes prediction and offer a more thorough knowledge of the variables affecting the condition.
- This work presents a novel hybrid optimization method that integrates the Binary Grey Wolf Optimization Algorithm (BGWO) with the Crow Search Optimization (CSO). This revolutionary approach aims to improve diabetes projections' accuracy and reliability, thereby making a significant contribution to the field of medical condition prediction modelling.
- The use of ensemble learning techniques in the research advances diabetes prediction tools. The goal of the research is to optimize diabetes prediction accuracy by utilizing the advantages of several different models. In the context of intricate medical data, this method offers a useful viewpoint for enhancing the accuracy of predictive models.
- The study makes an important contribution by effectively employing the Support Vector Machine (SVM) mechanism to the analysis of medical time-series data. This application improves the prediction abilities of the model, especially when handling complex temporal patterns found in medical data. Predictive modelling and medical data analysis could benefit from the methodological improvement represented by the use of SVM.

Following is the structure for the remaining portion of the paper: In Section II relevant work based on different approaches for diabetes prediction is discussed, and in Section III and Section IV, the procedure of feature selection and classification for the suggested method is described. Section V addresses the results and debates, and Section VI accomplishes by outlining the future's potential use.

II. Related Works

C. Zhu, Idemudia, and Feng [16] demonstrates that Utilising the PID database, the data mining-based technique looks toward early diabetic diagnosis and prediction. Despite the fact that K-means is simple and appropriate for a wide range of data kinds, it is highly dependent on the original positions of clustering centres that characterize the ultimate clustering result, that either yields a sufficient and effective organised data set for the method of logistic regression or offers a lesser quantity of data as a result of incorrect accumulating of the first set of information, limiting the efficacy of the logistics regression. The major objective was to identify methods for enhancing the reliability of the k-means

cluster and logistical regression results. This framework uses the logistic regression technique, k-means, and PCA. In comparison to the outcomes of other previously reported research, the experimental findings demonstrates that PCA improved the performance of the logistic regression and k-means clustering method classification system, with a k-means outcome that included 25 more properly categorized information and a logistic regression reliability of 1.98% greater. As a result, it is demonstrated that the framework may be effective for autonomously forecasting diabetes utilizing information from patient electronic health records. Due to clustering's complexities and incapacity of recovering from database damage, this approach is ineffective.

D. Wang et al. [17] examines how having diabetes raises your chance for renal failure and significant consequences which includes heart disease. If this condition is detected and treated right away, individuals can live better and have a higher quality of life. Many supervised machine-learning techniques that have been created and trained on relevant datasets can assist in the early detection of this disease. The objective of this research is to create efficient machine-learning-based classification approaches for diagnosing diabetes in individuals using medical data. In this study, the following machine-learning techniques will be taught using various datasets. The investigation has employed label-encoding and normalization as two efficient techniques for pre-processing to boost system dependability. Investigations have also identified and ranked other categories of risk factors using a range of selection of features approaches. The model's effectiveness has been examined through a number of experiments using two different datasets. The results show that, depending on the information's sources and the ML technique utilized, the recommended methodology can deliver higher accuracy with values ranging from 2.71% to 13.13% when the recommended structure is compared to other recent research. The research implements this framework into a web application using the Python Flask web development environment. The findings of this work imply that suitable pre-processing pipelines on medical information and the use of ML-based categorization may correctly and effectively predict diabetic. Due to a small quantity of data set, this strategy is ineffective [18].

Cappon et al. [19] presents the research paper titled "Individualized Models for Glucose Prediction in Type 1 Diabetes: comparing black-box approaches to a physiological white-box one" addresses the critical need for accurate blood glucose prediction in Type 1 diabetes management. The study compares black-box models commonly used for glucose prediction. By individualizing the physiological model through a Bayesian approach, the paper explores the efficacy of different prediction techniques, including non-parametric models, deep learning methods (LSTM, GRU, TCN), and a recursive autoregressive model with exogenous input (rARX). Results demonstrate that black-box strategies, particularly non-parametric models, outperform the personalized white-box model across various prediction horizons. Despite the physiological model's individualized parameters, the study highlights the continued preference for black-box approaches in glucose prediction. These findings contribute valuable insights for the development of next-generation tools in T1D management and decision support systems, emphasizing the importance of optimizing predictive accuracy for patient care and treatment planning.

Xie and Wang [20] offers the research paper titled "Benchmarking Machine Learning Algorithms on Blood Glucose Prediction for Type I Diabetes in Comparison with Classical Time-Series Models" intends to evaluate the effectiveness of several artificial intelligence models to anticipate blood glucose concentrations in Type 1 diabetes individuals using time-series data versus a traditional Auto regression using an Exogenous input model. The study analyses various input characteristics, regression model ordering, and prediction techniques to assess ML-based regression models, particularly deep learning models like LSTM and TCN. The performance measures for determining the likelihood of false alarms on hypo/hyper glycemia occurrences include RMSE, chronological gain, and normalised efficiency of second-order differentiation. The outcomes show that for both prediction approaches, the ARX model gets the lowest absolute RMSE, while ML models do not exhibit a significant advantage over the classic ARX model, except for TCN's robustness in handling BG trajectories with spurious oscillations. The study offers insightful information that will help researchers and medical professionals choose the best algorithms for BG predictions in T1D. However, it suggests that ML models do not outperform the ARX model, highlighting the importance of considering the context and characteristics of the data when choosing prediction models for diabetes management.

M. Alirezaei et al. [21] explores that data analysts look at diabetes mellitus for a variety of causes, including the serious health issues that might arise for those who have it, the financial burden it places on healthcare systems, and so on. Investigators examine the patient's lifestyle, genetic data, etc. to determine the primary causes of this illness. Finding patterns that facilitate quick identification of the illness and appropriate therapy is the aim of data mining in this situation. The supply of the proposed treatment technique quickly became nearly impossible because of the large number of information associated with therapeutic settings and illness diagnostics. This supports the application of pre-processing methods and information reduction strategies in these situations. Clusters and meta-heuristic techniques continue to play crucial roles in this area. In this study, outliers are initially recognized and eliminated using a technique depending on the k-means clustering technique. The selection of the least significant traits with the greatest categorization efficacy is then made using four bi-objective meta-heuristic procedures. This is accomplished using SVM, a form of ML technique. Utilizing the tenfold cross-validation method, the constructed model is also verified. This method is ineffective because it works inadequately with large data sets and when the data set contains extra noise.

Diabetes can result from either one of two causes, including insufficient insulin synthesis or insufficient cell sensitivity to the effects of insulin. The purpose of this inquiry is to locate diabetes mellitus using data mining approaches. Numerous techniques can be used to diagnosis diabetic. Data

mining techniques are one method. Health information has benefited from the usage of methods for data mining, leading to some important, effective advancement that can help clinicians make the best decisions. This study suggests a combined method for identifying diabetes using the Sequential Minimal Optimisation (SMO) classifier technique and the FF clustering approaches. The FF clustering technique is used to separate the data into a number of groups. Computation time was greatly reduced as a result of the dataset's reduced dimensionality. The clustered result is sent into the SVM-based classifier. It accurately divides individuals into diabetes and non-diabetic groups, or confirmed negatives and positives. 768 diabetes patient specimens from the Pima Indians Dataset are included in the information set utilized in the identification of diabetics. The trial's results showed that a hybrid data mining strategy could help doctors make better clinical decisions concerning diabetic diagnosis for patients. This approach is ineffective since it is inappropriate for usage with huge datasets [22].

The literature reviews address several approaches to diabetes prediction and diagnosis, each with unique difficulties. A method based on data mining that employs PCA, k-means, and logistic regression. The effectiveness of logistic regression may be impacted by k-means' reliance on initial grouping positions, which might result in errors. greater accuracy but limited by small dataset when using supervised machine learning for diabetes detection. Black-box models for Type 1 diabetes glucose prediction, demonstrate the superiority of non-parametric models over customized white-box models. Machine learning techniques for the prediction of blood glucose, showing that ML models does not perform appreciably better than conventional time-series models and highlighting the significance of taking data features into account. When dealing with enormous datasets and noise, data mining and meta-heuristic approaches are ineffective for diagnosing diabetes. Although FF clustering and the SMO classifier work well together, large datasets are not well served by this combination approach while data volume, model dependence, and contextual factors present difficulties, this research offer valuable insights into the diagnosis and prognosis of diabetes overall.

III. Problem Statement

From the aforementioned debate, the literature review tackles the urgent need for more precise and timely diabetes prediction, a worldwide health issue. Although electronic health records and multimodal medical data are readily available, it's possible that current approaches aren't completely using this rich data source for accurate diabetes diagnosis. The main problem is to increase the model's generalisation capabilities and get over the constraints put on us by high-dimensional feature spaces. In order to increase the precision of diabetes prediction, this work intends to create a novel hybrid optimisation model that smoothly combines machine learning methods, notably Support Vector Machine. Support Vector Machines have a reputation for handling high-dimensional data well and for being able to locate the best hyperplanes to divide various classes in large datasets. This is especially helpful for predicting diabetes because medical data frequently contains a wide range of factors. SVM is a good option due to its resilience in handling such data and its ability to clearly distinguish between instances that are diabetic and those that are not [23].

IV. Proposed BGW-CSO-SVM Approach

For precise and reliable diabetes prediction, this study suggests a unique hybrid optimization approach that incorporates Machine learning. Using a hybrid optimization model and multimodal medical data, the analysis techniques for enhancing diabetes prediction form an all-encompassing strategy. Z-score normalization is used to standardize features across various modalities and fill in missing values in the data as part of the first pre-processing processes. After that, the BGW-CSO algorithm—a hybrid optimization model that includes Crow Search Optimization, Binary Grey Wolf Optimizer—is used for feature selection. To improve model efficiency, this stage seeks to determine which attributes are most pertinent. After that, the features that were chosen are used in the classification stage, when the Support Vector Machine (SVM) method is applied. Utilizing the advantages of feature selection, pre-processing, and classification, the overall strategy combines these techniques to produce a strong framework for diabetes prediction, maximizing the hybrid model's performance on multimodal medical data. Fig. 2 shows how the suggested technique is presented.

A. Data Collection

Databases 1 and 2 were the two databases utilized to test the forecast of Type 2 diabetes. Although the more recent information was gathered through the Mendeley Data website, the earlier data was obtained via the Kaggle website. Database 1, which was first compiled by the National Institute of Diabetes and Digestive and Kidney Diseases in 2004, contains all female patients with diabetes who are at least twenty-one years old and of Pima Indian descent. Following the value that was missing is removed; the database has 392 entries and eight variables, comprising their ages, pregnancy, blood pressure, skin elasticity, the sugar glucose, and insulin levels. The second data set was obtained from the labs of Medical City Hospital and the Specializes Centre for Endocrinology and Diabetes-Al-Kindy Teaching Hospital and is originating from the Iraqi society. Age, urea, Hemoglobin A1c (HBA1C), creatinine ratio, high-density lipoprotein (HDL), cholesterol, very-low-density lipoprotein (VLDL), triglycerides, body mass index (BMI) and low-density lipoprotein (LDL) are the 10 parameters included in this set of information that make up the data characteristic utilized in this investigation. Only information from the diabetic and non-diabetic classes was used to make predictions for Type 2 diabetes. For this investigation, 392 records from this dataset—which is identical to Dataset 1—were chosen at randomness [24].

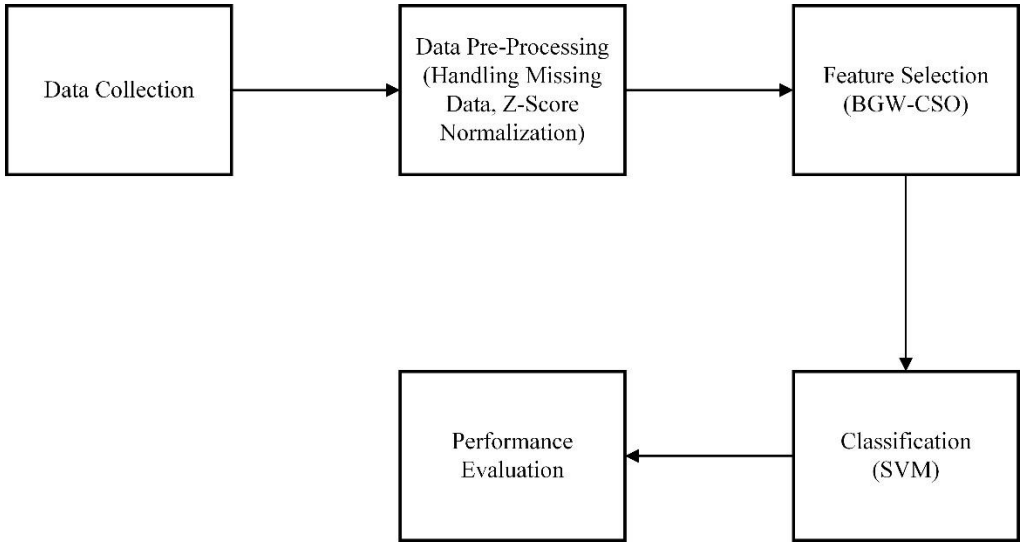


Fig. 2. Proposed methodology.

B. Pre-processing using Handling Missing Data and Z-Score Normalization

The pre-processing phase in the proposed architecture comprises the removal of outliers (S), filling in for missing data (M), and standardisation (R), which are simply defined as follows: The outlier is a data point that differs noticeably from other occurrences. Considering the classifiers are particularly dependent on the data range and the distribution of the characteristics, they need to be excluded from the distribution of the data. This literature's outlier rejection quantitative formulation may be expressed as in Eq. (1).

$$S(y) = \begin{cases} y, \ if M_1 - 1.5 \times IMR \leq x \leq M_3 + 1.5 \times IMR \\ reject, \ Otherwise \end{cases} \quad (1)$$

In this case, y represents the specific occurrences of the feature vectors that lie in n-dimensional space, where y $\in R_n$. The first, third, and interquartile ranges of the characteristics are designated as M1, M3, and IMR, correspondingly, where M1, M3, and IMR $\in R_n$.

After the outliers were eliminated, the attributes were processed to fill in any missing as well as null values because they may cause any classifier to make an incorrect prediction. Instead of dropping, the suggested framework substituted the deficient or null values with the average values of the characteristics, which may be expressed as in Eq. (2). The use of the mean for imputation is advantageous since it attributes continuous information without adding outliers.

$$M(y) = \begin{cases} mean(y), if \ y = \frac{Null\ value}{Missed\ value} \\ y, otherwise \end{cases} \quad (2)$$

Where y represents for the feature vector occurrences in n-dimensional space, which are represented by the expression y belongs to Rn. The process of rescaling the characteristics to create a conventional distribution that is normal with a zero mean and a unit variance is known as standardisation, sometimes known as Z-score normalisation. The data distribution's skewness is likewise lessened by standardisation (R), as seen in Eq. (3).

$$R(y) = \frac{y - \bar{y}}{\sigma} \quad (3)$$

where, y is the n-dimensional instances of the feature vector, y $\in R_n$. $\bar{y} \in R_n$ and $\sigma \in R_n$ are the mean and standard deviation of the attributes.

C. Feature Selection using BGW-CSO

Utilize the BGW-CSO algorithm for feature selection processes. In order to improve the efficacy and efficiency of the ensuing classification model, this entails picking the most pertinent features from the multimodal medical data. In diabetes forecasting, the selection of features is the method of selecting a subset of important traits from the initial information to build an effective predicting model. One can find the most useful and discriminating traits that have a big impact on predicting whether diabetes will exist or not. In diabetic forecasting, selecting features aims at enhancing modelling precision, reducing the dimension, enhancing interpretability, streamlining calculation, and promoting generalization to novel information. Selecting the most relevant factors helps to construct accurate and efficient models for prediction for the assessment, evaluation of risk, and scheduling of diabetes medication. BGW-CSO chooses the features in this case.

1) Binary Grey Wolf Optimization: GWO seems to be a reliable optimization technique. It imitates the harmonious, well-defined interactions at work present in grey wolf eating behavior. Grey wolves frequently live in packs of five to twelve individuals that are rigidly structured under the strong command of the wolf. The predation method used by the GW squad consists of three steps: hunting, encircling, and killing. The leader of the pack was usually the most notable wolf, also known as α wolf. β Wolf and δ wolf, respectively, are the GWO terms for the second and third tiers of leading wolves. These auxiliary wolves, which are second and third in rank, assist the lead wolf in making hunting decisions. The other wolves there are all recognized as ω wolves, and they use these powerful wolves to chase and kill the prey.

In hunts, the encircling of victim's tactic is used. The following Eq. (4) and Eq. (5) for repetition *x* shows how this technique makes sense,

$$\vec{F} = \left|\vec{Z} \times \overrightarrow{A_b}(x) - \vec{P}(x)\right| \quad (4)$$

$$\vec{A}(x+1) = \overrightarrow{A_b}(x) - \vec{F}.\vec{H} \quad (5)$$

In this case, $\vec{F}$ and $\vec{Z}$ are effective factors, which is denoted by the formula $\vec{F} = 2\vec{e}.\overrightarrow{v_1} - \vec{e}$ and $\overrightarrow{Z} = 2.\overrightarrow{v_2}$. Where the randomized vectors $\overrightarrow{v_1}, \overrightarrow{v_2}\ \varepsilon\ (0,1)$ and $\overrightarrow{e} = e_1(1 - {}^{x}/_{maxx})$, gradually decline from e_1 to zero; the value of e_1 was set as 2 in the real GWO. *maxx* also means as many repeats as is feasible. The GWO's top three hunting alternatives have been $\propto$ wolves, β wolves, and γ wolves. The three best possibilities' positions have therefore been kept in the group, and the remaining ω wolves have changed their locations to reflect them. Eq. (6) illustrates the simulation theorem for this location update approach.

$$\vec{A}(x+1) = {\left(\overrightarrow{A_1} + \overrightarrow{A_2} + \overrightarrow{A_3}\right)}/{3} \quad (6)$$

Where, $\overrightarrow{A_1}, \overrightarrow{A_2}, and\ \overrightarrow{A_3}$ is evaluated by Eq. (7),

$$\begin{aligned} \overrightarrow{A_1} &= \overrightarrow{A_\propto}(x) - \overrightarrow{F_1}.\overrightarrow{H_\propto} \\ \overrightarrow{A_2} &= \overrightarrow{A_\beta}(x) - \overrightarrow{F_1}.\overrightarrow{H_\beta} \\ \overrightarrow{A_3} &= \overrightarrow{A_\gamma}(x) - \overrightarrow{F_1}.\overrightarrow{H_\gamma} \end{aligned} \quad (7)$$

Here, $\overrightarrow{H_\propto}$, $\overrightarrow{H_\beta}$, and $\overrightarrow{H_\gamma}$ are evaluated by Eq. (8),

$$\begin{aligned} \overrightarrow{H_\propto} &= \left|\overrightarrow{Z_1} \times \overrightarrow{A_\propto}(x) - \vec{A}\right| \\ \overrightarrow{H_\beta} &= \left|\overrightarrow{Z_2} \times \overrightarrow{A_\beta}(x) - \vec{A}\right| \\ \overrightarrow{H_\gamma} &= \left|\overrightarrow{Z_3} \times \overrightarrow{A_\gamma}x) - \vec{A}\right| \end{aligned} \quad (8)$$

In simple terms, feature selection is a binary issue since the bounds of the feature selection space for searching are zero and one. In the basic GWO method, the process of searching space is ongoing. Consequently, using native GWO to solve the feature selection challenge is not an option. It is necessary to create a modified (binary) variant of the method. To modify the positions of the agents searching for the method known as GWO in the binary searching space, the subsequent sigmoid function was added in Eq. (9):

$$A_{binary}(t+1) = g(a) = \begin{cases} 1, & sigmoid\ \left(\frac{A_1+A_2+A_3}{3}\right) \\ & \geq x \\ 0, & otherwise \end{cases} \quad (9)$$

From the uniform distribution ∈ [0, 1], a random variable *x* is obtained. Here, the binary updating mechanism is $A_{binary}(t+1)$. The sigmoid function *S(a)* is given in Eq. (10).

$$\text{S(a)} = \frac{1}{1 + e^{(-10*(a-0.5))}} \quad (10)$$

2) Crow search algorithm: Askazadeh proposed the Crow search algorithm (CSA), a method with natural form cues [25]. The method is used in based on population evolutionary computation models crow birds' behavior and social relationships. Undoubtedly intelligent creatures with larger-than-average neurons, crows are clever. They dwell in groups called flocks and hide their food in places that may still be found and retrieved a few months later. Additionally, they experience self-consciousness when carrying out the mirror test. They are able to recall appears, and if a poor one is noticed, they can converse with one another incomprehensibly to alert the other crows. Similar to other social animals, crows occasionally steal by carefully determining where other crows store their meals and then taking them. Whenever a crow suspects that someone is tracking it in an effort to deceive a thief, it is going to a new position that is distant from where the food is.

There are N solutions in the population, and the task's parameters are (overall of crow). The vector $l_j^t = \left[l_{j1}^t, l_{j2}^t, \ldots\ldots\ldots, l_{je}^t\right]$b or j=1, 2, 3... N labels the locations of each crow *j* at cycle t, where l_j^t is the potential possible placement options for crow *j* in dimensions e.

If a crow *j* proclaims that it is interested in robbing another crow *i*, one of two things may happen:

According to Eq. (11), Crow *j* will locate Crow *i*'s food store and update Crow *i*'s location rather than really following Crow *i*.

$$l_j^{(t+1)} = l_j^{(t)} + r_j * ck_i^{(t)} * (n_i^{(t)} - l_j^{(t)}) \quad (11)$$

Where the journey distance, *ck* is shown. r_i was randomly selected from the range [0, 1].

i pursue the crow to see where its food is concealed after realizing it is an endangered species. The crow *i* in the scenario does irregular dancing to trick the crow *j*.

In reality, the two parts may be mixed numerically as seen below in Eq. (12);

$$l_j^{(t+1)} = \begin{cases} l_j^{(t)} + r_j * ck_i^{(t)} * \left(n_i^{(t)} - l_j^{(t)}\right), r_i \geq BP_j^t \\ Select\ a\ random\ position, \quad Otherwise \end{cases} \quad (12)$$

where, *j* and *i* are random integers between 0 and 1, BP_j^t is the probability that Crow *i* at Iteration *t* will be aware, and r_j and r_i are random numbers. The ability of the crow to seek is affected by the values of *ck*. Low values of *ck* will aid in local optimal while high values will aid in global search.

Every crow is assessed as the technique is running using a well-defined fitness function. The crows then relocate themselves according to where they are most comfortable. The viability of every new role is evaluated. Using Eq. (13), the crows' memory is enhanced.

$$n_j^{(t+1)} = \begin{cases} l_j^{(t+1)} jg\ c\left(l_j^{(t+1)}\right) & is\ better\ than\ c\left(l_j^{(t)}\right) \\ n_j^{(t)}, & Otherwise \end{cases} \quad (13)$$

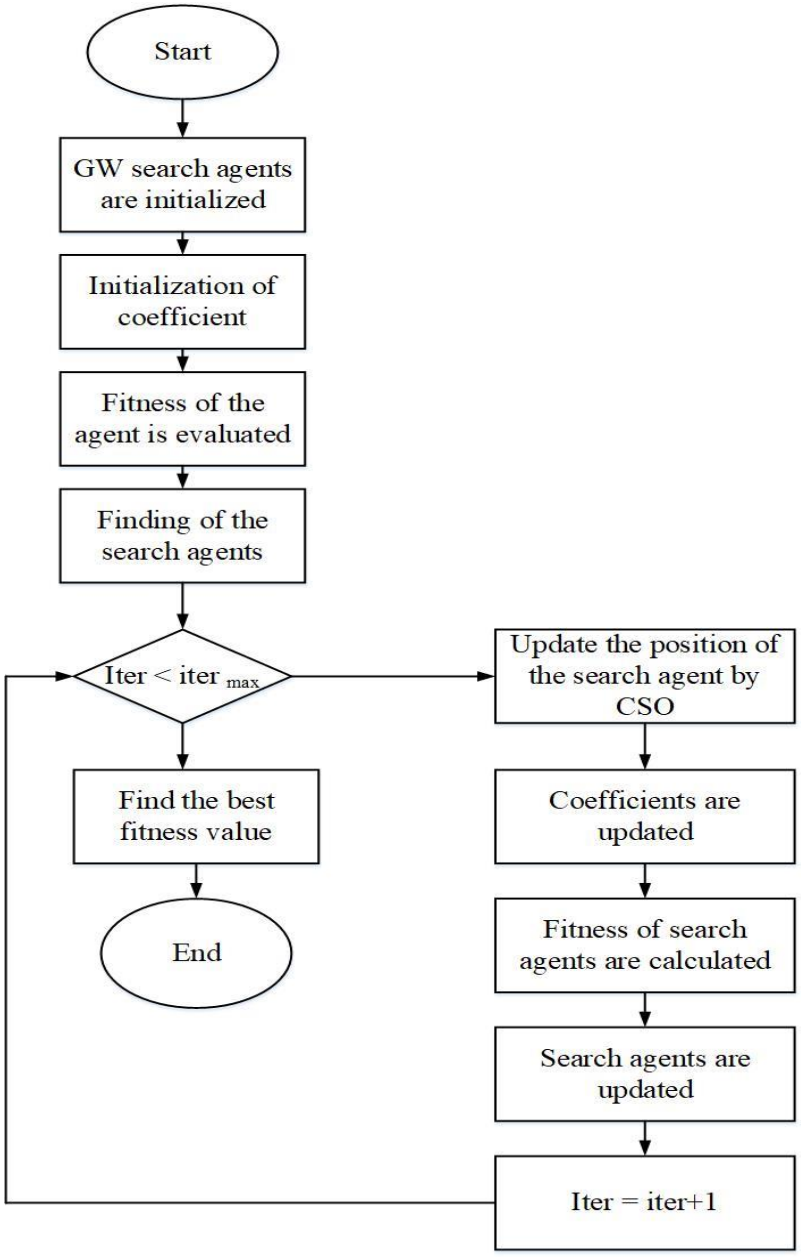


Fig. 3. Flow method of BGW-CS.

3) BGW-CS algorithm: To address the problem of immature integration, an improved research phase for BGWO is suggested in the current work. Through this method, BGWO and CSO algorithms are hybridized. In order to increase response accuracy, the created hybrid algorithm (BGW-CS) has a stronger propensity to pass over local optimal solutions. The flow methodology of BGW-CS is represented in Fig. 3. The equation of the hybrid BGW-CS is in Eq. (14),

$$A_{binary}(t+1) = \begin{cases} l_j^{(t+1)} jg\ c\left(l_j^{(t+1)}\right) & is\ better\ than\ c\left(l_j^{(t)}\right) \\ n_j^{(t)}, & Otherwise \end{cases} \quad (14)$$

D. Classification using Support Vector Machine

Apply the classification algorithm known as Support Vector Machine. Utilizing the multimodal medical data, train the SVM model with the pre-processed and chosen features to forecast diabetes outcomes because SVM can distinguish intricate correlations between features and is efficient in managing high-dimensional data, it is a popular choice. The SVM is used to group and stratify diverse classes of heterogeneous medical data, allowing for more precise predictions and individualised healthcare interventions. SVM uses its ability to construct hyperplanes that maximise the segregation of points of information in a multidimensional space of features to help categorise records of patients, outcomes of tests, and other important medical data. This categorization strategy is a crucial component of our model,

enhancing its accuracy and effectiveness in diabetes forecasting and providing healthcare practitioners with crucial information for better patient care and management approaches. The structural risk reduction idea is applied in SVM. Here, the learning machine's error rate is thought to be constrained by the interaction between the training error rate and the Vapnik Chervonenkis (VC) 1 dimension term. The hyper-plane that optimally distinguishes the data points for N training sets (X_i , Y_i) with labels, where $X_i \in$ Rn and $Y_i \in$ {–1, 1}, is illustrated in Eq. (15):

$$f(X_q) = \sum_{i=1}^{N} Y_i \propto_i K(X_q, X_i) + b \quad (15)$$

The sign of f (X_q) is used to calculate X_q in cases where the kernel functions are represented as K (.) and indicates the membership of the query sample. Making an ideal hyperplane is comparable to figuring out all nonzero _i, which stands for the bias b and the support vectors. The least amount of loss is anticipated while making a decision.

Using BGW-CSO to optimize the feature set and SVM to accurately forecast diabetes are the three main components of the comprehensive analysis methodology. Utilizing the advantages of each technique, this hybrid strategy seeks to improve the predictive model's overall performance.

V. Result and Discussion

The offered information from a research article describes a study that suggests a unique hybrid optimisation approach for enhancing the reliability and accuracy of diabetes prediction. To use multimodal medical data from the Pima Indian and Mendeley diabetes datasets, the model combines a Machine learning mechanism and an ensemble learning technique. To minimise the dimensions of the features, a hybrid technique known as Binary Grey Wolf-based Crow Search Optimisation (BGW-CSO) is utilized for selecting features. The selected features are then passed to a BGW-CSO-based Support Vector Machine (SVM) with enhanced hidden neurons. Utilising multiple performance criteria, the suggested model, BGW-CSO-SVM, is assessed and contrasted with existing methodologies. The findings show that the suggested methodology improves diabetes prediction accuracy, enabling early detection of at-risk patients and facilitating individualised patient management. According to the study, the hybrid optimisation model has the potential to significantly advance the science of diabetes prediction and help medical professionals make wise choices.

A. Evaluation of Performance Metrics

Accuracy, F1-score, precision, and recall were the four assessment measures used in the experiment to evaluate the models. These particular definitions of these parameters are given in Eq. (16) to Eq. (19):

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (16)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (17)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (18)$$

$$\text{F1score} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall+precision}} \quad (19)$$

The total amount of data which were correctly classified as positive out of all the positive data is referred to as the TP. The number of data which were incorrectly classified as negative out from one of is referred to as the TN. FN, Is it common for the model to mistakenly classify positive data as negative when, in fact, they were positive in the dataset. FP, Is it common for the model to mistakenly classify data as positive when, in reality, they were negative in the dataset. Recall is the ratio of the strategy's correctly classified positive pieces of information to those correctly classified positive pieces of information in the data set. In terms of the total quantity of variables that were classified as positive, precision is the proportion of data that the algorithm correctly recognised as positive. The F1 score is the harmonic average of recall and accuracy. Table I compares the performance of various classification methods or models. The methods evaluated include Soft Voting Classifier, Random Forest, DMP_MI, Bootstrap Aggregation, and a Proposed BGW-CSO-SVM.

Among the efficacy metrics that are assessed are F1-score, recall, accuracy, and precision. Accuracy is a measure of how well the strategy projections as a whole were predicted. Precision is the proportion of accurately predicted positive occurrences among all positive forecasts, whereas recall measures the ability to recognise positive examples. A single statistic called the F1-score integrates accuracy and recall. The Presented BGW-CSO-SVM approach obtained the greatest accuracy, precision, recall, and F1-score, showing greater efficiency when compared with the other techniques, as can be shown in Table I.

In Fig. 4, the Proposed BGW-CSO-SVM demonstrates the highest accuracy with a score of 96.62%. Following closely behind is the Random Forest method with an accuracy of 94.1%. The Bootstrap Aggregation technique achieves an accuracy of 94.62%, while the DMP_MI method achieves 87.1% accuracy. The Soft Voting Classifier has the lowest accuracy among the methods, with a score of 79.08%.

TABLE I. Comparison of Performance Metrics with Proposed Model

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Soft Voting Classifier [26]	79.08	73.13	70	71.56
Random Forest [27]	94.1	97.6	94.3	95.9
DMP_MI[28]	87.1	80.6	85.4	83.0
Bootstrap Aggregation [29]	94.62	94.7	94.6	94.6
Proposed BGW-CSO-SVM	96.62	98.54	96.3	97.8

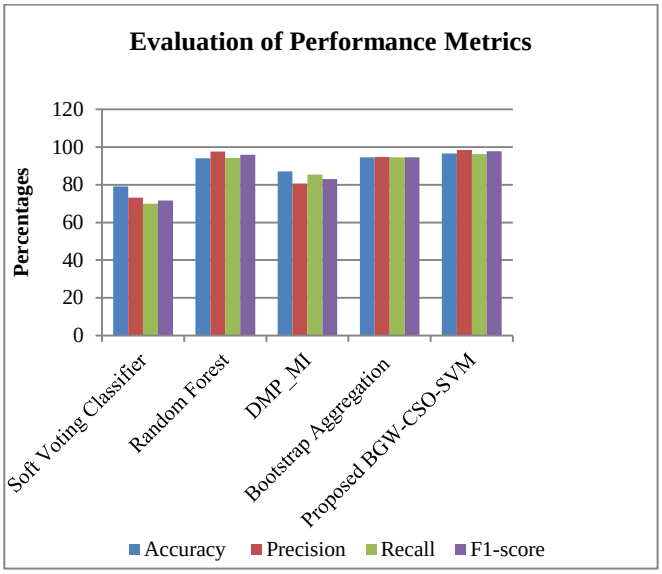


Fig. 4. Comparison graph of evaluation parameters.

In Table II, the proposed BGW-CSO-SVM model achieves the highest ROC value of 0.98, indicating strong discrimination capability between positive and negative instances in diabetes prediction. The Soft Voting method also demonstrates a high ROC value of 0.96, followed by LightGBM with a value of 0.95. XGBoost and Random Forest have ROC values of 0.93 and 0.94, respectively. AdaBoost shows a slightly lower ROC value of 0.92. Fig. 5 depicts the performance assessment of ROC curve. These results suggest that the proposed BGW-CSO-SVM model outperformed the other methods in terms of ROC value, indicating its potential for accurate prediction and effective management of diabetes.

The effectiveness of a classifier is often assessed using the ROC values in binary categorization tasks. Greater numbers denote greater performance. It shows a model's capacity to discriminate among both positive and negative examples.

A. Dataset Comparison

In Table III, the efficacy of the suggested BGW-CSO-SVM model in comparison to the approach presented by Taz, Islam, and Mahmud [30] is evaluated using two datasets: the Mendeley Diabetes Dataset and the PID Dataset. Fig. 6 depicts the proposed system dataset compared with existing approach. The evaluation metrics used are accuracy, precision, recall, and F1-score. The proposed BGW-CSO-SVM model consistently achieves higher accuracy, precision, recall, and F1-score on both the Mendeley Diabetes Dataset and the PID Dataset compared to the approach.

Table IV and Fig. 7 presents the results of a diabetes prediction study using different methods and their respective Root Mean Square Error (RMSE) values. The methods evaluated are Vanila-LSTM, BI-LSTM, and the Proposed Model.

TABLE II. COMPARISON OF PROPOSED SYSTEM ROC VALUE WITH EXISTING APPROACH

Methods	ROC value
Light GBM [30]	0.95
XGBoost [30]	0.93
AdaBoost [30]	0.92
Random Forest [30]	0.94
Soft Voting [30]	0.96
Proposed BGW-CSO-SVM	0.98

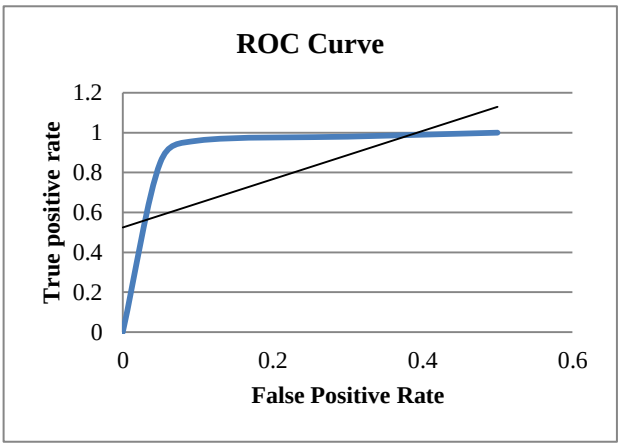


Fig. 5. ROC curve for the proposed model.

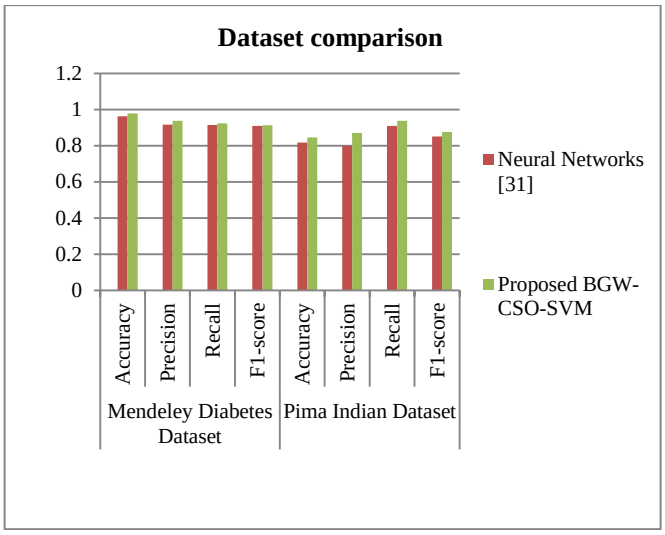


Fig. 6. Dataset comparison.

TABLE III. DATASET COMPARISON MDD AND PID

Methods	Mendeley Diabetes Dataset				Pima Indian Dataset			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Neural Networks [30]	96.24%	91.65%	91.46%	90.96%	81.82%	80.00%	90.91%	85.11%
Proposed BGW-CSO-SVM	97.9%	93.8%	92.36%	91.4%	84.63%	87.1%	93.8%	87.6%

TABLE IV. ERROR RATE COMPARISON

Method	RMSE
Vanila-LSTM [31]	15.43
BI-LSTM [31]	15.22
Proposed SVM	14.58

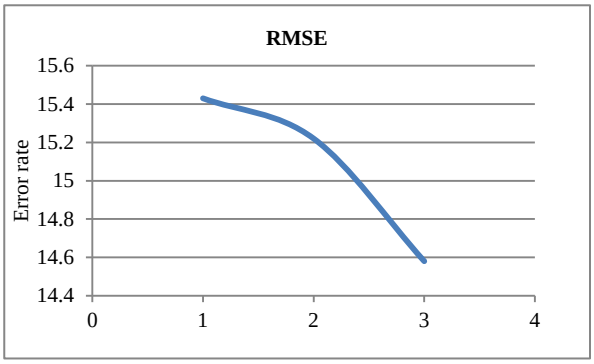


Fig. 7. Error rate graph.

B. Discussion

The study employs a feature selection technique, minimizes feature size, and ensures error-free classifications to effectively solve the difficulty of high-dimensional feature space in diabetes prediction. The research provides a novel methodology for feature selection, improving the model's capacity for generalization. It does this by introducing the revolutionary Binary Grey Wolf-based Crow Search Optimization (BGW-CSO) method, which combines Binary Grey Wolf Optimization (BGWO) with Crow Search Optimization (CSO). By optimizing the number of hidden neurons in the Support Vector Machine (SVM) through the use of BGW-CSO, the study improves diabetes prediction models overall and enhances the effectiveness of conventional SVMs. The suggested BGW-CSO-SVM model consistently outperforms other existing techniques, demonstrating superior accuracy, precision, recall, and F1-score through rigorous assessments on the Mendeley Diabetes Dataset and Pima Indian Dataset. The study's objective of creating a reliable diabetes prediction model that can be adjusted to fit a variety of datasets is perfectly aligned with this performance. Most importantly, the hybrid optimization model BGW-CSO-SVM greatly improves diabetes prediction accuracy, allowing for early at-risk individual identification, timely interventions, and individualized patient management. The results of the study highlight how well the objective of increasing the accuracy and consistency of diabetes prediction was met, and they may have ramifications for better patient outcomes in medical settings.

VI. CONCLUSION AND FUTURE WORK

The study addresses the pressing global health challenge of accurate and timely diabetes diagnosis. It introduces a novel hybrid optimization strategy that seamlessly integrates machine learning techniques to enhance the precision of diabetes prediction. This advancement is achieved by harnessing the potential of multimodal medical data, electronic health records, and advancements in data analytics. To ensure precise classification, two distinct datasets from the Pima Indian diabetes databases are used, alongside a rigorous feature selection method. The introduction of the BGW-CSO approach, a fusion of BGWO and CSO, bolsters feature selection capabilities. This innovation not only addresses the challenges posed by high-dimensional feature spaces but also significantly improves the system's ability to generalize. The performance of conventional Support Vector Machines (SVMs) benefits greatly from optimizing SVM techniques using the newly devised BGW-CSO approach. This proposed strategy, referred to as BGW-CSO-SVM, demonstrates substantial enhancements in diabetes prediction accuracy, as evidenced by a comprehensive examination using various performance metrics and comparisons with existing methodologies. This breakthrough enables the rapid identification of individuals at risk, paving the way for personalized and effective treatment interventions. In order to ensure the applicability of our methodology across a wide range of healthcare contexts, future research endeavours should prioritize testing its usefulness across varied datasets and demographics. It is important to prioritize creating solutions that are easy to use, accessible, and readily embraced by medical professionals. By facilitating a smooth transition into clinical settings, this strategy hopes to improve patient care and diabetes diagnosis on a larger scale. The research can demonstrate the robustness and generalizability of the suggested technique, increasing the likelihood of its widespread adoption and beneficial effects on healthcare practices. This can be achieved by expanding the validation to include a broad variety of datasets and demographic differences.

REFERENCES

[1] L. Garcia-Molina, A.-M. Lewis-Mikhael, B. Riquelme-Gallego, N. Cano-Ibanez, M.-J. Oliveras-Lopez, and A. Bueno-Cavanillas, “Improving type 2 diabetes mellitus glycaemic control through lifestyle modification implementing diet intervention: a systematic review and meta-analysis,” Eur. J. Nutr., vol. 59, no. 4, pp. 1313–1328, 2020, doi: https://doi.org/10.1007/s00394-019-02147-6.

[2] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting diabetes mellitus with machine learning techniques,” Front. Genet., vol. 9, p. 515, 2018, doi: https://doi.org/10.3389/fgene.2018.00515.

[3] T. Latchoumi, J. Dayanika, and G. Archana, “A comparative study of machine learning algorithms using quick-witted diabetic prevention,” Ann. Romanian Soc. Cell Biol., pp. 4249–4259, 2021.

[4] J. J. Wong et al., “Depression in context: Important considerations for youth with type 1 vs type 2 diabetes,” Pediatr. Diabetes, vol. 21, no. 1, pp. 135–142, 2020, doi: https://doi.org/10.1111/pedi.12939.

[5] G.-M. Huang, K.-Y. Huang, T.-Y. Lee, and J. T.-Y. Weng, “An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients,” in BMC bioinformatics, BioMed Central, 2015, pp. 1–10. doi: https://doi.org/10.1186/1471-2105-16-S1-S5.

[6] J. Zhu, Q. Xie, and K. Zheng, “An improved early detection method of type-2 diabetes mellitus using multiple classifier system,” Inf. Sci., vol. 292, pp. 1–14, 2015, doi: https://doi.org/10.1016/j.ins.2014.08.056.

[7] L. A. Sleeper et al., “Evaluation of Kawasaki disease risk-scoring systems for intravenous immunoglobulin resistance,” J. Pediatr., vol. 158, no. 5, pp. 831–835, 2011, doi: https://doi.org/10.1016/j.jpeds.2010.10.031.

[8] S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, and T. Saba, “Current techniques for diabetes prediction: review and case study,” Appl. Sci., vol. 9, no. 21, p. 4604, 2019, doi: https://doi.org/10.3390/app9214604.

[9] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017, doi: https://doi.org/10.1016/j.csbj.2016.12.005.

[10] A. J. Vickers, A. M. Cronin, E. B. Elkin, and M. Gonen, "Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers," BMC Med. Inform. Decis. Mak., vol. 8, pp. 1–17, 2008, doi: https://doi.org/10.1186/1472-6947-8-53.

[11] C. Chopra, S. Sinha, S. Jaroli, A. Shukla, and S. Maheshwari, "Recurrent neural networks with non-sequential data to predict hospital readmission of diabetic patients," in proceedings of the 2017 International Conference on Computational Biology and Bioinformatics, 2017, pp. 18–23. doi: https://doi.org/10.1145/3155077.3155081.

[12] S. Bashir, U. Qamar, and F. H. Khan, "IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework," J. Biomed. Inform., vol. 59, pp. 185–200, 2016, doi: https://doi.org/10.1016/j.jbi.2015.12.001.

[13] P. P. Debata and P. Mohapatra, "Diagnosis of diabetes in pregnant woman using a Chaotic-Jaya hybridized extreme learning machine model," J. Integr. Bioinforma., vol. 18, no. 1, pp. 81–99, 2020, doi: https://doi.org/10.1515/jib-2019-0097.

[14] ISTI, "Exploring Diabetes Epidemic in India | India Science, Technology & Innovation - ISTI Portal." Accessed: Jun. 23, 2023. [Online]. Available: https://www.indiascienceandtechnology.gov.in/featured-science/exploring-diabetes-epidemic-india

[15] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Inform. Med. Unlocked, vol. 16, p. 100203, 2019, doi: https://doi.org/10.1016/j.imu.2019.100203.

[16] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," Inform. Med. Unlocked, vol. 17, p. 100179, 2019, doi: https://doi.org/10.1016/j.imu.2019.100179.

[17] S. Cui, D. Wang, Y. Wang, P.-W. Yu, and Y. Jin, "An improved support vector machine-based diabetic readmission prediction," Comput. Methods Programs Biomed., vol. 166, pp. 123–135, 2018, doi: https://doi.org/10.1016/j.cmpb.2018.10.012.

[18] N. Ahmed et al., "Machine learning based diabetes prediction and development of smart web application," Int. J. Cogn. Comput. Eng., vol. 2, pp. 229–241, 2021, doi: https://doi.org/10.1016/j.ijcce.2021.12.001.

[19] G. Cappon, F. Prendin, A. Facchinetti, G. Sparacino, and S. D. Favero, "Individualized Models for Glucose Prediction in Type 1 Diabetes: Comparing Black-box Approaches To a Physiological White-box One," IEEE Trans. Biomed. Eng., pp. 1–11, 2023, doi: 10.1109/TBME.2023.3276193.

[20] J. Xie and Q. Wang, "Benchmarking Machine Learning Algorithms on Blood Glucose Prediction for Type I Diabetes in Comparison With Classical Time-Series Models," IEEE Trans. Biomed. Eng., vol. PP, Feb. 2020, doi: 10.1109/TBME.2020.2975959.

[21] M. Alirezaei, S. T. A. Niaki, and S. A. A. Niaki, "A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines," Expert Syst. Appl., vol. 127, pp. 47–57, 2019, doi: https://doi.org/10.1016/j.eswa.2019.02.037.

[22] R. D. H. Devi, A. Bai, and N. Nagarajan, "A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms," Obes. Med., vol. 17, p. 100152, 2020, doi: https://doi.org/10.1016/j.obmed.2019.100152.

[23] K. S. Prasad, N. C. S. Reddy, and B. Puneeth, "A framework for diagnosing kidney disease in diabetes patients using classification algorithms," SN Comput. Sci., vol. 1, no. 2, p. 101, 2020, doi: https://doi.org/10.1007/s42979-020-0096-7.

[24] P. Nuankaew, S. Chaising, and P. Temdee, "Average Weighted Objective Distance-Based Method for Type 2 Diabetes Prediction," IEEE Access, vol. 9, pp. 137015–137028, 2021, doi: 10.1109/ACCESS.2021.3117269.

[25] A. Askarzadeh, "A novel metaheuristic method for solving constrained engineering optimization problems: crow search algorithm," Comput. Struct., vol. 169, pp. 1–12, 2016.

[26] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," Int. J. Cogn. Comput. Eng., vol. 2, pp. 40–46, Jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.

[27] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," Procedia Comput. Sci., vol. 167, pp. 706–716, 2020, doi: 10.1016/j.procs.2020.03.336.

[28] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng, and D. N. Davis, "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values," IEEE Access, vol. 7, pp. 102232–102238, 2019, doi: 10.1109/ACCESS.2019.2929866.

[29] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, "An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study," Sensors, vol. 22, no. 14, p. 5247, Jul. 2022, doi: 10.3390/s22145247.

[30] N. H. Taz, A. Islam, and I. Mahmud, "A Comparative Analysis of Ensemble Based Machine Learning Techniques for Diabetes Identification," in 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), DHAKA, Bangladesh: IEEE, Jan. 2021, pp. 1–6. doi: 10.1109/ICREST51555.2021.9331036.

[31] H. Butt, I. Khosa, and M. A. Iftikhar, "Feature Transformation for Efficient Blood Glucose Prediction in Type 1 Diabetes Mellitus Patients," Diagnostics, vol. 13, no. 3, p. 340, Jan. 2023, doi: 10.3390/diagnostics13030340.

A Hybrid Movies Recommendation System Based on Demographics and Facial Expression Analysis using Machine Learning

Mohammed Balfaqih
Department of Computer and Network Engineering, College of Computer Science and Engineering
University of Jeddah, Jeddah, 23890, Saudi Arabia

***Abstract*—Cinemas and digital platforms offer an extensive array of content requiring tailored filtering to cater to individual preferences. While recommender systems prove invaluable for this purpose, conventional movie recommendations tend to emphasize specific attributes, leading to a reduction in overall accuracy and reliability. Notably, the extraction process of facial temporal attributes exhibits a suboptimal level of accuracy, thereby influencing the classification of attributes and the overall accuracy of the recommendation system. This article introduces a hybrid recommender system that seamlessly integrates collaborative filtering and content-based methodologies. The system takes into account crucial factors such as age, gender, emotion, and genre attributes. Films undergo an initial categorization based on genre, with a subsequent selection of the most representative genres to ascertain group preferences. Ratings for these selected movies are then predicted and organized in descending order. Employing Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models, the system achieves real-time extraction of facial attributes, particularly enhancing the accuracy of emotion attribute extraction through sequential processing. The CNN model demonstrates a commendable 55.3% accuracy score, the LSTM model excels with a 59.1% score, while the combined CNN and LSTM models showcase an impressive 60.2% accuracy. The performance of the recommendation system is rigorously evaluated using standard metrics, including precision, recall, and F1-measure. Results underscore the superior performance of the proposed system across various testing scenarios compared to the established benchmark. Nevertheless, it is noteworthy that the precision of the benchmark marginally surpasses the proposed system in the age groups of 8-14 and 15-24.**

***Keywords*—Recommender system; movies recommendation; emotion prediction; k-means clustering; deep learning**

I. Introduction

Recommender systems play a pivotal role in facilitating user exploration and item selection within the expansive choices available on web or electronic platforms. These systems employ advanced strategies, including content-based, collaborative, and hybrid approaches [1-4], to systematically filter and prioritize information, delivering clients tailored and pertinent data. Content-based filtering, a component of these systems, tailors recommendations based on individual preferences, effectively addressing the challenging cold start problem. Concurrently, collaborative filtering relies on user similarities or machine learning algorithms to recommend items. Hybrid recommender systems integrate both content-based and collaborative approaches, synergistically optimizing performance and honing the precision of recommendations. The amalgamation of these techniques empowers recommender systems to furnish users with suggestions that are not only personalized but also highly accurate [5-8].

A consumer's emotional state influences their decision-making process. Emotion is an unconscious mental state that arises spontaneously, accompanied by physiological and psychological changes in human organs and tissues, such as heart rate, facial expression, and the brain [9]. However, the recommendation process typically neglects the viewer's emotional state due to the intricate interplay of physiological signals with emotions, making subtle emotional expressions easily misunderstood. Prior research has predominantly focused on understanding user emotions through various means, such as ratings, comments, and helpfulness votes, among others [10, 11].

The Affective Video Recommender System (AVRS) has emerged as a prominent research area within recommender systems, diverging from traditional text, image, and speech emotion recognition. Focused on analyzing emotional states within videos and discerning emotions in distinct scenes [12], AVRS strategically recommends video content to viewers based on identified emotional states. An insightful study revealed gender-based disparities in movie preferences, with men exhibiting variations between mood and movie choices, while women tend to demonstrate a more congruent pattern [13]. To provide personalized recommendations in theaters or movie platforms devoid of recorded user data, the incorporation of face recognition becomes imperative. This facilitates the identification of attributes such as age, gender, and emotion estimation.

Within the domain of pattern recognition and computer vision, face recognition involves the identification of individuals by assessing distances between key facial points or the angles formed by facial components [14]. The creation of an efficient face recognition system necessitates considerations for speed, accuracy, scalability for system updates, and improvements in subject recognition. Fundamental face recognition approaches include holistic matching, feature-based (structural) methods that analyze local features, and hybrid techniques combining both holistic and feature

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant No. (UJ-21-DR-47). The author, therefore, acknowledges with thanks the University of Jedah technical and financial support.

extraction methodologies [15]. The culmination of these efforts results in the development of a functional and practical face recognition system.

An exhaustive review of literature on movie recommendation systems has identified two pivotal research challenges. Firstly, existing solutions for tracking and extracting facial temporal attributes exhibit substantial computational complexity and diminished accuracy, impacting attribute classification and diminishing the precision of recommendation systems. Secondly, prevalent movie recommendation systems concentrate on specific attributes, leading to a decline in overall accuracy and reliability [16-23]. As an extension of our previous work in [24], this study seeks to elevate the accuracy and efficacy of movie recommendations by crafting a precise face recognition system capable of discerning age, gender, and emotion from video data. Additionally, it discerns the most accurate deep learning models by incorporating multiple attributes, including user emotion and gender.

- A hybrid movies recommendation system based on demographics and facial expression analysis using machine learning. The system effectively captures user age, gender, and emotion in real-time through CNN and LSTM models to solve information overload and data sparsity problems.
- The system ensures the suitability for both new and existing users, providing effective movie recommendations, regardless of the presence of historical records or ratings. Movies are initially grouped by genre, selecting the most representative of preferred genres as the group's choice. Ratings for these movies are then predicted and listed in descending order.
- The study conducts experiments using real-world facial expressions data to demonstrate the excellent performance of the proposed methodology in the existing recommendation approach. It also identifies facial expressions as a crucial factor in movie recommendations.

The paper's structure is organized as follows: Section II addresses the research background, encompassing recommendation systems and face attribute recognition techniques, along with an exploration of related works and distinctions from existing systems. Section III provides a comprehensive overview of the proposed hybrid movie recommendation system, detailing the components of face attribute extraction, movie clustering, and recommendations. Section IV delves into the implementation setup and dataset usage, while Section V extensively discusses the findings derived from the proposed system. Finally, the paper concludes, summarizing the key insights and contributions.

II. Research Background

In this section, we explore the primary techniques employed by the proposed system, encompassing recommendation systems and facial attribute recognition. We also delve into recent research in these areas. In the final sub-section, we examine existing systems focused on integrating facial attributes recognition with movie recommendations, highlighting their limitations, and identifying research gaps.

A. Recommendation Systems

Recommendation systems, designed to address the information overload issue, filter pertinent information based on a user's interests, preferences, or observed behavior for a specific item [25, 26]. This problem arises as data volume increases, hindering effective decision-making [27]. The recommendation systems aim to provide meaningful user-specific recommendations for various items or products [28].

In the domain of movie recommendations, a multitude of recommendation systems has been explored in academic literature. Reddy S. et al. [16] proposed a framework akin to collaborative filtering techniques, integrating genre similarity and content-based filtering to enhance personalized recommendations. User feedback on films and genres significantly influences categorization, contributing to the customization of recommendations. Katarya Rahul [17] developed a hybrid recommender system utilizing the MovieLens dataset, incorporating the k-means clustering algorithm with bio-inspired artificial bee colony optimization.

Taking a distinct approach, [18] introduced an object-based collaborative filtering method, delving into the user's item rating matrix to establish connections among items for personalized recommendations. The author in [19] devised an efficient Graph Convolutional Network (GCN) algorithm, merging random walks and graph convolutions to generate embeddings. The author in [20] presented a comprehensive hybrid recommender system that integrates collaborative filtering via the Singular Value Decomposition (SVD) algorithm, a content-based system, and a fuzzy expert system. This expert system evaluates movie significance based on factors such as average rating and the number of ratings.

In contrast, [21] proposed a dynamic weighted hybrid recommender system, adapting the blend of collaborative filtering (CF) and content-based filtering (CBF) dynamically, deviating from fixed weights. Film-Conseil employs a machine learning algorithm to assess consumer advisory capability, replacing explicit movie scores [22]. Additionally, a movie recommendation system utilizes inductive learning, a machine learning method that effectively reduces sparsity and enhances scalability through experiments [23]. Nevertheless, current recommendation solutions exhibit limitations, lacking consideration for user emotions and demographics. Consequently, there is an imperative need to accurately and efficiently capture real-time user facial expressions.

B. Face Attributes Recognition Techniques

Feature detection and image matching are crucial tasks in machine vision, with varying computational efficiency and accuracy depending on the chosen feature detector and descriptor extraction algorithm. It is essential to select an appropriate algorithm for specific feature matching tasks [29].

In recent research, a significant focus has been placed on real-time analysis of age, gender, and emotional states. A pivotal study [30] employed a Convolutional Neural Network (CNN) to achieve a remarkable 95 percent accuracy in age and

gender identification using the IMDB-WIKI dataset. Simultaneously, the CNN demonstrated a 66 percent accuracy in emotion detection, leveraging the FER dataset. An alternative approach [31] utilized an artificial neural network, achieving a commendable 70.5 percent accuracy in age and gender recognition.

Imane et al. [32] introduced an innovative paradigm integrating the HAAR cascade and CNN for facial recognition, incorporating the FER2013 dataset for normalization and emotion detection. The integration of K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) algorithms resulted in a 70% accuracy rate. Rajesh et al. [33] contributed to the field with a real-time emotion detection system based on a nine-layer CNN, demonstrating an approximate 90 percent accuracy in categorizing seven distinct emotions.

Additionally, a method proposed by [34] deployed the Local Binary Pattern (LBP) classifier, achieving impressive results with a score of 94.39 percent using the CK+ dataset and 92.22 percent with the JAFFE dataset. Lastly, the method introduced by [34] implemented Exploratory Data Analysis (EDA), SVM, and demographic classification strategies, achieving an outstanding 99 percent accuracy and efficiency.

C. Related Works

Several systems focused on integrating facial attributes detection with items recommendations have been proposed in the literature. The authors in [36] introduced a video recommendation system centered on emotion detection, with the potential to address various conditions through a focus on human emotions and cognition. This system suggests YouTube videos based on captured emotions in images, videos, or webcam feeds, considering emotion intensity. For instance, individuals expressing happiness receive recommendations for funny videos, while those displaying sadness are guided to motivational content. The system has achieved an average emotion detection accuracy of 56%. For the same purpose, a dataset was created in [37] with five classes and then compared with an alternative dataset, showcasing state-of-the-art performance. The results revealed that, apart from CNN and DenseNet201, VGG16, InceptionV3, and MobileNetV2 exhibited superior accuracy compared to the collected dataset, affirming the excellence of our dataset over the alternative one. The primary limitations of these works involve accuracy concerns and a lack of consideration for user age and gender, which may not fully align with users' preferences.

Haar cascade and Local Binary Patterns Histogram (LBPH) algorithms were utilized in [38] for face detection, feature extraction, and emotion detection. Emotion detection relies on the FER 2013 dataset, while age and gender detection use the Adience dataset. For web application development, they implemented the Django framework. The video recommendation system adopts a content-based approach, customizing recommendations based on detected emotions, age, and gender. These personalized suggestions are sourced from the internet, prioritizing videos aligned with the target emotion, popularity, or user interactions. Video categories encompass music videos, motivational speeches, quotes, movies, cartoons, humor, action, and lifestyle.

To automate the process of identifying users and deducing their preferences from their content feedback of TV applications, a solution based on face detection and recognition services was proposed in [39]. Demographic characteristics (age and gender) classified the user, addressing the cold start problem. Smiles and emotions detected served as automatic feedback during content consumption. Accurate results were achieved with a frontal view of the face, while deviations from this angle and suboptimal lighting conditions could hinder face detection and recognition, particularly if parts like the eyes or mouth were not clearly visible. In [40], an innovative approach was introduced to address the lack of affective data for newly added videos on platforms like YouTube. It used reinforcement learning and deep bidirectional recurrent neural networks to process videos, gather affective annotations, and integrate emotion and affective intensity aspects, refining as user feedback was collected. Both implicit and explicit interactions, including facial expressions in real-time video streams, were tracked to train personalized reinforcement learning models for short-term affective behavior learning. This approach also highlighted the value of sequencing videos in different contexts to understand long-term affective trends using context-aware features. Its effectiveness was tested in experiments on two diverse video datasets.

An advertising video recommendation process was introduced in [41], leveraging computer vision and deep learning to gauge users' emotional responses to ads in real time by analyzing their facial expressions. This involved a CNN-based predictive model for rating predictions and a real-time SIFT algorithm-based similarity model to identify users with similar preferences. Instead of relying on users' historical records, the approach continuously updated a dynamic user profile based on real-time facial expression changes. Experimental tests using food advertising videos showcased the superiority of this method compared to conventional approaches like random recommendations, average ratings, and traditional collaborative filtering, offering improved recommendations for both existing and new users in the realm of advertising video recommendations.

Table I summarizes the existing movies/video recommendation systems based on face feature extraction. While there have been previous studies on the integration of videos/movies recommendation systems with face attributes extraction techniques, there are still gaps in the literature that call for further research. It can be concluded that the primary limitations to be addressed in this study pertain to the suboptimal accuracy of recommendation systems due to inadequate data, and the complexity of data analysis, which results in elevated computational costs.

TABLE II. A SUMMARY OF THE MOST RELATED MOVIES RECOMMENDATION SYSTEMS

Attributes	Customer detection	Purpose	Age	Gender	Emotion	Wild environment
Bokhare, A., & Kothari, T., 2023 [1].	Image	Video recommendation	Not consider	Not consider	7 categories	Not considered
Elias, T., et al., 2022 [2].	Image	Movies recommendation	Not considered	Not consider	5 categories	Not considered
Babanne, V., et al., 2020 [3].	Image	Video recommendation	Age groups	Considered	7 categories	Not considered
De Pessemier, T., 2016 [4].	Image	User's non-verbal affective feedback	Accurate ages	Considered	7 categories	Considered
Tripathi, A., et al., 2019 [5].	Video	User's non-verbal affective feedback	Age group	Considered	3 categories	Considered
Kim, G., et al., 2021 [6].	Video	Video advertisement recommendation	Age groups	Considered	-	Not considered
Proposed system	Video & image	Movies recommendation	Age groups	Considered	7 categories	Considered

III. PROPOSED HYBRID MOVIES RECOMMENDATION SYSTEM

In this section, an overall description of the proposed hybrid movies recommendation system. The outline of the proposed system framework is illustrated in Fig. 1. The proposed system will be implemented on a mobile/web platform for use in movie theaters or personal digital devices with embedded cameras. The system is composed of three individual modules, i.e., Face attributes extraction module, Movies clustering model based on movies attributes, and recommendation system. The first module is used to extract the demographic and emotion attributes of the users, while the second module clusters the movies into 19 different genres based on their attributes. The outputs from these modules are utilized by the recommendation system module to provide movie cluster recommendations tailored to specific user groups, considering age, gender, and emotion. Furthermore, the list of movies is organized based on the predicted ratings for all movies within the users' groups. The next subsections describe each module in detail.

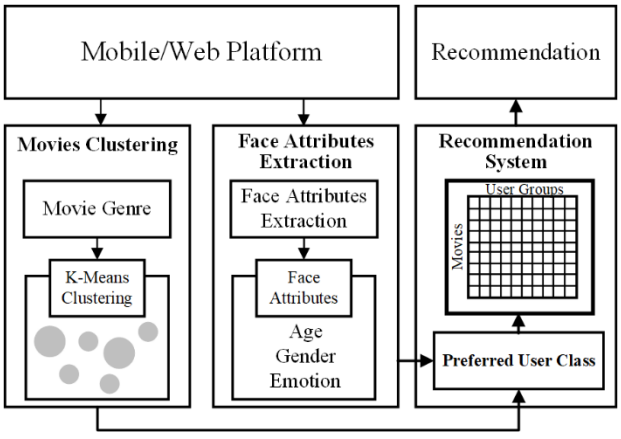


Fig. 1. A block diagram of the proposed system.

A. Face Attributes Extraction

The facial attribute extraction module is delineated into two discrete constituents: the CNN-based attributes extraction model and the LSTM-based attributes extraction model, as depicted in Fig. 2. This system facilitates real-time extraction of facial attributes, encompassing age, gender, and emotion. Positioned as a video-based application, its efficacy in capturing emotion attributes is heightened within a temporal sequence [42].

The CNN-based attributes extraction model concentrates predominantly on non-temporal features, with its output assuming a pivotal role in the ultimate synthesis. Notably, the Inception Net [43] and DenseNet [44] architectures are harnessed for their exceptional performance within this model. In contrast, the LSTM-based attribute extraction model is deployed for extracting vital expressive features, utilizing the VGG architecture to scrutinize emotional nuances within the temporal sequence. The symbiotic alignment of LSTM with the challenge's objectives manifests in competitive outcomes. The prognostications of facial attributes from the employed models are amalgamated by assigning weights to each method based on their performance metrics on the Acted Facial Expression in Wild (AFEW) validation set. The datasets utilized in this proposed system will be expounded upon in Section IV. The emotions considered for classification encompass Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise.

In both models, the initial step involves resizing the image to different scales, creating an image pyramid, which serves as input for the subsequent three-stage cascaded framework outlined in [45]. For face detection, candidate facial windows and their associated bounding box regression vectors are acquired using a fully convolutional network referred to as the proposal network (P-Net). Calibration of the candidates is performed based on the estimated bounding box regression vectors, followed by the application of non-maximum suppression (NMS) to consolidate highly overlapping candidates. Moving to the second stage, all candidates are directed through another CNN known as the refine network (R-Net), which serves to further eliminate a significant number of erroneous candidates, refine bounding boxes through regression, and perform NMS for accuracy. The final stage mirrors the second stage, with a focus on face regions that receive more supervision, leading to the network outputting the positions of facial landmarks.

1) CNN-based attributes extraction model: As shown in Fig. 2, the diagram of the model is divided into three main sections: Frames feature extraction, Frame-level feature aggregation, and classification. In frames feature extraction,

four networks are fine-tuned for the prediction of individual static images, specifically Inception V3, DenseNet121, DenseNet161, and DenseNet201. These networks were employed due to their efficient feature extraction and high image recognition accuracy. Inception-V3 is a popular CNN model known for its deep architecture and unique inception modules, enabling efficient feature extraction across multiple scales. It is optimized for CPU and GPU usage and has achieved accuracy rates exceeding 78.1% on ImageNet. On the other hand, DenseNet models are characterized by dense interconnections between layers. They vary in the number of layers, with DenseNet201 being the deepest. DenseNet models excel in image classification and feature extraction, delivering state-of-the-art results.

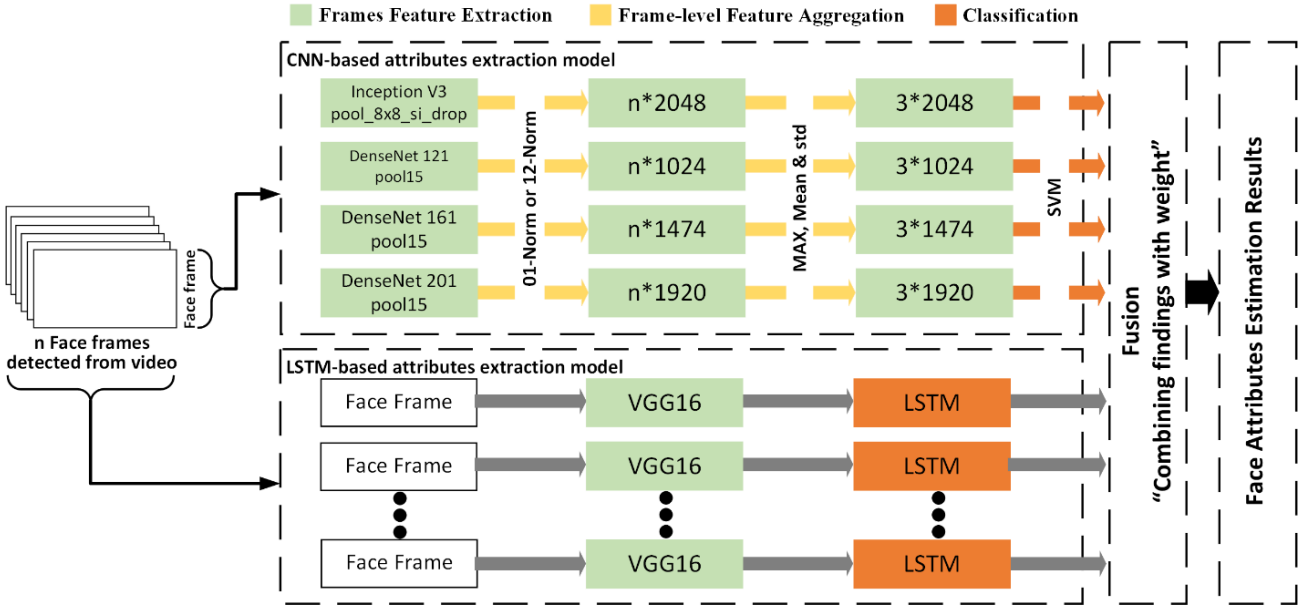


Fig. 2. Face attributes extraction module.

On the Real-world Affective Faces (RAF) validation test, these networks achieved accuracy scores of 82.74%, 83.84%, 83.25%, and 79.73%, with corresponding feature dimensions of 2048*3, 1024*3, 1474*3, and 1920*3, respectively. Subsequently, fine-tuned models extract features from the final layers of aligned faces, using them as the foundational representation. The Number of Features section shows the number of features extracted by each CNN layer. However, since the feature dimension in each video is directly linked to the number of detected faces and the layer dimension, normalization is used to standardize feature dimensions. The Video Features section shows the video features extracted by the CNN layers.

For frame-level feature aggregation, two normalization methods are applied separately to the features extracted by various CNN models from each aligned face in video frames using mean, max, and standard deviation. The video feature tripled compared to the initial CNN extraction. Following that, the RootSIFT and a normalization method ranging from [0,1] are applied to process the original feature [46]. Here's the calculation process:

$$F_i^{l2} = \frac{|F_i^v|}{\sum_j^n (F_j^v)^2} \quad (1)$$

$$F_i^{01} = \frac{F_i^v - \min(F^v)}{max(F^v) - min(F^v)} \quad (2)$$

Where F^v is the original feature extracted from a video, F_i^v is the index of feature in the feature vector. $max(F^v)$ and $min(F^v)$ stand for the maximum and minimum values in the video feature vector. F_i^{l2} and F_i^{01} are the features we normalized by l2-norm and 01-norm.

The classification step shows the output of the SVM. Linear SVMs were trained with various extracted features by the two normalization methods and four networks. Parameters were evaluated using 5-fold cross-validation. The results on the AFEW validation set for different features, based on SVM models, are notably lower than those in the RAF static image set. This difference can be attributed to the richer information in video clips regarding the expression process.

2) LSTM-based attributes extraction model: For face detection, the model is trained using VGG facial features on the AFEW training dataset. Maintaining stable face tracking is essential for optimal model performance in the analysis of the time sequence. Given that most video frames contain at least one face, the face detection threshold is set to a lower value to capture more faces while minimizing errors. A comprehensive set of face landmarks is leveraged to precisely locate the primary character's face, typically the one with the largest facial area. Additionally, a larger window is employed to capture finer details from regions like the forehead and chin.

For frame feature extraction, VGG-16 architecture is employed due to its exceptional quality of the FER2013 dataset in grayscale and its ample collection of meticulously selected faces. VGG-16 stands out for its depth, comprising 16 weight layers, uniform architecture with small 3x3 convolutional

filters, robustness, and high accuracy in image classification. The pre-trained VGG-Face Model is used on the FER2013 emotion dataset to fine tune the network and enhance its performance [14]. The images are resized to 224x224 in grayscale. Through the application of data augmentation techniques, 70.96% accuracy on the validation dataset is achieved which considered high level of performance.

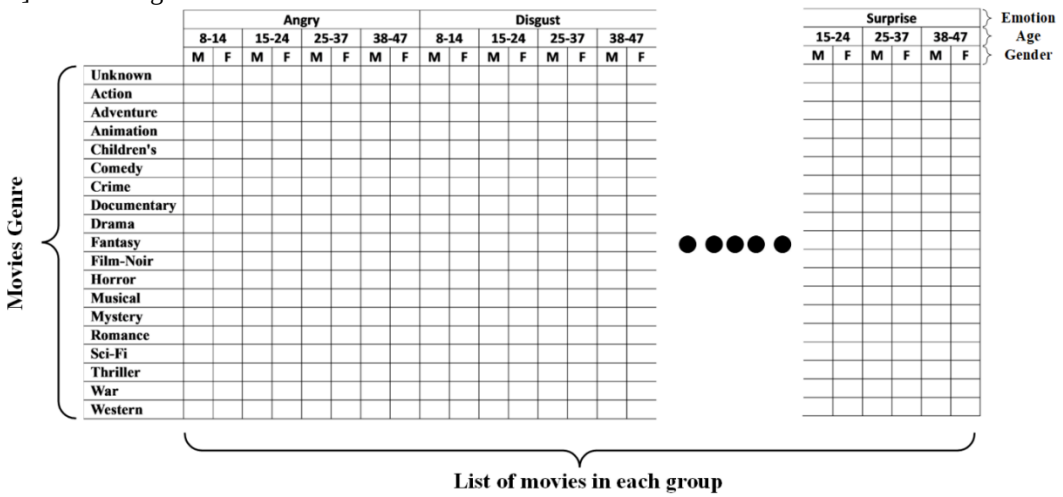


Fig. 3. Preferred movies based group clusters.

Finally, the classification is done using classic LSTM architecture, incorporating memory cell, input gate, output gate, and forget gate, was chosen. The implementation strategy aligns with [16]. LSTMs excel at handling long-term dependencies in sequential data, making them suitable for tasks like time series forecasting. Training videos have been segmented into 16-frame clips. An essential data augmentation step is the overlapping of clips by 8 frames, a well-established and effective technique for both training and testing. Additionally, mirror and multi-scale methods are employed. Temporal features from continuous video frames of facial expressions are extracted by the LSTM layer, utilizing a single LSTM layer with 128 embedding outputs. Notably, the final emotion prediction accuracy on the AFEW validation dataset achieves 46.21%. In contrast, utilizing an LSTM-256 layer instead of 128 leads to a slightly reduced accuracy of 43.07% in the validation dataset.

B. Movies Clustering based on Movie Attributes

Film characteristics will be derived from the Movielens dataset [47] to undergo system testing. Each film is delineated by its unique movie ID, title, release date, IMDb URL, and is classified across 19 genres, encompassing unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, and Western. To systematically categorize these films into distinct groups, the k-means algorithm will be employed—an unsupervised machine learning method.

This algorithm extracts insights from datasets through vectors, operating independently of labeled outcomes. It establishes a predetermined number (k) of cluster centroids within the dataset, computing the distances between each object and these centroids. Objects are subsequently assigned to the nearest cluster based on these distances, and the averages of all clusters are recalculated iteratively until the criterion function is satisfied. Attribute similarity is assessed using the Euclidean similarity approach, wherein objects with analogous attributes exhibit smaller dissimilarity distances, while those with disparate attributes display larger dissimilarity distances.

For a matrix X with i quantitative variables, the Euclidean distance d between two features, x_1 and x_2, can be computed as

$$d(x_1, x_2) = \sqrt{\sum_{i=1}^{n}(x_{1n}, x_{2n})^2} \quad (3)$$

The determination of the optimal K value in K-Means involves executing the algorithm with different k values and assessing variance. The selected K value corresponds to the point where variance is minimized. Variance, in this context, is computed as the cumulative sum of distances between each centroid and the items within its assigned cluster. The process includes plotting the variance against various K clusters, enabling the identification of the elbow point. The elbow point signifies the threshold at which the reduction in variance becomes notably stagnant. This allows for a systematic evaluation of K values based on variance metrics, providing a quantitative basis for determining the most suitable number of clusters in the K-Means clustering algorithm.

C. Movies Recommendation

Individuals are stratified based on attributes such as age group, gender, and emotional states to furnish tailored movie recommendations, as delineated in Fig. 3. The designated age cohorts encompass 8-14, 15-24, 25-37, and 38-47 years. Subsequently, the movies selected by each user are allocated to corresponding clusters. The cluster exhibiting the utmost representation of favored movie genres is identified as the group's preference. This determination relies on the cluster with the highest prevalence among all movie clusters, exemplified in a group with movie clusters [1, 1, 1, 2, 3, 3, 4, 1, 1], where cluster 1 is acknowledged as the favored group cluster.

Following this, the algorithm illustrated in Fig. 4 is employed to prognosticate movie ratings within each cluster. The system employs the Singular Value Decomposition (SVD) technique [47], a renowned Collaborative Filtering method. SVD, a matrix factorization method grounded in linear algebra, dissects a real matrix *X* into three matrices: *U*, *S*, and *V*. The SVD outcome encompasses matrix *U*, signifying user vectors, and matrix V^T, signifying movie vectors, with the singular values of *X* on the diagonal of matrix *S*, as depicted in the equation.

Algorithm: Ratings_prediction_of_movies

g: users' group (i.e., specific age, gender, and emotion)
m_g: a movie selected by a user group
m_c: a cluster of movies created by clustering module
C: representative movies cluster to specific user's group
m_i: a movie in the representative movies cluster
$\hat{X}$: rating of recommended movies
U: user vectors matrix
V^T: movies vectors matrix
S: diagonal matrix containing the singular values
m_p: movies preference for each *C*
Begin
for each users group *g* **do**
    selected movies $m_g \in m_c$ are listed
    identify *C* for each *g* by calculating maximum select movies cluster m_c
end for
for each movie $m_i \in C$ **do**
    rate the movies using $\hat{X} \approx U.S.V^T$
    list the movies preference m_p
end for
return m_p
End

Fig. 4. Algorithm of ratings prediction of the movies in each cluster.

$$X = U \times S \times V^T \tag{4}$$

SVD is selectively applied solely to movies in the favored group cluster, crafting a matrix where rows symbolize users in the same gender, age, and emotion category as active users, and columns signify chosen movies in a descending order. Matrix values denote the frequency of a specific movie being chosen. Ultimately, recommended movies with the highest ratings are ascertained through the dot product of *U*, *S*, and V^T, as elucidated in the equation.

$$\hat{X} \approx U.S.V^T \tag{5}$$

IV. Implementation Setup and Datasets

The computational framework is developed in Python and executed using Jupyter Notebook. Experimental procedures are conducted on a MacBook Pro featuring an Intel Core i7 processor and 8GB of RAM. Convolutional Neural Network (CNN) models, employed for facial attributes extraction, undergo pre-training utilizing the RAF and FER2013 datasets. The FER2013 dataset encompasses 28,709 training images, 3,589 validation images, and 3,589 testing images. Similarly, the RAF dataset comprises 12,271 training samples and 3,068 testing samples. Additionally, the AFEW dataset is applied for video clips emotion recognition, serving as a dynamic temporal facial expressions data repository derived from cinematic contexts. This dataset incorporates 957 samples, spanning six expression classes, and features neutral, natural head pose movements, occlusions, and a diverse array of subjects representing varied races, genders, ages, and other demographic characteristics. The Movielens 100k dataset is employed to facilitate the training and evaluation of the proposed system. Comprising 100,000 ratings across a 1 to 5 scale, the dataset involves 943 users rating 1682 movies, with each user contributing assessments for a minimum of 20 movies. The performance evaluation of the system involves the execution of 150 experimental observations.

V. Results and Discussion

This section first presents the outcomes of implementing a face attributes extraction model, followed by a movie recommendation model. The results of implementing face attributes extraction using CNN and LSTM based models individually and collaboratively are presented. Variations in sample sizes across classes highlight differences in category significance. To address this, class weights are employed, scaling scores by the square root of sample numbers, improving model performance in easily distinguishable categories. Model weights were determined through experiments on the validation set. The CNN-based model achieved a 55.3% score, while the LSTM-based model reached 59.1%. The combination of CNN and LSTM models achieved an impressive 60.07% accuracy which represents the overall accuracy of predicting the characteristics of the user correctly. However, the prediction accuracy of age, gender, and emotion are 86.3%, 87%, and 85%, respectively. Table II summarizes the accuracy findings of the proposed face attributes extraction model. Gender estimation errors mostly occurred among younger individuals aged 8-14. This may be due to the inherent difficulty in accurately predicting gender in children. On the other hand, age estimation exhibited a relatively high number of prediction errors in individuals aged [19-20]. This can be explained by the distinctiveness of aging patterns among individuals, which also varies based on gender.

Fig. 5. Preferred movies based group clusters.

In Fig. 5, an illustrative overview of movie recommendations incorporating estimated age, gender, and emotion is showcased. The assessment of movie recommendations' performance is undertaken, drawing comparisons with the methodology introduced in a prior study [32]. Assessment metrics include precision, recall, and F1-score. Precision, indicating the accuracy of movie predictions, is calculated as the ratio of true positive predictions to all positive predictions [49]. Recall, representing the model's

ability to predict occurrences, is computed as true positive predictions divided by actual positive predictions [49]. The F1-score, a harmonic mean of precision and recall, is determined using the specified formula [49].

The outcomes of these performance metrics are succinctly presented in Table II, underscoring the superior efficacy of the proposed system across various scenarios. While the benchmark system exhibits slightly higher precision in the 8-14 age group, the proposed system outperforms in recall and F1-score. Notably, the precision in the 15-24 age group reaches around 0.903, surpassing the proposed system's precision of approximately 0.862. Furthermore, in the 25-37 age group, the system achieves an average precision, recall, and F1-score of 0.873, 0.894, and 0.88, respectively.

TABLE III. THE ACCURACY OF THE PROPOSED FACE ATTRIBUTES EXTRACTION MODEL

		Emotion							Average (%)
Gender	**Age**	**Angry**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**	
Male	**8-14**	55.42	53.12	56.45	56.34	53.22	56.65	57.37	55.51
	15-24	58.34	56.23	57.40	62.35	61.44	56.32	58.43	58.64
	25-37	63.21	60.75	61.21	64.23	64.23	64.15	63.75	63.07
	38-47	62.56	60.34	60.13	64.56	64.25	65.43	64.35	63.08
Female	**8-14**	55.40	54.02	56.34	55.64	52.95	56.67	56.77	55.39
	15-24	58.02	55.97	57.82	62.15	61.64	55.81	58.82	58.60
	25-37	64.51	59.63	62.19	62.43	63.73	64.03	63.82	62.90
	38-47	62.77	62.43	59.94	64.02	65.02	64.86	64.85	63.41
Average (%)		60.02	57.81	58.93	61.46	60.81	60.49	61.02	60.07

While the proposed system demonstrates relatively high accuracy, it may not be suitable for users who cover their faces (e.g., wearing Hijab), as facial features cannot be extracted under such circumstances. To address this limitation, the system could enhance its capabilities by integrating demographic feature extraction through voice analysis. Furthermore, the computational efficiency and complexity of the face attribute extraction module could be enhanced by leveraging the advantages of the You Only Look Once (YOLO) algorithm due to its superior efficiency and suitability for real-time applications [50].

VI. CONCLUSION

Film venues and digital platforms provide an extensive array of cinematic options, often overwhelming consumers faced with a multitude of choices. To address this issue and enhance the efficiency of the decision-making process for clients, sophisticated recommendation systems have been devised. Existing movie recommendation systems, characterized by intricate computational processes, exhibit a notable deficiency in accurately tracking and extracting temporal facial attributes. This limitation compromises the precision of attribute classification, consequently diminishing the overall accuracy of the recommendation system. Furthermore, prevalent systems tend to overlook various attributes, contributing to a reduction in overall accuracy and dependability. This study introduces a novel hybrid recommender system that seamlessly integrates collaborative filtering and content-based methodologies. The system takes into account diverse attributes such as age, gender, emotion, and genre to optimize movie recommendations. Initially, movies undergo categorization based on genre, with a preference determined by selecting the most representative genres. Ratings for these films are subsequently predicted and presented in descending order. Employing Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models, the system conducts real-time extraction of facial attributes, enhancing accuracy in emotion attribute extraction by ensuring sequential extraction of attributes. Results from the system's implementation reveal its superior performance compared to the benchmark system across various test scenarios. However, the benchmark system exhibits marginally higher precision in the age groups of 8-14 and 15-24. Despite the proposed system's commendable accuracy, it faces limitations when users conceal their faces, necessitating improvements through the integration of demographic feature extraction via voice analysis and optimization of the computational efficiency of the face attribute extraction module utilizing the YOLO algorithm.

REFERENCES

[1] M. Elmisery, & D. Botvich, “Agent based middleware for private data mashup in IPTV recommender services,” In In 2011 IEEE 16th international work- shop on computer aided modeling and design of communication links and networks (CAMAD), pp. 107–111, 2011.

[2] B. Barragáns-Martínez, E. Costa-Montenegro, and J. Juncal-Martínez, “Developing a recommender system in a consumer electronic device,” Expert Systems with Applications, vol. 42, no. (9), pp. 4216–4228, 2015.

[3] M. Balfaqih, W. Jabbar, M. Khayyat, and R. Hassan, “Design and development of smart parking system based on fog computing and internet of things,” Electronics, vol. 10, no. 24, pp. 3184, 2021.

[4] A. Subasi, M. Balfaqih, Z. Balfagih, and K. Alfawwaz, “A comparative evaluation of ensemble classifiers for malicious webpage detection,” Procedia Computer Science, 194, pp. 272-279, 2021.

[5] M. Balfaqih, and S. A. Alharbi, “Associated Information and Communication Technologies Challenges of Smart City Development,” Sustainability, vol. 14, no. 23, pp. 16240, 2022.

[6] V. Shepelev, A. Glushkov, I. Slobodin, and M. Balfaqih, “Studying the Relationship between the Traffic Flow Structure, the Traffic Capacity of Intersections, and Vehicle-Related Emissions,” Mathematics, vol. 11, no. 16, pp. 3591, 2023.

[7] Yusof, M. H. M., Mokhtar, M. R., Zain, A. M., & Maple, C. "Embedded feature selection method for a network-level behavioural analysis detection model," International Journal of Advanced Computer Science and Applications, vol. 9, no. 12, 2018.

[8] M. M. H. Khan, E. W. K. Loh, and P. T. Singini, "Stabilization of tropical residual soil using rice husk ash and cement," International journal of applied environmental sciences, vol. 11, no. 1, pp. 73-87, 2016.

[9] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, (2018). "A review of emotion recognition using physiological signals," Sensors, vol. 18, no. 7, pp. 2074, 2018.

[10] S. Roy, and S. C. Guntuku, "Latent factor representations for cold-start video recommendation," In Proceedings of the 10th ACM conference on recommender systems, pp. 99-106, 2016.

[11] C. Orellana-Rodriguez, E. Diaz-Aviles, and W. Nejdl, "Mining affective context in short films for emotion-aware recommendation," In Proceedings of the 26th ACM Conference on Hypertext & Social Media, pp. 185-194, 2015.

[12] S. Zhang, X. Zhao, and Q. Tian, "Spontaneous speech emotion recognition using multiscale deep convolutional LSTM," IEEE Transactions on Affective Computing, vol. 13, no. 2, pp. 680-688, 2019.

[13] M. B. Devlin, L. T. Chambers, and C. Callison, "Targeting mood: Using comedy or serious movie trailers," Journal of Broadcasting & Electronic Media, vol. 55, no. 4, pp. 581-595, 2011.

[14] C. A. Hansen, "Face Recognition," Institute for Computer Science University of Tromso, Norway, 2009.

[15] R. Jafri, and H. R. Arabnia, "A survey of face recognition techniques," journal of information processing systems, vol. 5, no. 2, pp. 41-68, 2009.

[16] S. Reddy, S. Nalluri, S. Kunisetti, S. Ashok, and B. Venkatesh, "Content-Based Movie Recommendation System Using Genre Correlation. In: Smart Intelligent Computing and Applications," Smart Innovation, Systems and Technologies. Springer, Singapore, pp. 391–397, 2019.

[17] R. Katarya, "Movie recommender system with metaheuristic artificial bee. Neural Computing and Applications," vol. 30, no. 6, pp. 1983-1990, 2018.

[18] L. T. Ponnam, S. D. Punyasamudram, S. N. Nallagulla, and S. Yellamati, "Movie recommender system using item based collaborative filtering technique," In In 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), pp. 1–5, 2016.

[19] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, J. Leskovec, "Graph Convolutional Neural Networks forWeb-Scale Recommender Systems," In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '18. Association for Computing Machinery, London, United Kingdom, pp. 974–983, 2018.

[20] B. Walek, V. Fojtik, "A hybrid recommender system for recommending relevant movies using an expert system," Expert Systems with Applications, vol. 158, pp. 113452, 2020.

[21] H. Q. Do, T. H. Le, and B. Yoon, "Dynamic Weighted Hybrid Recommender Systems," In 2020 22nd International Conference on Advanced Communication Technology (ICACT), pp. 644-650, 2020.

[22] P. Perny, and J. D. Zucker, "Preference-based search and machine learning for collaborative filtering: the "film-conseil"movie recommender system," Informa- tion, Interaction, Intelligence, vol. 1, no. 1, pp. 9–48, 2001.

[23] P. Li, and S. Yamada, "A movie recommender system based on inductive learning," In IEEE conference on cybernetics and intelligent systems, pp. 318–323, 2004.

[24] M. Balfaqih, A. Altwaim, A. A. Almohammedi, and M. H. M. Yusof, "An Intelligent Movies Recommendation System Based Facial Attributes Using Machine Learning," In 2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA), pp. 1-6, 2023.

[25] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, "Recommendation systems: Principles, methods and evaluation," Egyptian informatics journal, vol. 16, no. 3, pp. 261-273, 2015.

[26] K. Haruna, M. Akmar Ismail, S. Suhendroyono, D. Damiasih, A. C. Pierewan, H. Chiroma, T. Herawan, "Context-aware recommender system: A review of recent developmental process and future research direction," Applied Sciences, vol. 7, no. 12, pp. 1211, 2017.

[27] J. Gantz, and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the future, vol. 2007, no. 2012, 1-16, 2012.

[28] P. Melville, and V. Sindhwani, "Recommender systems. Encyclopedia of machine learning, vol. 1, pp. 829-838, 2010.

[29] E. Karami, S. Prasad, and M. Shehata, "Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images," arXiv preprint arXiv:1710.02726, 2017.

[30] M. J. Uddin, P. C. Barman, K. T. Ahmed, S. A. Rahim, A. R. Refat, and M. Abdullah-Al-Imran, "A convolutional neural network for real-time face detection and emotion & gender classification," SR Journal of Electronics and Communication Engineering, vol. 15, no. 3, pp. 37-46, 2020.

[31] T. R. Kalansuriya, and A. T. Dharmaratne, "Neural network based age and gender classification for facial images," The International Journal on Advances in ICT for Emerging Regions, vol. 7, no. 2, 2014.

[32] I. Lasri, A. R. Solh, and M. El Belkacemi, "Facial emotion recognition of students using convolutional neural network," In 2019 third international conference on intelligent computing in data sciences (ICDS), pp. 1-6, 2019.

[33] G. A. Rajesh Kumar, and R. K. K. G. Sanyal, "Facial Emotion Analysis using Deep Convolutional Neural Network," 2017 International Conference on Signal Processing and Communication (ICSPC), pp. 369-374, 2017.

[34] S. L. Happy, and A. Routray, "Automatic facial expression recognition using features of salient facial patches," IEEE transactions on Affective Computing, vol. 6, no. 1, pp. 1-12, 2014.

[35] R. Azarmehr, R. Laganiere, W. S. Lee, C. Xu, and D. Laroche, "Real-time embedded age and gender classification in unconstrained video," In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 57-65, 2015.

[36] A. Bokhare, and T. Kothari, "Emotion Detection-Based Video Recommendation System Using Machine Learning and Deep Learning Framework," SN Computer Science, vol. 4, no. 3, pp. 215, 2023.

[37] T. Elias, U. S. Rahman, and K. A. Ahamed, "Movie Recommendation Based on Mood Detection using Deep Learning Approach," In 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), pp. 1-6, 2022.

[38] V. Babanne, M. Borgaonkar, M. Katta, P. Kudale, and V. Deshpande, "Emotion based personalized recommendation system," Int. Res. J. Eng. Technol.(IRJET), vol. 7, pp. 701-705, 2020.

[39] T. De Pessemier, D. Verlee, and L. Martens, "Enhancing recommender systems for TV by face recognition," In 12th international conference on web information systems and technologies (WEBIST 2016), vol. 2, pp. 243-250, 2016.

[40] A. Tripathi, T. S. Ashwin, and R. M. R. Guddeti, "EmoWare: A context-aware framework for personalized video recommendation using affective video sequences," IEEE Access, vol. 7, pp. 51185-51200, 2019.

[41] G. Kim, I. Choi, Q. Li, and J. Kim, "A CNN-based advertisement recommendation through real-time user face recognition," Applied Sciences, vol. 11, no. 20, pp. 9705, 2021.

[42] C. Liu, T. Tang, K. Lv, and M. Wang, "Multi-feature based emotion recognition for video clip," In Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 630-634, 2018.

[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826, 2016.

[44] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708, 2017.

[45] K. Zhang, Z. Zhang, Z., Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, 2016.

[46] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video, arXiv preprint arXiv:1711.04598, 2017.

[47] F. M. Harper, and J. A. Konstan, "The movielens datasets: History and context," Acm transactions on interactive intelligent systems (tiis), vol. 5, no. 4, pp. 1-19, 2015.

[48] M. G. Vozalis, and K. G. Margaritis, "Using SVD and demographic data for the enhancement of generalized collaborative filtering," Information Sciences, vol. 177, no. 15, pp. 3017-3037, 2007.

[49] Z. Omary and F. Mtenzi, "Machine learning approach to identifying the dataset threshold for the performance estimators in supervised learning," International Journal for Infonomics (IJI), vol. 3, no. 3, pp. 314-325, 2010 .

[50] K. S. Dixit, M. G. Chadaga, S. S. Savalgimath, G. R. Rakshith, and M. N. Kumar, "Evaluation and evolution of object detection techniques YOLO and R-CNN," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2S3, 2019.

Analysis of Ransomware Impact on Android Systems using Machine Learning Techniques

Anfal Sayer M. Al-Ruwili[1], Ayman Mohamed Mostafa[2]
Department of Computer Science-College of Computer and Information Sciences, Jouf University, Saudi Arabia[1]
Department of Information Systems-College of Computer and Information Sciences, Jouf University, Saudi Arabia[2]
Department of Information Systems-College of Computers and Informatics, Zagazig University, Egypt[2]

***Abstract*—Ransomware is a significant threat to Android systems. Traditional methods of detection and prediction have been used, but with the advancement of technology and artificial intelligence, new and innovative techniques have been developed. Machine learning (ML) algorithms are a branch of artificial intelligence that have several important advantages, including phishing detection, malware detection, and spam filtering. ML algorithms can also be used to detect ransomware by learning the patterns and behaviors associated with ransomware attacks. ML algorithms can be used to develop detection systems that are more effective than traditional signature-based methods. The selection of the dataset is a crucial step in developing an ML-based ransomware detection system. The dataset should be large, diverse, and representative of the real-world threats that the system will face. It should also include a variety of features that are informative for ransomware detection. This research presents a survey of ML algorithms for ransomware detection and prediction. The authors discuss the advantages of ML-based ransomware detection systems over traditional signature-based methods. They also discuss the importance of selecting a large, diverse, and representative dataset for training ML algorithms. Two datasets are applied during the conducted experiments, which are SEL and ransomware datasets. The experiments are repeated with different splitting ratios to identify the overall performance of each ML algorithm. The results of the paper are also compared to recent methods of ransomware detection and showed high performance of the proposed model.**

Keywords*—Ransomware; machine learning; malware detection; phishing detection; spam filtering

I. Introduction

During today's rapidly evolving technological landscape, the menace of malware remains a formidable challenge, with ransomware at its forefront. Cybercriminals consistently innovate to breach computer systems, propelling the need for more advanced detection and prediction methods. Particularly, ransomware is a dire threat to Android systems, prompting the exploration of innovative strategies driven by the surge of artificial intelligence and machine learning (ML).

Ransomware is a pernicious form of malware that encrypts valuable data, demanding a ransom for decryption [1]. This cybersecurity threat spans servers, computers, and smartphones, jeopardizing critical personal data and daily operations [2]. Android systems, in particular, are a prime target due to their open-source nature, enabling attackers to encrypt and hold data hostage, thereby escalating the impact of ransomware attacks. The advent of cryptocurrencies like Bitcoin has further complicated tracking both the attackers and their extorted funds [3].

The integration of artificial intelligence particularly ML, has emerged as a potent tool in the fight against ransomware. ML algorithms, distinguished by their effectiveness in various domains, excel in detecting phishing attempts, identifying malware, and filtering spam [3]. These algorithms can be trained on extensive datasets containing benign and malicious software to discern the unique behavioral patterns that set ransomware apart. Once trained, they can recognize new ransomware variants, by analyzing these distinctive behavioral patterns. The advantages of using ML for ransomware detection over traditional methods are profound. ML algorithms can identify new, previously unknown ransomware variants, adapt to evolving threat patterns, and minimize false positives. This is achieved by focusing on behavior patterns rather than static signatures or predefined rules [4].

Ransomware behavior is evolving rapidly and it targets many important assets, including critical systems such as Android. Therefore, it poses a challenge to detect and prevent it, and automated learning methods are effective ways to detect malicious ransomware behavior.

This paper presents an analysis of the malicious behavior of ransomware targeting the Android system using several machine learning algorithms such as Naive bayes, Support vector machine, Decision tree, k-nearest neighbors, Random forest, and Logistic Regression. This was conducted on various data sets containing many of the most common types of ransomware, and the data was divided into different proportions to ensure the effectiveness of the ML models.

II. Related Work

As presented in study [1], ransomware is malicious software that seizes a victim's data, encrypts data, and extorts money from the victim in exchange for their data. It evolves rapidly, making its detection challenging requiring continuously evolving detection tools. The authors of [2] categorized Ransomware into three main methods: Computer locker, I/O centric Locker, and Crypto miner. It follows a five-step process, including infection delivery, environment verification, hiding to avoid detection, target selection, and displaying a blackmail message to the victim.

As presented in study [4], the authors proposed an automated education-based approach for detecting

ransomware through three key stages: data collection related to ransomware, extraction of shared ransomware behaviors, and precise identification of harmful ransomware behaviors. The authors also emphasized the effectiveness of automated education in evaluating and verifying information. Ransomware targets a wide range of categories, including individuals, due to the personal value of their data, business databases, and commercial companies. It also targets local servers to damage multiple systems potentially. Additionally, there are general guidelines for ransomware protection, such as encrypting backups, utilizing updated firewalls, using the latest antivirus software, and implementing a strategy of reduced user privileges [5].

As presented in study [6], ransomware has increased in recent years, primarily due to its profitability, and it has targeted various sectors, including healthcare, industry, and education. As explained in study [7], ransomware can be categorized into encryption-focused and screen-locking variants. The increase in ransomware attacks is attributed to its availability as a service, with some attackers offering ransomware creation tools and taking a 20% cut of the ransom. Victims facing such attacks have four options: pay the ransom, restore data from backups, and attempt to guess the decryption key through brute force, or lose the data.

As presented in [8], the methods for addressing ransomware are categorized into two key aspects: prevention, including measures like backups, and detection, further divided into four categories. These categories are behavior analysis of data to identify suspicious changes and trigger alarms and to compare current operations to past ransomware behaviors. Finally, event-based detection includes traffic and API monitoring and detection through automated learning algorithms. As shown in study [9], conventional intrusion detection systems fall short in countering advanced attacks, thus necessitating more sophisticated detection programs. It highlighted one advanced approach, the honey pot, which is a system purposefully crafted to emulate a genuine system luring in potential attackers. As presented in study [10], numerous mechanisms exist to safeguard against ransomware infections. These include maintaining up-to-date system updates, which address vulnerabilities with each release. Additionally, employing the latest ransomware detection tools is crucial.

As explained in study [11], several reasons contributed to the intensive increase of ransomware: Encryption algorithms are a double-edged sword used for privacy and attacks, and Electronic currencies that allow the attacker to be anonymous. With the rapid development of ransomware, it has become easy and available to obtain. As presented in study [12], a detection-assisting proposal called PEAD is based on the API of detecting the attack before encryption.

As presented in study [13], many victims find themselves compelled to pay the ransom. Email fraud is the primary method of victim targeting, accounting for 59%, followed by websites at 24%. Furthermore, it introduced a ransomware detection tool utilizing machine learning. This tool monitors CPU usage, detecting deviations from normal performance as indicators of potential ransomware activity. Additionally, it scrutinizes file extensions executed on the device, issuing warnings for suspicious programs. Notably, it successfully alerts users during an attack, displaying 0 for benign behavior and 1 for harmful behavior when detected. As presented in [14], ransomware poses a significant cybersecurity threat and is a prevalent form of malware in cybercrime. Numerous variants characterize ransomware and represent a criminal innovation primarily driven by monetary gains, often utilizing cryptocurrencies such as Bitcoin.

As presented in study [15], a strategy that relies on dynamic analysis of prevalent ransomware families, such as WannaCry, was devised. This strategy involves monitoring the real-time impact on a system, including adding or deleting files. Furthermore, it involves observing packet behavior in Wireshark; a change in the Multiplex ID indicates a potential infection.

As proposed in study [16], the researchers introduced a static analysis technique for identifying ransomware. It gathered executable code samples from various ransomware families, categorized them based on their characteristics and employed automated machine learning for classification. Ransomware detection can be approached through various primary methods. The first method is signature-based detection, which involves comparing malicious signatures with known ones. The second method is inferential disclosure, which relies on comparing malicious code [17]. The conducted experiments in virtual environments are used to assess ransomware detection techniques. It observed that these techniques exhibited improved performance after a 24-hour period. Dynamic analysis emerged as a more accurate method, while signature detection and inferential detection were found to be less effective in identifying new and mysterious malicious families.

As presented in study [18], ransomware typically leaves victims with limited recourse for addressing the attack, often necessitating a ransom payment. The study detailed an examination of various ransomware variants and established a virtual environment for scrutinizing DNS activity during such attacks. The researchers employed a trace capture tool both before and following the execution of the attack. As authors of [19, 20], the ransomware follows a specific lifecycle. It initiates by constructing a malicious program, followed by propagation, reaching the target device, identifying the data to encrypt, performing encryption or locking, and ultimately resorting to blackmail.

Ransomware employs various methods to infiltrate victims and compromise their critical data. One of the primary tactics involves encrypting the victim's data to seize control and another method employs a lock screen approach. It's essential to recognize that cybercriminals pursue their malicious objectives, such as extortion, sabotage, and financial gain. Understanding the consequences of ransomware and being knowledgeable about defenses against this threat can enhance user and asset security. Therefore, in this section we will conduct a comparative analysis of the studies, focusing on four main aspects: the ransomware methods employed in each paper, the objectives behind ransomware infection and its impact, and the countermeasures utilized.

TABLE I. COMPARATIVE ANALYSIS OF RANSOMWARE AND THEIR COUNTERMEASURES

Ref	Ransomware Method	Objective	Ransomware Effect	Security Countermeasure
[1]	Ransomware in general	Profitability –Disruptive – blackmail.	Encryption of victims data	A model that analyzes the level of risk using inferential detection.
[2]	Crypto-Ransomware infection	Profitability –Disruptive – blackmail.	Encryption of victims data	Inferential Behavior & API Linking.
[4]	Crypto Ransomware	Profitability –Disruptive – blackmail.	Encryption of victims data	Automated learning algorithm for ransomware prevention.
[5]	Lock screen or encryption	Profitability –Disruptive – blackmail.	Data damage either by encryption or lock screen	A proposal that prevents the attack, slows the encryption and reduces the impact.
[6]	Cryptographic Ransomware	Profitability –Disruptive – blackmail.	Encryption of victims data	Survey of ransomware detection techniques.
[7]	Ransomware in general	Profitability –Disruptive – blackmail.	Data damage either by encryption or lock screen	Diagram with instructions for dealing with ransomware.
[8]	Crypto Ransomware Attack	.Profitability –Disruptive – blackmail	Encryption of victims data	Proposal for early detection.
[9]	Ransomware in general	.Profitability –Disruptive – blackmail	Data damage either by encryption or lock screen.	A three-layer proposal based on a honey pot.
[10]	Lock screen or encryption	.Profitability –Disruptive – blackmail	Data damage either by encryption or lock screen.	A proposal based on three-tiered security.
[11]	Ransomware in general	.Profitability –Disruptive – blackmail	Data damage either by encryption or lock screen	A proposal that uses machine learning algorithms.
[12]	Crypto-Ransomware	Profitability –Disruptive – blackmail.	Encryption of victims data	PEDA pre-encryption algorithm.
[13]	Ransomware in general	Profitability –Disruptive – blackmail.	Data damage either by encryption or lock screen	A proposal to identify benign or harmful behavior with an API-based proposal.
[14]	Bitcoin and the Financial Impact of Ransomware	Profitability –Disruptive – blackmail.	Financial impact awareness and vision that contributes to solutions against attack	NA
[15]	WannaCry Ransomware	.Profitability –Disruptive – blackmail	Encryption of victims data	Dynamic analysis to collect malware indicators
[16]	Lock screen or encryption	Profitability –Disruptive – blackmail.	Data damage either by encryption or lock screen	Static analysis based proposal.
[17]	Lock screen or encryption	Profitability –Disruptive – blackmail.	Data damage either by encryption or lock screen	Dynamic analysis and code comparison.
[18]	WannaCry Ransomware	Profitability –Disruptive – blackmail.	Encryption of victims data	Conducting an analysis only contributes to the manufacture of mechanism.
[19]	Lock screen or encryption	Profitability –Disruptive – blackmail.	Data damage either by encryption or lock screen	Analysis of detection mechanisms such as honey pot and dynamic analysis based on API programming.
[20]	Ransomware in general	Profitability –Disruptive – blackmail.	Data damage either by encryption or lock screen	A proposal based on machine learning that compares the neural network.

In Table I, it proposes a comparative analysis of different ransomware methods by explaining the main objective of each method, its main effect and the proposed security countermeasure for preventing the threat.

III. MACHINE LEARNING METHODS FOR RANSOMWARE APPLICATIONS

Machine learning (ML) is a powerful tool that can be used to detect ransomware applications. ML algorithms can be trained on large datasets of both benign and malicious software to learn the behavioral characteristics that distinguish ransomware from legitimate software. Once trained, these algorithms can be used to identify new and previously unseen variants of ransomware, including zero-day attacks, based on their behavioral patterns. ML-based ransomware detection has several advantages over traditional signature-based and heuristic-based detection methods. First, ML algorithms can detect new or unknown ransomware variants that do not match existing signatures or patterns. Second, ML algorithms can adapt to changing ransomware behavior patterns over time. Third, ML algorithms are less prone to false positives than signature-based and heuristic-based detection, as they rely on detecting actual behavior patterns rather than static code signatures or predefined rules [21].

Ransomware threatens the Android system significantly, as we have seen its impact in the previous part of this research. Many traditional methods have been applied to

detect and predict this malicious attack. However, with the recent development of technology and artificial intelligence, modern and innovative methods have been developed to detect ransomware attacks. The application of machine learning algorithms is one of the branches of artificial intelligence that has several important advantages, including phishing detection, malware detection, and spams filtering and contributes to commercial tasks.

It can be classified into different categories. One such method is supervised machine learning, which uses algorithms that require outside supervision in order to provide a data set for testing and training. As a result, the model uses decision trees and support vector machines to enable categorization and prediction. The other form of algorithm is unsupervised education, where the algorithms produce data based on their models K-Means. Combining the first two forms, semi-supervised machine learning is the third type. The final type is reinforcement machine learning, in which the algorithm responds to good or bad signals by repeating the task performance [22].

There are numerous techniques to implement ML-based ransomware detection. Utilizing ML algorithms to search for suspicious activities in network traffic is a typical strategy. The usage of ML algorithms, for instance, can be utilized to spot ransomware attacks by spotting surges in encryption activity. Utilizing ML algorithms to examine the behavior of active processes is an alternative strategy. For instance, ML algorithms can be used to spot processes that attempt to connect to known ransomware servers or that encrypt a huge number of files. ML-based ransomware detection is a rapidly evolving field, and new techniques are being developed all the time. As ransomware attacks become more sophisticated, ML will continue to play an increasingly important role in ransomware detection and prevention [23].

IV. Proposed Methodology

In this paper, ML algorithms can be used to detect ransomware by learning the patterns and behaviors associated with ransomware attacks. ML algorithms can be used to develop detection systems that are more effective than traditional signature-based methods. As illustrated in Fig. 1, the ransomware dataset is split into training and testing where the training dataset is preprocessed and then different ML algorithms are applied to measure accuracy. The testing dataset is then applied to identify the best accuracy for each splitting ratio.

The objective of using machine learning (ML) in ransomware detection is to develop systems that are more effective and efficient at detecting ransomware attacks than traditional signature-based methods. Ransomware attacks are becoming increasingly sophisticated and difficult to detect using traditional signature-based methods. ML algorithms can learn to identify the patterns and behaviors associated with ransomware attacks, and they can be used to develop detection systems that are able to detect new and emerging ransomware variants. ML-based ransomware detection systems can also be more efficient than traditional signature-based methods. Traditional signature-based methods require security vendors to maintain and update databases of signatures for known ransomware variants [24]. The overall methodology for managing ransomware attacks is presented in the following steps:

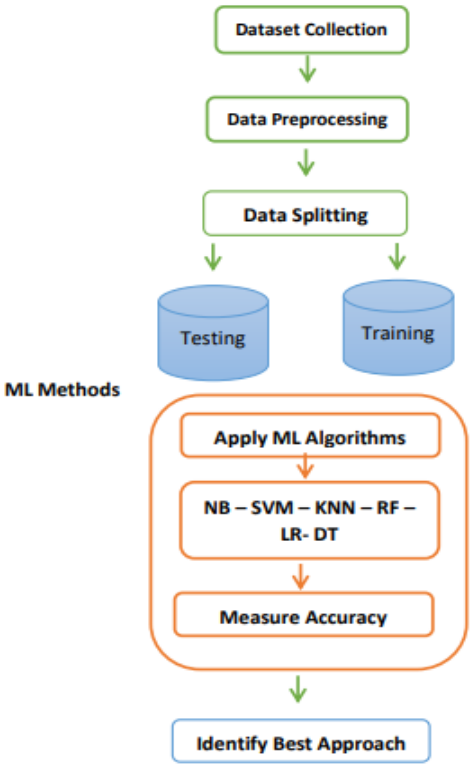


Fig. 1. Proposed ransomware methodology.

A. Dataset Collection

The selection of the dataset is a crucial step in developing a machine learning (ML)-based ransomware detection system. The dataset should be large, diverse, and representative of the real-world threats that the system will face. It should also include a variety of features that are informative for ransomware detection.

1) Security Engineering Lab (SEL) dataset: Ransomware threatens the Android system significantly, Many traditional methods have been applied to detect and predict this malicious attack. However, with the recent development of technology and artificial intelligence, modern and innovative methods have been developed to detect ransomware attacks. Applying machine-learning algorithms is a major advantage for phishing, malware, and spam detection. The Security Engineering Lab (SEL) dataset built in [25] based on 10153 samples of Android apps was used to be one of the latest data sets related to Android ransomware and benign software. As presented in Table I, the dataset contains 500 ransomware applications from several sources, such as the Ransom Proper Project, a timeline that classifies ransomware based on 15 families. It also contains 9653 benign applications collected from several reliable sources, such as the official Android store Google Play , that are distributed as presented in Table II.

TABLE II. DISTRIBUTION OF SEL DATASET

Type	Benign Programs	Ransomware Programs	Total
Number of Programs	9653	500	10153

2) Android ransomware dataset: In this dataset, 10 types of the latest ransomware for Android Taken from Kaggle [26], such as Pletor, Sim blocker, Wanna Locker, Jisut, SV peng, Porn Droid, Koler, Ransom BO, Charger, and Locker pin. The dataset contains 392034 records of benign data for Android programs , that are distributed as presented in Table III.

TABLE III. DISTRIBUTION OF RANSOMWARE DATASET

Attack Name	Number of Records
SVpeng	54161
PornDroid	46082
Koler	44555
Benign	43091
RansomBO	39859
Charger	39551
Simplocker	36340
WannaLocker	32701
Jisut	25672
Lockerpin	25307
Pletor	4715

B. Data Preprocessing

To analyze machine learning algorithms more accurately and reliably and to make sure there are no duplicate, wrong, or corrupted data in the dataset, it must be cleaned. This could lower the level of analysis and produce results that are inaccurate. Therefore, the cleaning procedure raises the data's quality. As can be seen in Table IV, methods have thus been used to enhance the data-cleansing process.

TABLE IV. DATA CLEANING OF SEL AND RANSOMWARE DATASETS

Process	SEL Dataset	Ransomware Dataset
Data Duplication	The dataset is of high quality and does not contain duplicate data	The dataset is of high quality and does not contain duplicate data
Unnecessary Data	-	Unnecessary data are eliminated
Missing Values	Removing missing values from the columns	-
Validation	The data has been validated	The data has been validated
Data Conversion	-	The dataset is converted into binary classification

C. Selection of ML Algorithms

This paper explains varieties of ML algorithms that can be used for ransomware detection as follows:

1) Naïve Bayes (NB): Ransomware detection is one of the many classification tasks that can be performed using the straightforward yet effective machine learning method known as Naive Bayes (NB). Based on the likelihood that each feature in a given data point belongs to a particular class, NB determines the likelihood that a given data point belongs to that class [24].

2) Support Vector Machine (SVM): Another effective machine learning approach for ransomware detection is SVM. Finding a hyperplane in the feature space that divides the data points into two classes (harmless and malevolent) is how SVMs operate. You would first need to compile a dataset containing both benign and dangerous software in order to employ SVMs for ransomware detection [24]. The file type, file size, file permissions, and file contents should all be included in this dataset. The model can be used to categorize fresh data points as benign or malicious after it has been trained. The model would determine the distance between the new data point and the hyperplane in order to accomplish this. The data are only used if the distance exceeds a predetermined threshold.

3) Decision Tree (DT): A supervised machine learning approach called decision trees (DT) is useful for both classification and regression tasks. DTs divide the feature space recursively into smaller and smaller sections until each zone only contains data points from one class [22]. After the dataset has been gathered, a DT model would need to be trained using the information. In order to achieve this, the feature space is recursively divided into smaller and smaller sections, until each zone only contains data points from a single class.

4) K-nearest neighbors (KNN): It is a simple and effective machine-learning algorithm that can be used for classification and regression tasks. KNN works by finding the K most similar data points to the new data point and assigning the new data point to the class of the majority of the K most similar data points [21].

5) Random Forest (RF): An ensemble learning approach called random forests (RFs) combines the predictions of various decision trees to create a forecast that is more accurate [22]. A huge number of decision trees are built using random selections of the data using RFs, and their predictions are then averaged. After gathering your dataset, you would need to use the information to train an RF model. To do this, many decision trees are built using random subsets of the data, and their predictions are then averaged.

6) Logistic Regression (LR): A machine learning approach for classification tasks is logistic regression (LR). By applying a logistic function to the data, LR operates [22]. A real integer is entered into the logistic function, a sigmoid function that returns a probability between 0 and 1. After gathering your dataset, you would need to use the information to train an LR model. In order for the model to forecast the likelihood that a data point would be malicious given its characteristics, the data must be fitted using a logistic function.

V. Experimental Results

The experimental results on two ransomware datasets are promising and suggest that ML algorithms can be used to develop effective ransomware detection systems. The accuracy, precision, recall, and F1-score are measured in both SEL and Ransomware datasets to explore the main performance of each dataset with different ML algorithms.

The accuracy is used to measure the performance of a ML model in predicting the correct class TP_c of a new data point. It is defined as the percentage of correct predictions made by the model as given in Formula (1).

$$Accuracy = \frac{TP_c+TN_c}{TP_c+FP_c+TN_c+FN_c} \quad (1)$$

A measure of precision in ML is the percentage of positive predictions that are in fact accurate. It is calculated by dividing the total number of accurate positive forecasts by the number of true positives as given in Formula (2).

$$Precision = \frac{TP_c}{TP_c+FP_c} \quad (2)$$

Recall is a metric used in ML that evaluates the percentage of real positives that are properly expected. It is calculated by dividing the total number of real positives by the number of true positives as given in Formula (3).

$$Recall = \frac{TP_c}{TP_c+FN_c} \quad (3)$$

The precision and recall measures for ML are combined into a single statistic called the F1-score. Given that it is defined as the harmonic mean of the precision and recall scores, both precision and recall are given equal weight as given in Formula (4).

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

A. Results of SEL Dataset

The experimental results are executed for training and testing the dataset using the predefined ML algorithms. The dataset is split based on two ratios 80:20 and 70:30. For the dataset split ratio of 80:20, the 80% is applied for training and the 20% is applied for testing. For the dataset split ratio of 70:30, the 70% is applied for training and the 30% is applied for testing. As presented in Table V the data is split to 20% for testing 80% for training and the performance of different machine learning is measured based on the accuracy, recall, and F1-score.

TABLE V. Performance Analysis of ML Algorithms with 80:20 Splitting for SEL Dataset

ML Alg.	Accuracy	Recall	F1-score
SVM	99.26%	99.43%	99.61%
RF	97.68%	99.89%	98.79%
NB	95.02%	**100%**	97.44%
KNN	99.06%	99.48%	99.50%
DT	97.48%	97.97%	98.66%
LR	**99.31%**	99.63%	**99.63%**

As presented in Fig. 2, the LR algorithm achieved the highest accuracy of 99.31% on the SEL dataset, while the SVM algorithm recorded 99.26%. The KNN algorithm recorded the third-highest accuracy, with 99.06%. It is followed by the RF algorithm with an accuracy of 97.68%, while the DT algorithm recorded an accuracy of 97.48%. The NB algorithm achieved 95.02%, which is considered the lowest accuracy on the SEL dataset.

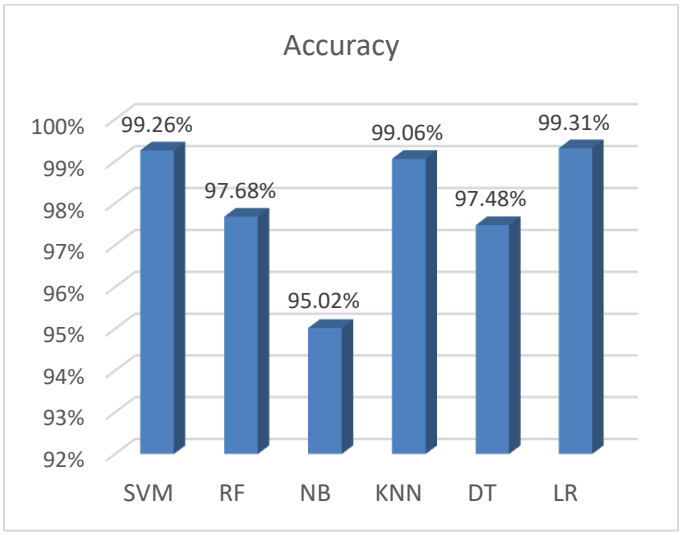


Fig. 2. Accuracy of different ML algorithms on SEL dataset with splitting ratio of 80:20.

As presented in Fig. 3, the NB algorithm achieved the highest recall of 100% on the SEL dataset, while the RF algorithm recorded 99.89%. The LR algorithm recorded 99.63%. It is followed by the KNN algorithm with a recall of 99.48%. The SVM algorithm recorded a recall of 99.43%, while the DT algorithm achieved 97.97%, which is considered the lowest recall on the SEL dataset.

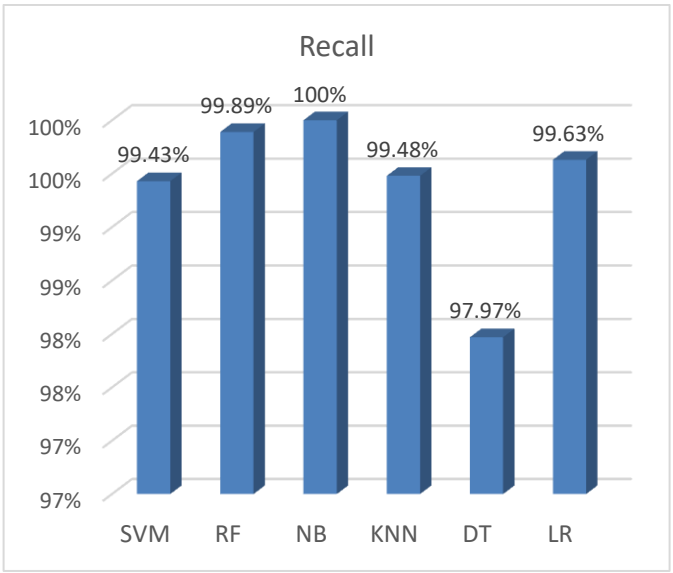


Fig. 3. Recall of different ML algorithms on SEL dataset with splitting ratio of 80:20.

As presented in Fig. 4, the LR algorithm achieved the highest F1-score of 99.63% on the SEL dataset, while SVM recorded 99.61%. The KNN algorithm recorded 99.50%, while the RF algorithm achieved an F1-score of 98.79%. The DT algorithm recorded an F1-score of 98.66%, while the NB algorithm achieved 97.44%, which is considered the lowest F1-score on the SEL dataset.

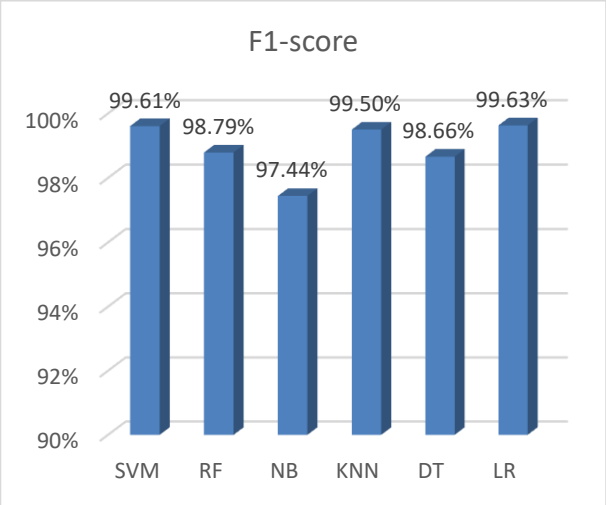


Fig. 4. F1-score of different ML algorithms on SEL dataset with splitting ratio of 80:20.

As presented in Table VI, the data is split to 30% for testing 70% for training and the performance of different machine learning is measured based on the accuracy, recall, and F1-score.

TABLE VI. PERFORMANCE ANALYSIS OF ML ALGORITHMS WITH 70:30 SPLITTING FOR SEL DATASET

ML Alg.	Accuracy	Recall	F1-score
SVM	99.24%	99.41%	99.60%
RF	98.32%	99.89%	99.12%
NB	95.27%	**99.96%**	97.5%
KNN	99.31%	99.65%	99.63%
DT	93.30%	93.17%	96.36%
LR	**99.47%**	99.72%	**99.72%**

As presented in Fig. 5, the LR algorithm achieved the highest accuracy of 99.47% on the SEL dataset, while the KNN algorithm recorded 99.31%. The SVM algorithm recorded the third-highest accuracy, with 99.24%. It is followed by the RF algorithm with an accuracy of 98.32%, while the NB algorithm recorded an accuracy of 95.27%. The DT algorithm achieved 93.30%, which is considered the lowest accuracy on the SEL dataset.

As presented in Fig. 6, the NB algorithm achieved the highest recall of 99.96% on the SEL dataset, while the RF algorithm recorded 99.89%. The LR algorithm recorded 99.72%. It is followed by the KNN algorithm with a recall of 99.65%. The SVM algorithm recorded a recall of 99.41%, while the DT algorithm achieved 93.17%, which is considered the lowest recall on the SEL dataset.

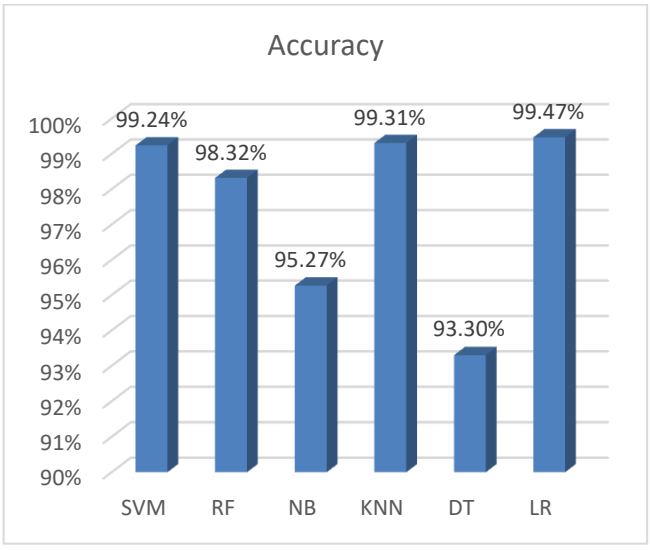


Fig. 5. Accuracy of different ML algorithms on SEL dataset with splitting ratio of 70:30.

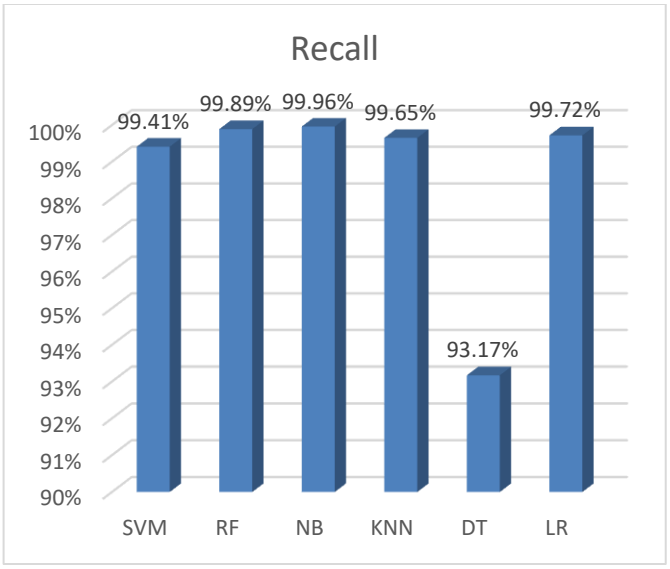


Fig. 6. Recall of different ML algorithms on SEL dataset with splitting ratio of 70:30.

As presented in Fig. 7, the LR algorithm achieved the highest F1-score of 99.72% on the SEL dataset, while KNN recorded 99.63%. The SVM algorithm recorded 99.60%, while the RF algorithm achieved an F1-score of 99.12%. The NB algorithm recorded an F1-score of 97.50%, while the DT algorithm achieved 96.36%, which is considered the lowest F1-score on the SEL dataset.

B. Results of Ransomware Dataset

The experimental results are conducted again on the Ransomware dataset to reevaluate the overall performance of the detection methodology. The Ransomware dataset is also split to 80:20 and 70:30 ratios to check whether the performance will be enhanced or not. As presented in Table VII, the data is split to 20% for testing 80% for training and the performance of different machine learning is measured based on the accuracy, recall, and F1-score.

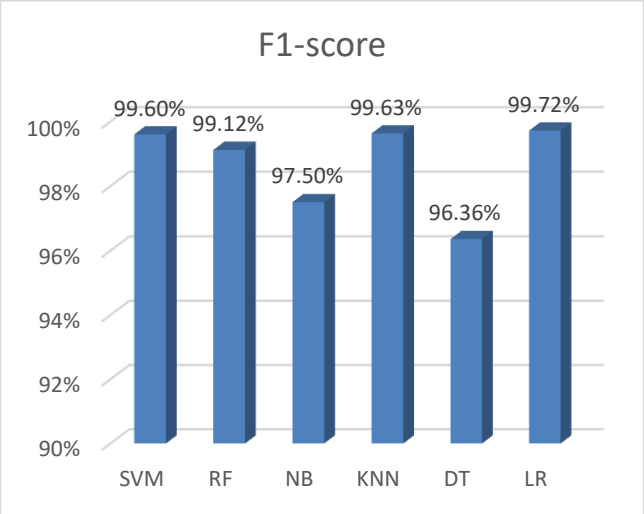


Fig. 7. F1-score of different ML algorithms on SEL dataset with splitting ratio of 70:30.

TABLE VII. PERFORMANCE ANALYSIS OF ML ALGORITHMS WITH 80:20 SPLITTING FOR RANSOMWARE DATASET

ML Alg.	Accuracy	Recall	F1-score
RF	90%	96.4%	94.7%
NB	89.01%	**100%**	94.18%
KNN	**93.25%**	97.12%	**96.24%**
DT	79.31%	84%	88%
LR	89.49%	96.9%	94.2%

As presented in Fig. 8, the KNN algorithm achieved the highest accuracy of 93.25% on the Ransomware dataset, while the RF algorithm recorded 90.00%. The LR algorithm recorded the third-highest accuracy, with 89.49%. It is followed by the NB algorithm with an accuracy of 89.01%. The DT algorithm achieved 79.31%, which is considered the lowest accuracy on the Ransomware dataset.

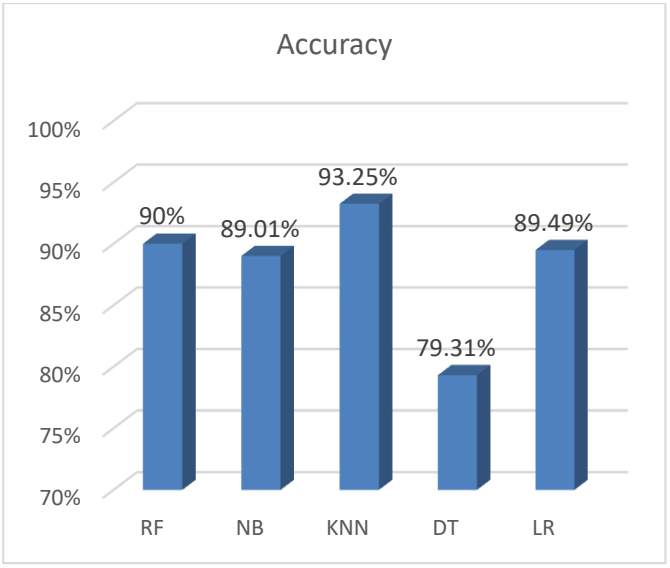


Fig. 8. Accuracy of ML algorithms on Ransomware dataset with splitting ratio of 80:20.

As presented in Fig. 9, the NB algorithm achieved the highest recall of 100% on the Ransomware dataset, while the KNN algorithm recorded 97.12%. The LR algorithm recorded the third-highest recall, with 96.90%. It is followed by the RF algorithm with a recall of 96.40%. The DT algorithm achieved 84%, which is considered the lowest recall on the Ransomware dataset.

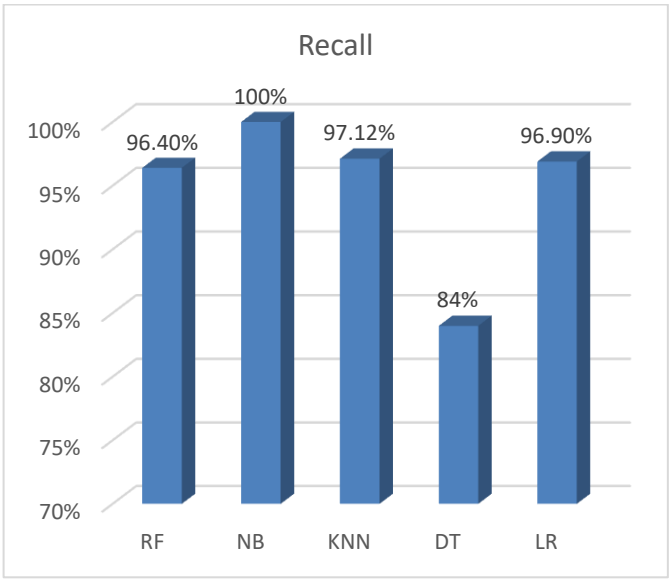


Fig. 9. Recall of ML algorithms on Ransomware dataset with splitting ratio of 80:20.

As presented in Fig. 10, the KNN algorithm achieved the highest F1-score of 96.24% on the Ransomware dataset, while the RF algorithm recorded 94.70%. The LR algorithm recorded the third-highest F1-score, with 94.20%. It is followed by the NB algorithm with an F1-score of 94.18%. The DT algorithm achieved 88%, which is considered the lowest F1-score on the Ransomware dataset.

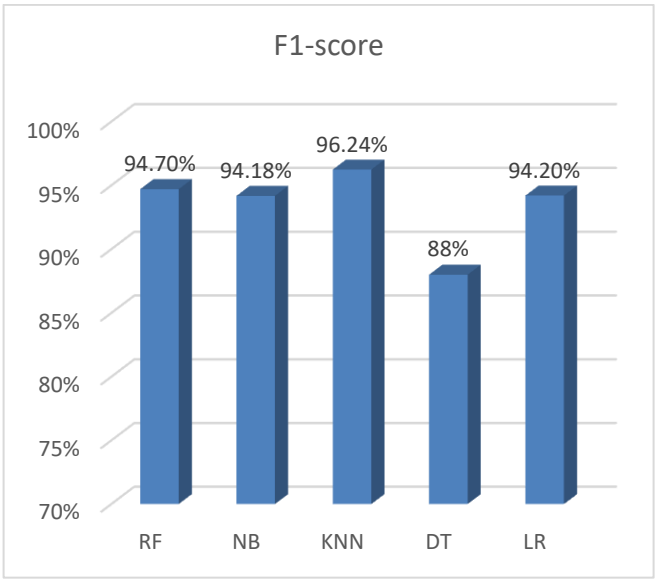


Fig. 10. F1-score of ML algorithms on Ransomware dataset with splitting ratio of 80:20.

TABLE VIII. PERFORMANCE ANALYSIS OF ML ALGORITHMS WITH 70:30 SPLITTING FOR RANSOMWARE DATASET

ML Alg.	Accuracy	Recall	F1-score
RF	89.32%	97.77 %	94.21%
NB	88.96%	**99.97%**	94.15%
KNN	**93.12%**	97.11%	**96.16%**
DT	82.88%	87.79%	90.12%
LR	89.45%	96.95%	94.24%

As presented in Table VIII, the data is split to 30% for testing 70% for training and the performance of different machine learning is measured based on the accuracy, recall, and F1-score.

As presented in Fig. 11, the KNN algorithm achieved the highest accuracy of 93.12% on the Ransomware dataset, while the LR algorithm recorded 89.45%. The RF algorithm recorded the third-highest accuracy, with 89.32%. It is followed by the NB algorithm with an accuracy of 88.96%. The DT algorithm achieved 82.88%, which is considered the lowest accuracy on the Ransomware dataset.

As presented in Fig. 12, the NB algorithm achieved the highest recall of 99.97% on the Ransomware dataset, while the RF algorithm recorded 97.77%. The KNN algorithm recorded the third-highest recall, with 97.11%. It is followed by the LR algorithm with a recall of 96.95%. The DT algorithm achieved 87.79%, which is considered the lowest recall on the Ransomware dataset.

As presented in Fig. 13, the KNN algorithm achieved the highest F1-score of 96.16% on the Ransomware dataset, while the LR algorithm recorded 94.24%. The RF algorithm recorded the third-highest F1-score, with 94.21%. It is followed by the NB algorithm with an F1-score of 94.15%. The DT algorithm achieved 90.12%, which is considered the lowest F1-score on the Ransomware dataset.

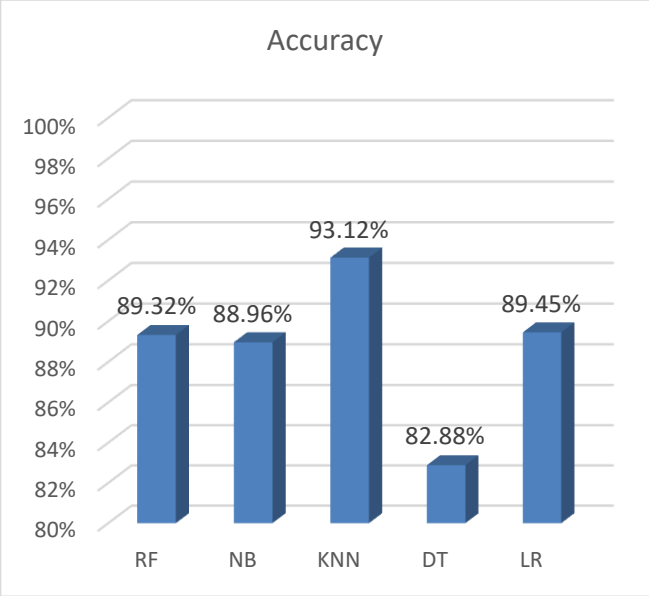


Fig. 11. Accuracy of ML algorithms on Ransomware dataset with splitting ratio of 70:30.

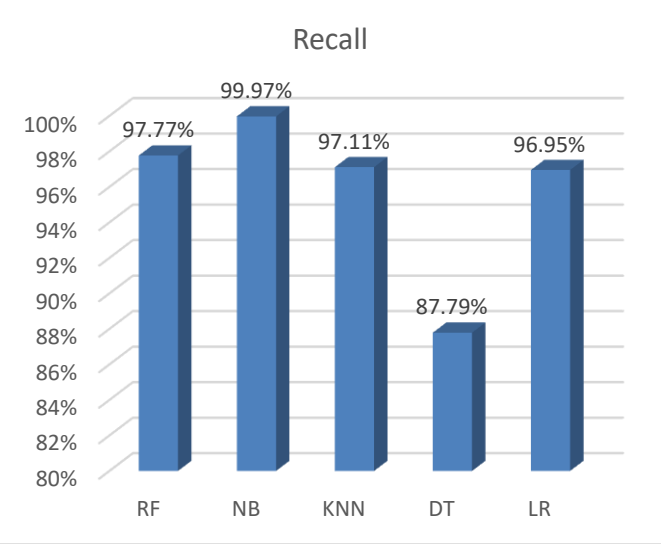


Fig. 12. Recall of ML algorithms on Ransomware dataset with splitting ratio of 70:30.

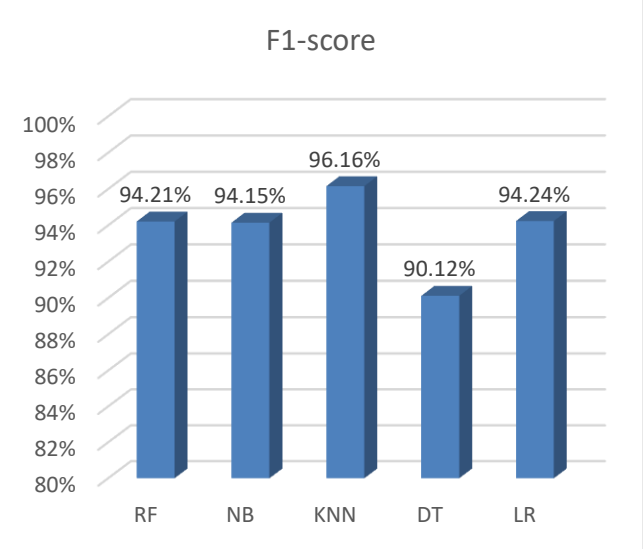


Fig. 13. F1-score of ML algorithms on Ransomware dataset with splitting ratio of 70:30.

VI. DISCUSSION

LR achieved the highest accuracy rate of 99.47%, and accuracy is the classifier's overall accuracy in correctly classifying samples as either ransomware or benign software. But if we were to verify the accuracy of the classifier in classifying the positive samples that are actually ransomware, we can see that Recall achieved the highest percentage of NB in the SEL dataset, at 100% in dividing the data 80:20 and 99.96% in dividing the data 70:30. It also achieved the highest accuracy rate in the KNN ransomware data set, at 93.25%, which is the overall accuracy of the classifier for classifying ransomware and benign samples. It is also worth mentioning that NB also achieved the highest Recall rate in the second data set at 100%. Therefore, NB is considered a good model to verify the accuracy of the classifier in detecting samples that are actually ransomware.

TABLE IX. COMPARISON OF RECENT ML ALGORITHMS ACCURACY WITH THE PROPOSED MODEL

Ref	ML Algorithm	Dataset	Accuracy
[4]	accuracy of classification algorithms	Dataset of 10 ransomware types	Not stated
[11]	regression and rule- based algorithms	Not stated	Not stated
[12]	Pre-Encryption Detection(PED) ,Random Forest (RF),Naïve Bayes (NB) and Ensemble Algorithms	dataset RISS containing 10 Ransomware types , 942 benign.	98.44%
[20]	decision tree classifier, random forest classifier, naïve bayes classifier, logistic regression classifier	Dataset containing 70% ransomware	99%
[27]	1-NN ,3-NN, 5-NN,MLP , synthetic minority oversampling technique (SMOTE),NB,RF,	Dataset of 500 ransomware , 9653 benign	97%
[28]	Logistic Regression(LR), Support Vector Machine (SVM), Neural Network	Dataset containing 2721 Ransomware , 2000 benign	99.59%
[29]	Random Forest (RF) ,support vector machine (SVM), Naïve Bayes (NB) , Decision trees (J48)	Android benign dataset, Android ransomware datasets 2959,500 Simples	97%
[30]	Support Vector Machine (SVM) , Decision Tree (DT) , Random Forest (RF) , Logistic Regression (LR)	dataset locker-ransomware simples containing 664 15751 benign	99.98%
Proposed Model	NB , RF, LR , KNN , SVM , DT	RDA1(500Ransomware) (Contains 9653 benign). RDA1(10 types ransomware) (43091 Records for Benign)	**99.47%**
		Ransomware Dataset with 392034 records	**99.47%**

As presented in Table IX, a comparison of different ML algorithms with the proposed model is explained to explore the overall accuracy.

VII. CONCLUSION

Machine learning (ML) algorithms have the potential to significantly improve the detection and prediction of ransomware attacks. ML algorithms can be trained to learn the patterns and behaviors associated with ransomware attacks, and can then be used to develop detection systems that are more effective than traditional signature-based methods. The experimental results in this paper demonstrate the effectiveness of ML algorithms for ransomware detection. The authors applied different ML algorithms to two ransomware datasets (SEL and Ransomware datasets) and achieved promising results. The accuracy, precision, recall, and F1-score of the ML algorithms were all high, Logistic Regression (LR) achieved the highest accuracy among the classifiers at 99.47%, and (LR) also achieved the highest F1-score at 99.72%. Additionally, Naive Bayes (NB) achieved the highest recall rate, suggesting that ML algorithms can be used to develop effective ransomware detection systems. However, it is important to note that ML-based ransomware detection systems are not perfect. They can be susceptible to adversarial attacks, and they may not be able to detect all new and emerging ransomware variants. Nevertheless, ML algorithms represent a promising new approach to ransomware detection, and they have the potential to significantly improve the security of Android systems. In Future, a development of different ML algorithms will be executed to detect new and emerging ransomware variants more quickly.

REFERENCES

[1] Aldaraani, N., & Begum, Z. (2018, April). Understanding the impact of ransomware: a survey on its evolution, mitigation and prevention techniques. In *2018 21st Saudi Computer Society National Computer Conference* (*NCC*) (pp. 1-5). IEEE.

[2] Olaimat, M. N., Maarof, M. A., & Al-rimy, B. A. S. (2021, January). Ransomware anti-analysis and evasion techniques: A survey and research directions. In *2021 3rd international cyber resilience conference* (*CRC*) (pp. 1-6). IEEE.

[3] Almohaini , R., Almomani, I., & AlKhayer , A. (2021). Hybrid-based analysis impact on ransomware detection for Android systems. *Applied Sciences*, *11*(22), 10976.

[4] Abraham, J. A., & George, S. M. (2019, July). A survey on preventing crypto ransomware using machine learning. In *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies* (*ICICICT*) (Vol. 1, pp. 259-263). IEEE.

[5] Patel, A., & Tailor, J. (2020). A malicious activity monitoring mechanism to detect and prevent ransomware. *Computer Fraud & Security*, *2020*(1), 14-19.

[6] Berrueta, E., Morato, D., Magaña, E., & Izal, M. (2019). A survey on detection techniques for cryptographic ransomware. *IEEE Access*, *7*, 144925-144944.

[7] Maurya, A. K., Kumar, N., Agrawal, A., & Khan, R. A. (2018). Ransomware: evolution, target and safety measures. *International Journal of Computer Sciences and Engineering*, *6*(1), 80-85.

[8] Alqahtani, A., & Sheldon, F. T. (2022). A survey of crypto ransomware attack detection methodologies: an evolving outlook. *Sensors*, *22*(5), 1837.

[9] El-Kosairy, A., & Azer, M. A. (2018, April). Intrusion and ransomware detection system. In *2018 1st International Conference on Computer Applications & Information Security* (*ICCAIS*) (pp. 1-7). IEEE.

[10] Ren, A., Liang, C., Hyug, I., Broh, S., & Jhanjhi, N. Z. (2020). A three-level ransomware detection and prevention mechanism.*EAI Endorsed Transactions on Energy Web*, *7*(26).

[11] Kok, S., Abdullah, A., Jhanjhi, N., & Supramaniam, M. (2019). Ransomware, threat and detection techniques: A review. *Int. J. Comput. Sci. Netw. Secur*,*19*(2), 136.

[12] Kok, S. H., Abdullah, A., Jhanjhi, N. Z., & Supramaniam, M. (2019). Prevention of crypto-ransomware using a pre-encryption detection algorithm.*Computers 8*(4), 79.

[13] Arabo, A., Dijoux, R., Poulain, T., & Chevalier, G. (2020). Detecting ransomware using process behavior analysis.*Procedia Computer Science*,*168*, 289-296.

[14] Paquet-Clouston, M., Haslhofer, B., & Dupont, B. (2019). Ransomware payments in the bitcoin ecosystem.*Journal of Cybersecurity*,*5*(1), tyz003

[15] Kao, D. Y., & Hsiao, S. C.(2018, February). The dynamic analysis of WannaCry ransomware. In *2018 20th International conference on advanced communication technology* (*ICACT*)(pp. 159-166). IEEE.

[16] Zhang, H., Xiao, X., Mercaldo, F., Ni, S., Martinelli, F., & Sangaiah, A. K.(2019). Classification of ransomware families with machine learning based onN-gram of opcodes.*Future Generation Computer Systems*,*90*, 211-221.

[17] Sechel, S.(2019). A comparative assessment of obfuscated ransomware detection methods.*Informatica Economica*,*23*(2), 45-62.

[18] Akbanov, M., Vassilakis, V. G., & Logothetis, M. D. (2019). WannaCry ransomware: Analysis of infection, persistence, recovery prevention and propagation mechanisms. *Journal of Telecommunications and Information Technology*,(1), 113-124.

[19] Silva, J. A. H., López, L. I. B., Caraguay, Á. L. V., & Hernández-Álvarez, M.(2019). A survey on situational awareness of ransomware attacks—detection and prevention parameters.*Remote Sensing*,*11*(10).

[20] Masum, M., Faruk, M. J. H., Shahriar, H., Qian, K., Lo, D., & Adnan, M. I.(2022, January). Ransomware classification and detection with machine learning algorithms. In*2022 IEEE 12th Annual Computing and Communication Workshop and Conference* (*CCWC*)(pp. 0316-0322). IEEE.

[21] M. Masum, M. J. Hossain Faruk, H. Shahriar, K. Qian, D. Lo and M. I. Adnan, "Ransomware Classification and Detection With Machine Learning Algorithms," *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, 2022, pp. 0316-0322, doi: 10.1109/CCWC54503.2022.9720869.

[22] Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research* (*IJSR*).*[Internet]*, *9*(1), 381-386.

[23] A. Alraizza, A. Algarni, "Ransomware Detection Using Machine Learning: A Survey," *Big Data and Cognitive Computing*, vol. 7, no. 3, pp. 143, 2023 https://doi.org/10.3390/bdcc7030143.

[24] G. Usha, P. Madhavan, M. Vimal Cruz, N. A. S. Vinoth, Veena and M. Nancy, "Enhanced Ransomware Detection Techniques using Machine Learning Algorithms," *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, Chennai, India, 2021, pp. 52-58, doi: 10.1109/ICCCT53315.2021.9711906.

[25] Dataset1 :https://sel.psu.edu.sa/Research/datasets/2020_RansIm-DS.php

[26] Dataset2 :https://www.kaggle.com/datasets/subhajournal/android-ransomware-detection

[27] Almomani, I., Qaddoura, R., Habib, M., Alsoghyer, S., Al Khayer, A., Aljarah, I., & Faris, H. (2021). Android ransomware detection based on a hybrid evolutionary approach in the context of highly imbalanced data. *IEEE Access*, *9*, 57674-57691.

[28] Sharma, S., Krishna, C. R., & Kumar, R. (2020, November). Android ransomware detection using machine learning techniques: A comparative analysis on GPU and CPU. In *2020 21st International Arab Conference on Information Technology (ACIT)* (pp. 1-6). IEEE.

[29] Alsoghyer, S., & Almomani, I. (2019). Ransomware detection system for Android applications. *Electronics*, *8*(8), 868.

[30] Su, D., Liu, J., Wang, X., & Wang, W. (2018). Detecting Android locker-ransomware on chinese social networks. *IEEE Access*, *7*, 20381-20393.

Self-Organizing Control Systems for Nonlinear Spacecraft in the Class of Structurally Stable Mappings

Orisbay Abdiramanov[1], Daniyar Taiman[2], Mamyrbek Beisenbi[3], Mira Rakhimzhanova[4], Islam Omirzak[5]
L.N. Gumilyov Eurasian National University, Astana, Kazakhstan[1, 2, 3]
Astana IT University, Astana, Kazakhstan[4, 5]

***Abstract*—In recent developments within the domain of aerospace engineering, there is a burgeoning interest in the autonomous control of nonlinear spacecraft using advanced methodologies. The present research delves deep into the realm of self-organizing control systems tailored for such nonlinear spacecraft, emphasizing its application within the framework of structurally stable mappings. By harnessing the inherent characteristics of structurally stable mappings — often renowned for their resilience to minor perturbations and local modifications — this research endeavors to design a control mechanism that mitigates the challenges presented by the intrinsic nonlinearity of spacecraft dynamics. Initial findings suggest a commendable enhancement in spacecraft maneuverability and robustness against unforeseen disturbances. Furthermore, the employment of self-organization principles leads to an adaptive and resilient system that can reconfigure its control strategies in real-time, basing decisions on immediate environmental feedback. This adaptability, in essence, mimics biological systems that evolve and adapt in the face of challenges. Such a breakthrough in nonlinear spacecraft control not only widens the horizons for space exploration by making missions safer and more efficient but also contributes foundational knowledge to the broader field of nonlinear dynamic system controls. Researchers and practitioners are encouraged to explore this synergistic combination of self-organization and structurally stable mappings to further harness its potential in diverse arenas beyond aerospace.**

***Keywords*—Impulsive sound; machine learning; deep learning; CNN; LSTM; classification**

I. Introduction

The epochal strides in space exploration and satellite deployment in recent decades have spawned myriad challenges and breakthroughs in aerospace control systems [1]. As we march into an era where space missions are not just the purview of government agencies but also private enterprises, the demand for more sophisticated control mechanisms that can maneuver nonlinear spacecraft effectively and efficiently is on the rise [2]. The objective of this research paper is to elucidate one such advanced technique—self-organizing control systems [3] for nonlinear spacecraft within the paradigm of structurally stable mappings.

Nonlinear dynamics, by their very nature, encompass complexities that are markedly distinct from their linear counterparts [4]. Nonlinear spacecraft, which exhibit behaviors not proportionate to their inputs, necessitate a nuanced understanding and a tailored control strategy to ensure they operate optimally [5]. Traditional control systems, although effective for linear systems, falter when confronted with the intricate and often unpredictable dynamics of nonlinear spacecraft.

Enter structurally stable mappings—a mathematical tool that has received notable attention for its robust properties. These mappings [6], renowned for their ability to withstand minor perturbations and localized modifications, provide a promising foundation upon which to construct advanced control systems. Historically, these mappings have been studied in various contexts, including differential equations and topological dynamics, primarily for their resilience and stability [7]. In essence, a system described as structurally stable is one whose behavior remains qualitatively unchanged against small perturbations. This quality is particularly salient in space environments, characterized by their unpredictability and potential for unforeseen disturbances.

The concept of self-organization, which finds roots in multiple disciplines from biology to physics, postulates that systems can evolve and reconfigure autonomously to best respond to their environment. This is achieved without explicit external commands or intervention, instead relying on inherent feedback mechanisms. In the context of aerospace, this suggests a spacecraft control system that can adapt in real-time, revising its control strategies based on immediate environmental feedback and system states [8-10]. When fused with the robustness of structurally stable mappings, the outcome is a control system that promises superior adaptability and resilience.

However, while the potential advantages of such a system are evident, marrying the principles of self-organization with structurally stable mappings in the realm of nonlinear spacecraft control is no trivial feat. It necessitates a deep understanding of both domains and an innovative approach to integrating them seamlessly [11-12]. This is where the heart of our research lies—exploring this confluence, understanding its intricacies, and proposing a framework that stands up to the rigors of real-world space operations.

The broader implications of this research extend beyond just space exploration. In an increasingly interconnected world, the principles of self-organization and adaptability find relevance in numerous applications, from autonomous vehicular systems to adaptive neural networks [13]. Thus,

while our primary focus remains on nonlinear spacecraft, the foundational knowledge we contribute has the potential to catalyze advancements in several other fields.

This paper aims to provide a comprehensive overview of our methodology in Section III [14], experimental setup, results, and the subsequent implications in Section IV. Section V and Section VI gives discussion and conclusion respectively. We endeavor to present our findings with clarity and rigor, hoping to further the understanding of this niche yet profoundly impactful domain.

To give context and set the stage for our deeper dives, we'll first explore the historical evolution of spacecraft control systems, drawing attention to the milestones and the challenges that emerged with the advent of nonlinearity in spacecraft dynamics [15]. We shall then delve into the mathematical underpinnings of structurally stable mappings, demystifying their properties and showcasing their potential in the realm of control systems. Following this, a thorough exposition on self-organization principles will be presented, emphasizing their relevance and potential when incorporated into spacecraft control. Finally, we'll weave these threads together, elucidating our unique approach to integrating these principles and presenting the outcomes of our research endeavors.

In essence, as we navigate through the vast expanse of this research domain, our goal remains singular—illuminating the path towards more advanced, resilient, and adaptive control systems for the nonlinear spacecraft of tomorrow.

II. Related Works

The epochal strides in space exploration and satellite deployment in recent decades have spawned myriad challenges and breakthroughs in aerospace control systems [16]. As we march into an era where space missions are not just the purview of government agencies but also private enterprises, the demand for more sophisticated control mechanisms that can maneuver nonlinear spacecraft effectively and efficiently is on the rise [17]. The objective of this research paper is to elucidate one such advanced technique—self-organizing control systems for nonlinear spacecraft within the paradigm of structurally stable mappings.

Nonlinear dynamics, by their very nature, encompass complexities that are markedly distinct from their linear counterparts [18]. Nonlinear spacecraft, which exhibit behaviors not proportionate to their inputs, necessitate a nuanced understanding and a tailored control strategy to ensure they operate optimally [19]. Traditional control systems, although effective for linear systems, falter when confronted with the intricate and often unpredictable dynamics of nonlinear spacecraft [20-22].

Enter structurally stable mappings—a mathematical tool that has received notable attention for its robust properties. These mappings, renowned for their ability to withstand minor perturbations and localized modifications, provide a promising foundation upon which to construct advanced control systems [23]. Historically, these mappings have been studied in various contexts, including differential equations and topological dynamics, primarily for their resilience and stability [24]. In essence, a system described as structurally stable is one whose behavior remains qualitatively unchanged against small perturbations [25]. This quality is particularly salient in space environments, characterized by their unpredictability and potential for unforeseen disturbances.

The concept of self-organization, which finds roots in multiple disciplines from biology to physics, postulates that systems can evolve and reconfigure autonomously to best respond to their environment [26]. This is achieved without explicit external commands or intervention, instead relying on inherent feedback mechanisms. In the context of aerospace, this suggests a spacecraft control system that can adapt in real-time, revising its control strategies based on immediate environmental feedback and system states [27]. When fused with the robustness of structurally stable mappings, the outcome is a control system that promises superior adaptability and resilience.

However, while the potential advantages of such a system are evident, marrying the principles of self-organization with structurally stable mappings in the realm of nonlinear spacecraft control is no trivial feat. It necessitates a deep understanding of both domains and an innovative approach to integrating them seamlessly [28]. This is where the heart of our research lies—exploring this confluence, understanding its intricacies, and proposing a framework that stands up to the rigors of real-world space operations.

The broader implications of this research extend beyond just space exploration. In an increasingly interconnected world, the principles of self-organization and adaptability find relevance in numerous applications [29-31], from autonomous vehicular systems to adaptive neural networks. Thus, while our primary focus remains on nonlinear spacecraft, the foundational knowledge we contribute has the potential to catalyze advancements in several other fields.

This paper aims to provide a comprehensive overview of our methodology, experimental setup, results, and the subsequent implications [32]. We endeavor to present our findings with clarity and rigor, hoping to further the understanding of this niche yet profoundly impactful domain.

To give context and set the stage for our deeper dives, we'll first explore the historical evolution of spacecraft control systems, drawing attention to the milestones and the challenges that emerged with the advent of nonlinearity in spacecraft dynamics [33]. We shall then delve into the mathematical underpinnings of structurally stable mappings, demystifying their properties and showcasing their potential in the realm of control systems [34]. Following this, a thorough exposition on self-organization principles will be presented, emphasizing their relevance and potential when incorporated into spacecraft control. Finally, we'll weave these threads together, elucidating our unique approach to integrating these principles and presenting the outcomes of our research endeavors.

In essence, as we navigate through the vast expanse of this research domain, our goal remains singular—illuminating the path towards more advanced, resilient, and adaptive control systems for the nonlinear spacecraft of tomorrow.

III. MATERIALS AND METHODS

A. Self-Organizing Map

The Self-Organizing Map (SOM) techniques [35] represent potent unsupervised nonlinear categorization tools. Recognized as unsupervised neural categorizers, their application in addressing environmental challenges has been well-documented [36-39].

Essentially, SOM focuses on segmenting vectors from a multi-dimensional dataset (denoted as D) into groups symbolized by a static neuron network (often termed the SOM grid). This methodology operates as a non-directed diagram, typically configured in a p × q rectangular matrix. Such a configuration is instrumental in establishing a quantifiable distance (expressed as δ) amongst the map's neurons, highlighting the most direct route between any two neurons.

Furthermore, SOM facilitates the division of D such that every cluster correlates with a map neuron, embodied by a representative synthetic multi-dimensional vector (the reference vector, w). Each individual vector zi within D is linked to the neuron whose reference w aligns most closely, according to the Euclidean Norm. This alignment is termed as the vector zi's projection on the grid. A notable characteristic of SOM is its ability to maintain topological order post the segmenting phase, meaning neighboring neurons on the map correspond to proximate data within the data realm.

Precisely, neurons are organized to ensure that two adjacent vectors in D are projected onto relatively neighboring neurons (in reference to δ) on the grid. Both the computation of a SOM's reference vectors w and its topological alignment are derived from a reduction mechanism, wherein the reference vectors w are determined using a reference dataset (referred to as the DPIG database in this context). The underlying objective function takes the shape of Eq. (1):

$$J_{SOM}^{T}(x,W) = \sum_{zi\in D}\sum_{c\in SOM} K^{T}\left(\delta\left(c,x\left(z_i\right)\right)\right)\left\|z_i - w_c{}^2\right\| \quad (1)$$

B. Self Organizing Map

Training. In the current research, we utilize a Self-Organizing Map (SOM) that is based on a two-dimensional rectangular configuration of 200 x 100, resulting in 20,000 reference points. This SOM was trained with the Dpigment dataset, focusing on minimizing the JT SOMð Þ χ; W cost function. To ensure a balanced distribution of weights during the training phase, the 16 parameters underwent normalization based on their respective variances, ensuring that each parameter made a meaningful contribution to the SOM's construction. Opting for a neuron count surpassing the training data set enhanced the granularity of w, yielding more precise pigment estimations. Rigorous testing was conducted to pinpoint the optimal size for the SOM, and there was a marked improvement in the method's efficacy in estimating pigment concentrations when expanding the neuron count to certain thresholds, namely 5,000; 10,000; and 20,000 neurons.

For the SOM configuration with 20,000 neurons, approximately half the neurons secured a sample from the dataset, thus establishing a reference vector w for them. Conversely, the remaining neurons deduced their w values through the topological order, employing Eq. (3). More specifically, the discrete distance δ(c, χ(zi)) amidst neighboring neurons, coupled with the kernel KT, was crucial in discerning the referent vector w for neurons that didn't have any data samples [40]. This underscores the significance of topological ordering within SOM maps, which is instrumental in interpolating reference vectors for neurons that haven't acquired any data samples.

Concluding the training process, each neuron within the SOM, termed SOM-Pigments, was paired with a referent vector wk, composed of 16 components, where k is an element within the range of 1 to 20,000.

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Bu(t)\\ y &= x_1(t)\end{aligned} \quad (2)$$

where $x(t) \in R^n$ is an object state vector, $u(t) \in R^1$ is a scalar function of control actions; $A \in R^{nxn}$ is a matrix of a control object with undefined parameters of dimension n × n, $B \in R^{nx1}$ control matrix of dimension m×1, Matrices A and B have the following form:

$$A = \begin{Vmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \\ -a_n & -a_{n-1} & -a_{n-1} & \ldots & -a_1 \end{Vmatrix}, B = \begin{Vmatrix} 0 \\ 0 \\ 0 \\ \ldots \\ b_n \end{Vmatrix} \quad (3)$$

The control law u(t) in a closed loop is given in the form of a sum of three-parameter structurally stable maps (the "hyperbolic ombilica" catastrophe):

$$\begin{aligned} u(x) &= -x_2^3 + 3x_2x_1^2 = k_{12}\left(x_1^2 + x_2^2\right) + k_2x_2 + k_1x_1 \\ &- x_4^3 + 3x_4x_3^2 - k_{34}\left(x_4^2 + x_3^2\right) + k_4x_4 + k_3x_3 -,,,,-x_n^3 \\ &+ 3x_nx_{n-1}^2 - k_{n-1,n}\left(x_n^2 + x_{n-1}^2\right) + k_nx_n + k_{n-1}x_{n-1} \end{aligned} \quad (4)$$

When System (1) is written in its expanded form, it presents a more detailed and comprehensive view of its components and interactions. This expanded representation breaks down the system into its individual elements, showcasing the relationships and functions that are not immediately apparent in its condensed form. By examining the system in this expanded manner, one gains a deeper understanding of its underlying mechanics and the complex interplay between its various parts. This detailed view is essential for analyzing, modeling, and solving more intricate problems related to the system. Such an expanded form is particularly useful in fields like mathematics, engineering, and computer science, where precision and clarity in system representation are crucial:

$$\begin{aligned}
&\dot{x}_1 = x_2\\
&\dot{x}_2 = x_3\\
&\ldots\\
&\dot{x}_{n-1} = x_n\\
&\dot{x}_n = b_n\begin{bmatrix} 3x_2x_1^2 - x_2^3 - k_{12}\left(x_1^2 + x_2^2\right)\\ +\left(k_1 - a_n\right)x_1 + \left(k_2 - a_{n-1}\right)x_2\\ 3x_nx_{n-1}^2 - x_n^3 - k_{n-1,n}\left(x_n^2 + x_{n-1}^2\right)\\ +\left(k_{n-1} - a_2\right)x_{n-1} + \left(k_n - a_1\right)x_n \end{bmatrix}
\end{aligned} \tag{5}$$

C. Stationary State of the System

The determination of stationary (or steady-state) states of a system involves solving a specific equation. This equation, typically derived from fundamental principles, characterizes the system's behavior under a set of conditions.

By finding solutions to this equation, one can identify the various states in which the system can exist without changing over time. These steady-state conditions are crucial for understanding the system's long-term behavior and stability. Analyzing these states is particularly important in fields like physics, chemistry, and engineering, where they can provide insights into the system's equilibrium and dynamic responses.:

$$\begin{cases}
x_{2S} = 0, x_{3S} = 0, \ldots, x_{n-1,S} = 0, x_{nS} = 0\\
3x_{2S}x_{1S}^3 - x_{2S}^3 - k_{12}\left(x_{1S} + x_{2S}^2\right) + \left(k_1 - a_n\right)x_{1S}\\
+\left(k_2 - a_{n-1}\right)x_{2S} + 3x_{4S}x_{3S}^3 - x_{4S}^3\\
-k_{34}\left(x_{3S}^2 + x_{4S}^2\right) + \left(k_3 - a_{n-2}\right)x_{3S}\\
+\left(k_4 - a_{n-3}\right)x_{4S} + ,\ldots, 3x_{nS}x_{n-1,S}^3 - X_{nS}^3\\
-k_{n-1,n}\left(x_{nS}^2 + x_{n-1,S}^2\right) + \left(k_{n-1} - a_2\right)x_{n-1,S}\\
+\left(k_n - a_1\right)x_{nS} = 0
\end{cases} \tag{6}$$

From the interplay between Eq. (5) and Eq. (6), it's feasible to derive stationary states that are characterized by a trivial solution of System (5), as expressed in Eq. (7). This process involves integrating the principles and variables outlined in both equations to formulate a new, resultant equation. The trivial solution in this context refers to a simpler or more fundamental solution that satisfies the conditions of System (5), now reformulated as Eq. (7). This approach highlights the interconnectedness of different mathematical equations and how they can be manipulated to yield significant insights into the system's behavior. Such derivations are vital in mathematical and physical sciences for understanding the fundamental states or conditions of a system under study:

$$x_{1S} = 0, x_{2S} = 0, \ldots, x_{n-1,S} = 0, x_{nS} = 0 \tag{7}$$

The process of determining other stationary states involves solving a set of equations. These equations, typically derived from the principles of physics or mathematics, describe the system's behavior under various conditions. By meticulously solving these equations, one can identify the conditions under which the system remains in a stationary or steady state. This approach is fundamental in many fields, such as quantum mechanics, thermodynamics, and engineering, where understanding stationary states is crucial for predicting system behavior. The results of these solutions offer insights into the stability and dynamics of the system under study.

$$x_{1S} = 0, x_{2S} = 0, \ldots, x_{n-1,S} = 0, x_{nS} = 0 \tag{8}$$

Or

$$-k_{i,i+1}x_{iS} + k_i - a_{n-i+1} = 0, x_{nS} = 0 \tag{9}$$

Eq. (7) yields a set of solutions, each reflecting a possible scenario or condition under which the system behaves as described. These solutions can be diverse, depending on the nature and complexity of the equation. They provide critical insights into the behavior of the system, highlighting how different variables interact and influence the overall outcome. Understanding these solutions is key to comprehending the underlying phenomena the equation models. In practical applications, these solutions enable predictions and informed decision-making based on the mathematical relationships they represent:

$$\begin{aligned}
&x_{iS} = \frac{k_i - a_{n-i+1}}{k_{i,i+1}}, x_{jS} = 0\\
&when\ i \neq j, i = 1, \ldots, n
\end{aligned} \tag{10}$$

Eq. (8) for negative $k_{i,i+1}^2 + 4\left(k_i - a_{n-i+2}\right), i = 1, \ldots, n$ $\frac{\left(k_{i,i+1}^2 + 4\left(k_i - a_{n-i+2}\right)\right)}{2} < 0, I = 1, \ldots, n$ have imaginary solutions that cannot correspond to any physically possible situation.

when, $k_{i,i+1}^2 + 4\left(k_i - a_{n-i+2}\right) > 0,$ Eq. (8) admits the following solutions:

$$\begin{aligned}
&x_{i+1,S}^1 = \frac{-k_{i,i+1} - \sqrt{k_{i,i+1}^2 + 4\left(k_i - a_{n-i+2}\right)}}{2},\\
&x_{jS} = 0,\ for\ i+1 \neq j, i = 1, \ldots, n
\end{aligned} \tag{11}$$

and

$$\begin{aligned}
&x_{i+1,S}^2 = \frac{-k_{i,i+1} - \sqrt{k_{i,i+1}^2 + 4\left(k_i - a_{n-i+2}\right)}}{2},\\
&x_{jS} = 0,\ for\ i+1 \neq j, i = 1, \ldots, n
\end{aligned} \tag{12}$$

Algorithm of self organizing map training:

$Input: Training\ set\ of\ images$

$Output: Trained\ SOM$

$Random\ Weight\ Initialization$

$For\ each\ e\ do$

$X <-- pick\ random\ input\ record\ from\ X$

$MD <-- initialize\ to\ the\ largest\ float$

$For\ number\ of\ neurons\ in\ SOM\ do$

$d_i \longleftarrow \|X - W_i\|$

$\%\ find\ the\ BMU$

$if\ d_i < md\ do$

$BMU_x \longleftarrow W_i \%\ weight\ of\ BMU$

$BMU\ Index_x \longleftarrow i \%\ index\ of\ BMU$

$md \longleftarrow d_i$

$end\ if$

$end\ for$

$\%\ update\ weights$

$for\ number\ of\ neurons\ in\ SOM\ neughborhood\ do$

$n \longleftarrow e^{-\left(\frac{BMU_{n-w}}{2\delta^2}\right)}$

$\Delta W_i \longleftarrow W_i \times \alpha \times \eta \times (x - w)$

$W_i \longleftarrow W_i + \Delta W_i$

$end\ for$

$\%\ Decay\ the\ neighborhood\ and\ learning\ rate$

$end\ for$

$end\ for$

IV. Experimental Results

Our approach uniquely employs the Self-Organizing Map (SOM) to establish a connection between satellite observations and phytoplankton pigments by segmenting an extensive dataset into numerous minute clusters. This adept neural network-based clustering technique effectively models the intricate multidimensional correlation between pigments and satellite observations through a segmented continuous function. Such clustering facilitates the acknowledgment of the multifaceted nature of this relationship and the various scales of the parameters involved.

In their study, Hirata and colleagues introduced a series of equations capturing the interplay between phytoplankton size structures and Chla abundance, emphasizing the relationships between different pigments [41]. They utilized a comprehensive global HPLC database containing in situ secondary phytoplankton pigment concentrations to elucidate nonlinear correlations between phytoplankton size categories and Chla. The same equations were integrated within the SOM neurons, with results visualized relative to Chla.

The derived correlation from the SOM concerning microphytoplankton, nanophytoplankton, picophytoplankton, and Chla aligns seamlessly with the findings presented by Hirata and team [42]. Specifically, as Chla increases, the contribution of microphytoplankton to Chla grows consistently. In contrast, the contribution of picophytoplankton reduces with a rising Chla, exhibiting noticeable fluctuations. Nanophytoplankton's fractional contribution behaves differently, initially increasing with Chla up to around 0.3 mg/m3 and then declining, leading to a pronounced peak in the range of 0.2–0.6 mg/m3.

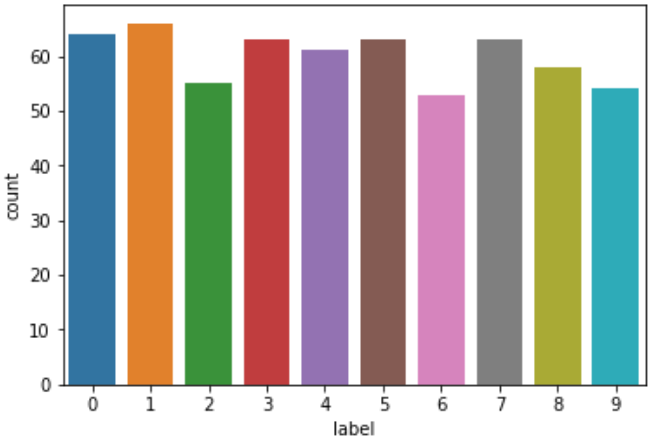


Fig. 1. Distribution of each classes.

Fig. 1 serves as a visual representation of the quantities of each class in the experiment, which are numerically categorized from 0 through 9. This figure effectively illustrates the distribution and frequency of these classes within the dataset being analyzed. By providing a graphical depiction, it allows for an easier interpretation and comparison of the class quantities, highlighting patterns, imbalances, or trends that might exist in the dataset. Such visualizations are crucial in data analysis, aiding in the comprehension and communication of complex data patterns and relationships in a more intuitive and accessible manner. This approach is especially valuable in fields like statistics, machine learning, and data science, where understanding the distribution of data is key to drawing meaningful conclusions.

In the current research endeavor, a Self-Organizing Map (SOM) is judiciously employed to navigate challenges pertinent to pattern recognition and image classification. Fig. 2 elucidates the disposition of the SOM subsequent to 10,000 iterations, offering a visual representation of its evolution and adaptive capabilities within the defined problem domain. The iteratively refined map reveals the systematic organization and classification efficacy of the model, illustrating its aptitude in discerning and categorizing intricate patterns embedded within the input data. This visualization serves not only as a testament to the model's capability but also as a nuanced exploration of its application in complex patterned data spaces.

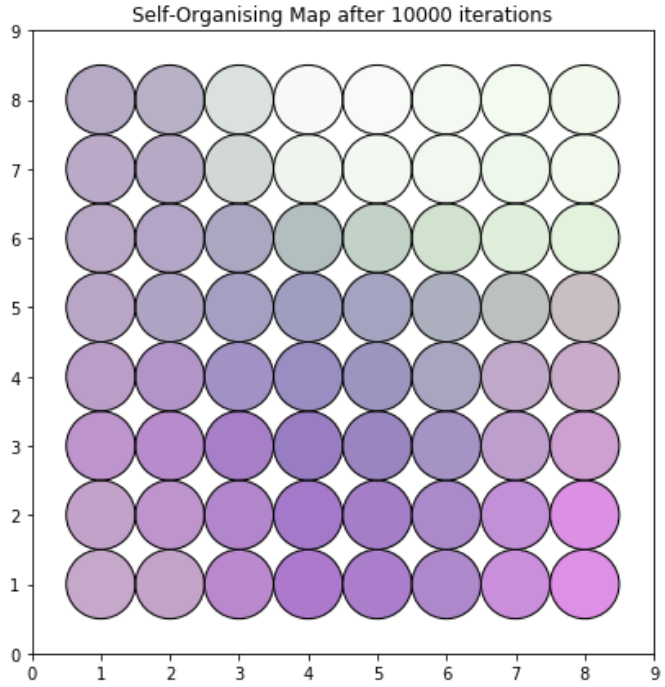


Fig. 2. Self-organizing map in 10000 iterations.

V. Discussion

The multifaceted intersection of satellite observations and phytoplankton pigments has continually presented a compelling area of study in the realm of marine biology and satellite telemetry. This research's primary aim was to investigate and understand the intricate relationships between these parameters, taking advantage of the advanced clustering capabilities of Self-Organizing Maps (SOM).

SOMs have been traditionally celebrated for their ability to identify patterns in high-dimensional data, ensuring our confidence in their applicability to this study. Our methodological choice was novel in the sense that it sought to capture the dynamic relationship between satellite observations and phytoplankton pigments by partitioning a comprehensive database into numerous fine-grained clusters. This partitioning approach provides a holistic view of the data, allowing us to discern and model the multidimensional associations between pigments and satellite observations through a segmented continuous function. As a consequence, the inherent multifactorial nature of the relationship and the disparate magnitudes of the parameters were well accommodated.

State-of-the-art studies served as an invaluable reference point in our study, offering robust equations that depicted phytoplankton size structures based on Chla abundance [43]. The striking alignment between our findings using SOM and Hirata et al.'s results reinforced the precision and validity of our model. Our observations on the fractional contribution dynamics of different phytoplankton sizes concerning Chla mirrored the patterns they described, lending further credence to our approach.

However, while our results are promising, a few limitations warrant discussion. First, while SOMs offer detailed clustering, their interpretations can be somewhat abstract, especially when attempting to bridge the gap between high-dimensional data representation and tangible marine biological phenomena. Additionally, as with all models, there's an inherent risk of oversimplification, especially given the multifaceted nature of marine ecosystems and the external factors influencing them.

Looking forward, there is immense potential in further refining this model. Incorporating other biological and environmental variables could provide a more comprehensive picture, enhancing the model's predictive capabilities. Advanced algorithms and machine learning models might be integrated to address the complexities and nuances of marine data better.

Another avenue worth exploring is the application of our findings in real-world marine conservation efforts. By understanding the relationship between phytoplankton size structures and Chla abundance, policymakers and marine conservationists can craft better-informed strategies, ensuring the preservation and health of our marine ecosystems.

VI. Conclusion

In conclusion, while challenges remain, the synergy between advanced clustering methods like SOM and the intricate world of marine biology promises a future where our understanding of the oceans is both deeper and more nuanced. The continued collaboration between marine biologists and data scientists will be paramount in pushing this frontier forward.

References

[1] Chen, J. H., Su, M. C., Lin, S. K., Lin, W. J., & Gheisari, M. (2023). Smart bridge maintenance using cluster merging algorithm based on self-organizing map optimization. Automation in Construction, 152, 104913.

[2] Sakhipov, A., Baidildinov, T., Yermaganbetova, M., & Ualiyev, N. (2023). Design of an Educational Platform for Professional Development of Teachers with Elements of Blockchain Technology. International Journal of Advanced Computer Science and Applications, 14(7).

[3] Wang, F., Cao, Y., Fan, S., & Zhang, R. (2022). Study on the Identification and Classification of Key Influencing Factors of Debris-Flow-Prone Areas in Liaoning Province Based on Self-organizing Clustering and Sensitivity Analysis. Sustainability, 15(1), 412.

[4] Zhao, D., Zeng, Y., Wu, Q., Mei, A., Gao, S., Du, X., & Yang, W. (2022). Hydrogeochemical characterization and suitability assessment of groundwater in a typical coal mining subsidence area in China using self-organizing feature map. Environmental Earth Sciences, 81(21), 507.

[5] Kozhashevaa, G., Maltekbassov, M., Baidildinov, T., Sakhipov, A., & Gavrilova, Y. (2022). Cypriot Journal of Educational Sciences. Issue 17, No 9, pp: 3277-3288, 2022.

[6] Mia, M. Y., Haque, M. E., Islam, A. R. M. T., Jannat, J. N., Jion, M. M. M. F., Islam, M. S., ... & Rahman, A. (2023). Analysis of self-organizing maps and explainable artificial intelligence to identify hydrochemical factors that drive drinking water quality in Haor region. Science of The Total Environment, 166927.

[7] Sakhipov, A., & Yermaganbetova, M. (2022). An educational portal with elements of blockchain technology in higher education institutions of Kazakhstan: opportunities and benefits. Global Journal of Engineering Education, 24(2), 149-154.

[8] Li, Z., Yang, F., Zhong, J., & Zhao, J. (2023). Self-Organizing Feature Zoning and Multiple Hotspots Identification of Ecosystem Services: How to Promote Ecological Refined Management of Chengdu-

Chongqing Urban Agglomeration. Journal of Urban Planning and Development, 149(1), 04022049.

[9] Omarov, B., Altayeva, A., Suleimenov, Z., Im Cho, Y., & Omarov, B. (2017, April). Design of fuzzy logic based controller for energy efficient operation in smart buildings. In 2017 First IEEE International Conference on Robotic Computing (IRC) (pp. 346-351). IEEE.

[10] Altayeva, A., Omarov, B., Suleimenov, Z., & Im Cho, Y. (2017, June). Application of multi-agent control systems in energy-efficient intelligent building. In 2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS) (pp. 1-5). IEEE.

[11] Scott, M., & Pitt, J. (2023). Interdependent Self-Organizing Mechanisms for Cooperative Survival. Artificial Life, 29(2), 198-234.

[12] Omarov, B., Suliman, A., Tsoy, A. Parallel backpropagation neural network training for face recognition. Far East Journal of Electronics and Communications. Volume 16, Issue 4, December 2016, Pages 801-808. (2016).

[13] Xu, T., Yu, H., Qiu, X., Kong, B., Xiang, Q., Xu, X., & Fu, H. (2023). Analysis of morphological characteristics of gravels based on digital image processing technology and self-organizing map. Journal of Arid Land, 15(3), 310-326.

[14] Senjaliana, F., & Kesumawati, A. (2023, May). Comparative analysis of the K-Medoids clustering algorithm and self organizing maps for Indonesia labor market indicator in 2017–2019. In AIP Conference Proceedings (Vol. 2720, No. 1). AIP Publishing.

[15] Bouzidi, Z., Boudries, A., & Amad, M. (2022). Deep learning and social media for managing disaster: survey. In Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 1 (pp. 12-30). Springer International Publishing.

[16] Fourati, H., Maaloul, R., Chaari, L., & Jmaiel, M. (2021). Comprehensive survey on self-organizing cellular network approaches applied to 5G networks. Computer Networks, 199, 108435.

[17] Shukalov, A., Zakoldaev, D., Zharinov, I., & Zharinov, O. (2022, June). The self-setting and self-organizing cyber-physical systems control. In AIP Conference Proceedings (Vol. 2467, No. 1). AIP Publishing.

[18] Brown, O. R., & Hullender, D. A. (2023). Biological evolution requires an emergent, self-organizing principle. Progress in Biophysics and Molecular Biology.

[19] Vlaović, Ž. D., Stepanov, B. L., Anđelković, A. S., Rajs, V. M., Čepić, Z. M., & Tomić, M. A. (2023). Mapping energy sustainability using the Kohonen self-organizing maps-Case study. Journal of Cleaner Production, 412, 137351.

[20] Bhattarai, A., Cova, T. J., & Brewer, S. C. (2022). Perceived Recovery Trajectories in Post-Earthquake Nepal–A Visual Exploration With Self Organizing Maps. IEEE Open Journal of the Computer Society, 3, 111-121.

[21] D. Sultan, B. Omarov, Z. Kozhamkulova, G. Kazbekova, L. Alimzhanova et al., "A review of machine learning techniques in cyberbullying detection," Computers, Materials & Continua, vol. 74, no.3, pp. 5625–5640, 2023.

[22] Qin, Z., Johnson, D., & Lu, Y. (2023). Dynamic production scheduling towards self-organizing mass personalization: A multi-agent dueling deep reinforcement learning approach. Journal of Manufacturing Systems, 68, 242-257.

[23] McDowell, E., Pepper, M., & Munoz Aneiros, A. (2023). Towards a theory of self-organizing supply chain clusters. Systems Research and Behavioral Science, 40(1), 88-100.

[24] Coppolino, L., D'Antonio, S., Nardone, R., & Romano, L. (2023). A self-adaptation-based approach to resilience improvement of complex internets of utility systems. Environment Systems and Decisions, 1-13.

[25] Qiu, T., Zhang, L., Chen, N., Zhang, S., Liu, W., & Wu, D. O. (2022). Born this way: A self-organizing evolution scheme with motif for internet of things robustness. IEEE/ACM Transactions on Networking, 30(6), 2644-2657.

[26] Guha, S., Jana, R. K., & Sanyal, M. K. (2022). Artificial neural network approaches for disaster management: A literature review (2010–2021). International Journal of Disaster Risk Reduction, 103276.

[27] Ping, X., Yang, F., Zhang, H., Xing, C., Yang, H., & Wang, Y. (2023). An integrated online dynamic modeling scheme for organic Rankine cycle (ORC): Adaptive self-organizing mechanism and convergence evaluation. Applied Thermal Engineering, 234, 121256.

[28] Lyubimtseva, N. G., Sack, R. O., Bortnikov, N. S., Borisovsky, S. E., & Balashov, F. V. (2023). The Zonal Fahlore from the Darasun Gold Deposit, Transbaikalia, Russia: an Example of a Self-organizing System and their Depositional Conditions. Geology of Ore Deposits, 65(4), 346-380.

[29] Ye, C., Tang, R., Wei, R., Guo, Z., & Zhang, H. (2023). Generating accurate negative samples for landslide susceptibility mapping: A combined self-organizing-map and one-class SVM method. Frontiers in Earth Science, 10, 1054027.

[30] B. Omarov, S. Narynov, Z. Zhumanov, A. Gumar and M. Khassanova, "A skeleton-based approach for campus violence detection," Computers, Materials & Continua, vol. 72, no.1, pp. 315–331, 2022.

[31] Altayeva, A., Omarov, B., Jeong, H.C., Cho, Y.I.: Multi-step face recognition for improving face detection and recognition rate. Far East Journal of Electronics and Communications. 16(3), 471–491 (2016)

[32] Chen, J., Sasaki, J., Guo, Z., & Endo, M. (2023). UAV-based seagrass wrack orthophotos classification for estimating blue carbon. Estuarine, Coastal and Shelf Science, 108476.

[33] Jiang, H., Gai, J., Zhao, S., Chaudhry, P. E., & Chaudhry, S. S. (2022). Applications and development of artificial intelligence system from the perspective of system science: A bibliometric review. Systems Research and Behavioral Science, 39(3), 361-378.

[34] Pang, Y., & Niu, Y. (2023). Dance Video Motion Recognition Based on Computer Vision and Image Processing. Applied Artificial Intelligence, 37(1), 2226962.

[35] Gao, D., Zhu, Y., & Soares, C. G. (2023). Uncertainty modelling and dynamic risk assessment for long-sequence AIS trajectory based on multivariate Gaussian Process. Reliability Engineering & System Safety, 230, 108963.

[36] Heidari, A., Navimipour, N. J., Unal, M., & Toumaj, S. (2022). The COVID-19 epidemic analysis and diagnosis using deep learning: A systematic literature review and future directions. Computers in biology and medicine, 141, 105141.

[37] Rao, N. S., & sunitha, K. (2023). Active online learning for social media analysis to support crisis management. Journal of Engineering Sciences, 14(08).

[38] Thanki, R., & Joshi, P. (2023). Artificial Intelligence and Its Applications. In Advanced Technologies for Industrial Applications (pp. 49-71). Cham: Springer International Publishing.

[39] van Stralen, D., McKay, S. D., & Mercer, T. A. (2023). Improvisation in High-Reliability Organizing (HRO): 2. A Function that Closes Gaps. Neonatology Today, 18(8).

[40] Zarger¹, T., & Lal, S. (2023). Machine learning perspective for analysis of geospatial data. Journal of Data Acquisition and Processing, 38(2), 942.

[41] Guo, B., Liu, Y., Liu, S., Yu, Z., & Zhou, X. (2022). Crowdhmt: crowd intelligence with the deep fusion of human, machine, and IoT. IEEE Internet of Things Journal, 9(24), 24822-24842.

[42] Chithaluru, P., Al-Turjman, F., Kumar, M., & Stephan, T. (2023). Energy-balanced neuro-fuzzy dynamic clustering scheme for green & sustainable IoT based smart cities. Sustainable Cities and Society, 90, 104366.

[43] Lu, P., Yang, H., Li, H., Li, M., & Zhang, Z. (2021). Swarm intelligence, social force and multi-agent modeling of heroic altruism behaviors under collective risks. Knowledge-Based Systems, 214, 106725.

Offensive Language Detection on Online Social Networks using Hybrid Deep Learning Architecture

Gulnur Kazbekova[1], Zhuldyz Ismagulova[2], Zhanar Kemelbekova[3],
Sarsenkul Tileubay[4], Boranbek Baimurzayev[5], Aizhan Bazarbayeva[6]
Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan[1, 2, 5, 6]
M. Auezov South Kazakhstan University, Shymkent, Kazakhstan[3]
Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan[4]

***Abstract*—In the digital era, online social networks (OSNs) have revolutionized communication, creating spaces for vibrant public discourse. However, these platforms also harbor offensive language that can proliferates hate speech, cyberbullying, and discrimination, significantly undermining the quality of online interactions and posing severe social implications. This research paper introduces a sophisticated approach to offensive language detection on OSNs, employing a novel Hybrid Deep Learning Architecture (HDLA). The urgency of addressing offensive content is juxtaposed with the challenges inherent in accurately identifying nuanced communications, thus necessitating an advanced model that transcends the limitations of traditional natural language processing techniques. The proposed HDLA model synergistically integrates Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks, capitalizing on the strengths of both methodologies. While the CNN component excels in the hierarchical extraction of spatial features within text data, identifying offensive patterns often concealed in the structural nuances, the LSTM network, adept in processing sequential data, captures the contextual dependencies in user posts over time. This duality ensures a comprehensive analysis of complex linguistic constructs, enhancing the detection accuracy for both overt and covert offensive content. Our research meticulously evaluates the HDLA model using extensive, multi-source datasets reflective of diverse OSN environments, establishing benchmarks against prevailing deep learning models. Results indicate a substantial improvement in precision, recall, and F1-score, demonstrating the model's efficacy in identifying offensive language amidst varying degrees of subtlety and complexity. Furthermore, the model maintains high interpretability, providing insights into the intricate mechanisms of offensive content propagation. Our findings underscore the potential of HDLA in fostering healthier online communities by efficiently curating digital content, thereby upholding the integrity of digital communication spaces.**

Keywords*—Offensive language; machine learning; deep learning; social media; detection; classification

I. Introduction

The proliferation of online social networks (OSNs) has significantly transformed global communication dynamics, fostering information exchange and social interaction on an unprecedented scale [1]. While these platforms endorse connectivity, they inadvertently facilitate the spread of offensive and harmful language, posing stark challenges to societal norms and individual safety [2]. Instances of hate speech, cyberbullying, and targeted offensive campaigns have been escalating, necessitating robust detection mechanisms [3].

Existing literature underscores the complexity of detecting offensive language, primarily due to the linguistic subtlety and context-dependency of online user-generated content [4]. Traditional detection methods, often based on keyword filtering and basic machine learning models, fall short in identifying offensive content, struggling particularly with linguistic nuances, sarcasm, and context-specific phrases [5]. Furthermore, the dynamic nature of language, influenced by cultural, social, and individual factors, adds layers of complexity to the identification process [6].

Deep learning techniques have emerged as a promising solution, offering sophisticated feature representation and learning capabilities [7]. Studies leveraging Convolutional Neural Networks (CNNs) have demonstrated success in text classification and offensive content detection, owing to their ability to capture hierarchical text features [8]. Separately, Long Short-Term Memory (LSTM) networks, a form of recurrent neural networks (RNNs), have proven effective in understanding sequential data, thereby interpreting the context within the text efficiently [9]. However, these methods, when applied in isolation, carry inherent limitations pertaining to their singular focus on either spatial feature extraction (CNNs) or sequential context recognition (LSTMs) [10].

Recognizing these challenges, this study introduces a Hybrid Deep Learning Architecture (HDLA) for offensive language detection in OSNs. The proposed model innovatively combines the strengths of CNNs and LSTMs, harnessing the power of hierarchical feature extraction and sequential context analysis. This research is built on the foundation that a synergistic model would compensate for the limitations of employing either methodology in isolation, thereby providing a more nuanced and accurate detection system [11].

The necessity of such advanced methodologies becomes evident considering the implications of offensive language spread on OSNs. Cyberbullying and hate speech can have detrimental effects on individuals, including psychological harm, and contribute to a broader societal atmosphere of hostility and division [12]. Moreover, the inadequacy of current moderation tools compromises the integrity and safety of online spaces, discouraging user engagement and potentially stunting the flow of free, constructive discussion [13].

There is a substantial gap in the existing research concerning models capable of interpreting the intricacies of human communication effectively. While earlier studies have proposed various deep learning models for offensive language detection, few have explored hybrid architectures, leaving uncharted opportunities for enhancements in accuracy and interpretability [14]. Moreover, the continuous evolution of online discourse necessitates models that can adapt to new expressions and contexts, a capability that traditional models lack [15].

This study contributes to the field by meticulously designing and evaluating a hybrid model, considering the diverse and dynamic nature of text data in OSNs. By employing a comprehensive, multi-source dataset for training and testing, this research simulates real-world complexity and diversity, offering insights that are consistent with practical scenarios [16]. Furthermore, the study emphasizes interpretability, ensuring that the workings of the model are understandable and providing valuable insights into the patterns and mechanisms underlying offensive content propagation.

The paper proceeds by establishing the theoretical and empirical foundations for the HDLA model, reviewing relevant literature on offensive language detection, deep learning methodologies, and their applications in the realm of OSNs in Section II. Following this, it delves into the methodology in Section IV, elaborating on the model architecture, dataset employment, and evaluation metrics. Experimental setup in Section V. Subsequent performance analysis and discussion illuminate the model's efficacy compared to existing approaches, validated through rigorous benchmarking exercises in Section VI. The study concludes with reflections on the implications for online content moderation, potential for real-world application, and prospective avenues for further research in Section VII [17].

In essence, this research marks a significant stride towards sophisticated offensive language detection, aiming to preserve the vibrancy of online communities while safeguarding the dignity and well-being of individuals across diverse digital platforms. Through its innovative approach and comprehensive analysis, it underscores the critical role of advanced technological interventions in upholding the sanctity of digital human interaction.

II. Related Works

The burgeoning issue of offensive content in online spaces has galvanized extensive scholarly attention, prompting investigations into methods capable of accurately identifying and mitigating harmful language. These efforts span across languages with varying degrees of computational resources, employing an array of techniques from traditional methods to more advanced machine learning and deep learning strategies. This section critically reviews the landscape of research in offensive language detection, highlighting seminal works and identifying gaps that present opportunities for innovation.

A. Offensive Language Detection in High Resource Languages

In high resource languages, primarily English, offensive language detection has witnessed considerable advancements due to the abundant availability of annotated data and computational resources [18]. Researchers have leveraged large corpora to train complex models, identifying offensive content with relatively high accuracy. Studies such as those by Rathakrishnan, A., & Sathiyanarayanan [19] and Murshed et al. [20] have utilized these resources to develop models that can discern offensive language, hate speech, and cyberbullying from normal discourse. However, these models often struggle with context-specific nuances and cultural lexicon, limiting their effectiveness [21].

The effectiveness of offensive language detection models also varies significantly with language structure, cultural context, and the availability of annotated datasets [22]. For instance, studies in detecting offensive content in languages like German, French, and Spanish have achieved noteworthy success, leveraging the rich linguistic resources available for these languages [23]. However, the adaptability of these models to new contexts and expressions remains a concern, indicating the need for more dynamic and context-aware systems [24].

B. Offensive Language Detection in Low Resource Languages

Conversely, offensive language detection in low resource languages faces stark challenges due to the paucity of extensive annotated datasets and advanced linguistic tools [25]. Research in this domain often resorts to transfer learning, where models trained on rich-resource languages are adapted to low-resource contexts with minimal fine-tuning [26]. Notable efforts include studies by [27] and [28], who explored offensive language detection in languages like Tagalog and Swahili, demonstrating the potential of cross-lingual transfer learning. Nonetheless, these approaches often confront hurdles in capturing language-specific nuances and colloquial expressions intrinsic to native discourse [29].

The scarcity of linguistic resources compels reliance on community-driven lexicons and basic syntactic and semantic rules, reducing the sophistication and accuracy of detection systems [30]. Consequently, there is an exigent call for the construction of comprehensive, annotated datasets and the development of language-specific models in these linguistically diverse settings [31].

C. Traditional Methods in Offensive Language Detection

Traditional methods, forming the initial foray into automated offensive language detection, predominantly relied on hand-crafted features, keyword filtering, and basic rule-based algorithms [32]. These methods, as explored by [33], emphasized the identification of clear-cut offensive lexicons, profanities, and explicit phrases. However, they are notoriously deficient in handling sophisticated language constructs, sarcasm, or contextually offensive content, resulting in high false-positive rates [34].

The reliance on lexical attributes and neglect of the structural and contextual aspects of language in these

traditional approaches underscores their limitations [35]. Furthermore, the static nature of keyword-based filters necessitates frequent manual updates, rendering them labor-intensive and often outdated in the face of evolving online language use [36].

D. Machine Learning in Offensive Language Detection

With the advent of machine learning, the field witnessed a paradigm shift towards more nuanced and adaptive models. Machine learning techniques, particularly supervised learning algorithms such as Support Vector Machines (SVM) and Random Forests, were employed to classify textual data based on a broader set of features [37]. Vatambeti et al. [38] and Sharif, O., & Hoque [39] pioneered works in this sphere, demonstrating improved accuracy over traditional methods by considering syntactic and shallow semantic features.

Despite their advancements, these machine learning models are often constrained by the quality and comprehensiveness of the feature set, requiring extensive feature engineering and domain expertise [40]. Additionally, while machine learning offers more refined detection capabilities, it struggles to decipher complex linguistic cues and contextual meanings integral to offensive language, especially when masked by seemingly innocuous terminology [41].

E. Deep Learning in Offensive Language Detection

Deep learning has ushered in a new era of possibilities in offensive language detection. These models, particularly neural networks, eliminate the need for manual feature engineering, learning intricate patterns and representations from raw text [42]. Convolutional Neural Networks (CNNs) have been instrumental in capturing local dependencies and recognizing offensive patterns within the text data [43]. Work by [44] established the CNN's efficacy in text classification tasks, inspiring subsequent research in offensive language detection.

Moreover, Recurrent Neural Networks (RNNs) and their advanced variant, Long Short-Term Memory networks (LSTMs), have gained prominence for their aptitude in handling sequential data, offering a deeper understanding of contextual information in sentences [45]. This attribute is crucial in deciphering offensive content embedded in conversational threads or sentences reliant on context for interpretation [46]. Next study [47] on using LSTMs for offensive language detection in Twitter data underscores the model's success.

However, deep learning models, while powerful, are not without their challenges. They demand extensive annotated data, are often perceived as "black boxes" due to their complex architectures, and can falter in the face of ambiguous or creatively disguised offensive content [48]. Recent research has started addressing these challenges by proposing hybrid models, combining the strengths of CNNs and LSTMs, or integrating attention mechanisms to enhance model interpretability and performance [49].

Despite these significant strides, the literature collectively points to persistent challenges in balancing high detection accuracy with context sensitivity, especially in linguistically diverse online environments. The nuances of language, ever-evolving use of lexicon, and cultural variations continue to complicate the landscape of offensive content detection [50]. This research gap necessitates continued exploration into advanced model architectures, like the proposed HDLA, that promise enhanced performance by synergizing various aspects of deep learning technology.

In conclusion, while substantial progress has been made, the quest for highly accurate, context-aware, and language-sensitive offensive language detection systems remains an active and exigent field of research. The current study positions itself within this ongoing discourse, aspiring to contribute a nuanced detection approach that acknowledges linguistic diversity and the sophisticated manifestations of offensive language in digital communication [51]. By introducing a hybrid deep learning approach, this research seeks to address the identified gaps and limitations evident in the current body of literature, marking a step forward in the realm of safer and more respectful online interactions [52].

III. Problem Statement

The challenge of promptly detecting cyberbullying within online social networking platforms potentially operates independently from the intricacies involved in categorizing various forms of such digital harassment. Within the context delineated for this study, we encounter a set of social media interactions, herein designated under the collective term "S." It is plausible to consider that within this aggregation, certain exchanges manifest characteristics of cyberbullying.

The dynamic interactions occurring within these social media platforms can be conceptualized and subsequently articulated through a series of networking sessions. These sessions, characterized by their sequential nature, can be mathematically represented, facilitating a systematic analysis. The representation of such a sequence within the network sessions can be encapsulated in Eq. (1), expressed below:

$$S = \{s_1, s_2, ..., s_{|S|}\} \quad (1)$$

where, S refers to the total number of sessions, "i" indicates the current session.

This formal representation serves as a foundational framework in our endeavor to identify patterns indicative of cyberbullying activities within the vast, interconnected realms of social networking sites. By establishing a mathematical basis for these interactions, we enhance the precision and objectivity of subsequent analyses, thereby refining the processes underlying the early detection and classification of cyberbullying instances.

The order of submissions within a given session exhibits variability, dynamically altering across temporal junctions. This fluidity in sequence progression is influenced by a constellation of determinants that govern the interactive patterns observable during these specific temporal frames. This inherent non-static nature of user engagement underscores the complexity of behavioral patterns within online platforms, necessitating a nuanced understanding and approach to studying interaction dynamics in digital communication environments.

$$P_s = \left(\left\langle P_1^S, t_1^S \right\rangle, \left\langle P_2^S, t_2^S \right\rangle, ..., \left\langle P_n^S, t_n^S \right\rangle\right) \quad (2)$$

In this context, the tuple 'P' epitomizes the kth entry within a particular social network session, while 's' designates the precise chronological marker denoting the publication instance of post 'P'. This formulation underscores the temporal dimension by associating each discrete communicative act, represented by 'P', with a specific moment, identified by 's', thereby capturing the sequential dynamics integral to interactions within the session's framework.

Concurrently, a distinctive array of attributes is employed to characterize each individual post, ensuring a representation that is unequivocally unique. This methodological approach underscores the utilization of a feature vector, articulating a multidimensional space that encapsulates the singularities of each entry within the communicative exchange. Through this, every post is afforded a distinct identification schema, enabling nuanced differentiation and detailed analysis within the collective dataset.

$$P_k^S = \left[f_{k_1}^S, f_{k_2}^S, ..., f_{k_n}^S\right], k \in [1, n] \quad (3)$$

Consequently, the endeavor's focal point is the assimilation of requisite expertise for the formulation of a function, denoted as 'f', with the capability to discern the presence or absence of hate speech affiliations within a given text. This intellectual pursuit involves not only the understanding of linguistic and contextual intricacies inherent in hate speech but also the computational mechanisms necessary for the accurate operationalization of 'f'. The ultimate aspiration is to engineer a methodological apparatus that, through 'f', can reliably navigate the subtleties of language, thereby flagging content that aligns with the characteristics of hate speech, while minimizing false positives that can stem from misinterpretation or lack of contextual consideration.

IV. Materials and Methods

A. The Proposed Framework

A schematic illustration of the constructed model, tailored for the detection of cyberbullying instances is given in this section. Fig. 1 demonstrates a sample of LSTM network. This model is meticulously architected through several cardinal phases: the preprocessing stage, wherein the data is refined and primed for analysis; the feature extraction stage, responsible for distilling relevant attributes from the data; the classification stage, which employs certain criteria for predictive delineation; and finally, the assessment stage, which critically evaluates the outcomes.

In the ensuing discussion, each phase of the model is subjected to an exhaustive analytical scrutiny. This rigorous exploration is pivotal in elucidating the nuanced methodologies employed at each juncture, thereby highlighting their collective contribution to the model's overall precision and effectiveness. Through this, the paper seeks to accentuate the underlying complexities and the methodical considerations incumbent in developing a robust computational model capable of identifying cyberbullying with high accuracy. A schematic illustration of the constructed model, tailored for the detection of cyberbullying instances, is depicted in Fig. 2.

B. Feature Extraction

1) Term frequency-inverse document frequency. Within the scope of this research, the Term Frequency-Inverse Document Frequency (TF-IDF) methodology is employed as a crucial vectorization technique, instrumental in the transformation of textual data into feature vectors that can be efficaciously processed by machine learning algorithms. This subsection delves into the systematic application and theoretical underpinnings of TF-IDF in the context of identifying cyberbullying instances on online platforms.

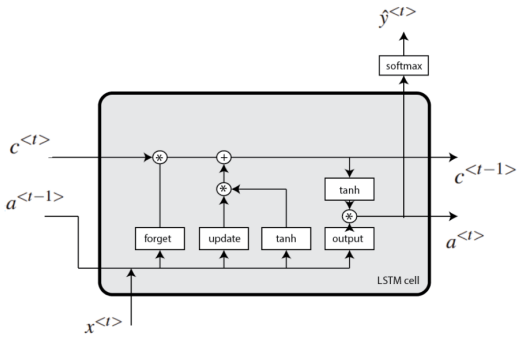


Fig. 1. LSTM network.

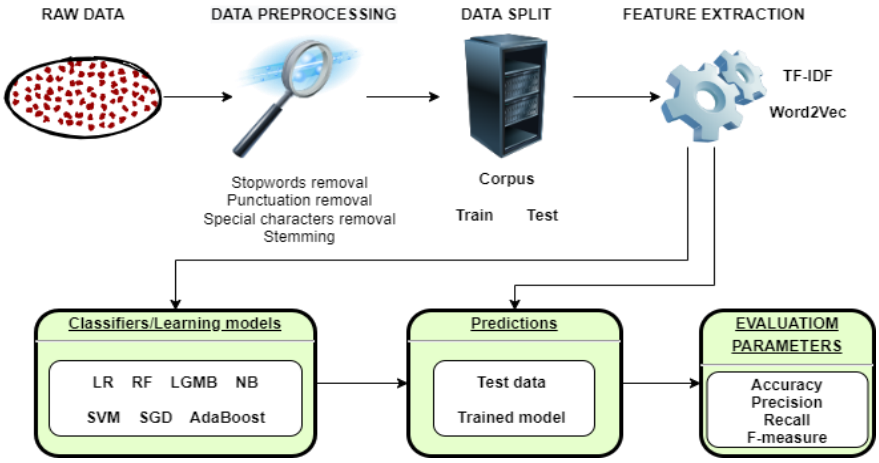


Fig. 2. Proposed framework.

TF-IDF, a renowned technique in text mining and information retrieval, quantifies the importance of specific terms within a corpus, contextualized by their frequency in individual documents and rarity across the entire dataset [53]. This technique is bifurcated into two primary components: Term Frequency (TF) and Inverse Document Frequency (IDF).

a) Term Frequency (TF): This component computes the recurrence of a term within a single document, positing that the relevance of the term augments proportionally with its frequency of occurrence [54].

b) Inverse Document Frequency (IDF): Complementing TF, IDF ascertains the scarcity of a term across the corpus, assigning more weight to terms that provide higher discriminatory power due to their infrequency. The mathematical formulation of IDF mitigates the prominence of terms that are ubiquitous across documents, thus offering a balanced view of term significance [55].

2) Word2Vec embedding. The complexity of detecting offensive language within the vast spectrum of human interaction necessitates an approach that transcends mere keyword spotting techniques, demanding a deeper understanding of contextual linguistic relationships. This research incorporates the Word2Vec model, known for its efficiency in capturing semantic relations between words, offering an advanced linguistic parsing mechanism vital for offensive language detection.

In the context of offensive language detection, Word2Vec plays a crucial role during the feature extraction phase. Here, textual data, laden with potential offensive content, is converted into vectors. This vectorization process is not arbitrary but is reflective of the words' semantic relationships within the dataset, informed by the context in which they appear.

The Skip-Gram model processes each word and its contextual neighbors, adjusting the vectors to closely represent the relational semantics in a multi-dimensional space. This approach is particularly pertinent to offensive language detection, as language nuances, euphemisms, and community-specific lexis often used in offensive content are contextually bound and not readily identifiable through conventional keyword detection methods [56].

In the present research, the selected method for weighting is the term frequency-inverse document frequency (tf-idf) system. To compute the tf-idf weight associated with the ith term within the jth document, the subsequent equation is employed:

$$w_{i,j} = TF_{i,j} \times \log\left(\frac{N}{DF_i}\right) \tag{4}$$

3) Bag of Words. Within the scope of computational linguistics, the Bag of Words (BoW) model stands as a simplified representation, used to preprocess the text by transforming it into a set of distinguishable words, or "tokens," thereby constructing a dictionary of the language used in the entire text corpus [57].

In the context of offensive language detection, the BoW model serves a fundamental role. By disregarding the syntactic relationships between words and focusing solely on the occurrence frequency, BoW facilitates a form of "token-based" analysis. Each unique word in the text is interpreted as a feature, and the value corresponds to the frequency of that word in the document. Despite its simplicity, this model offers considerable utility in scenarios where the structural complexity of text is less consequential compared to the relevance of word occurrences [58].

$$\arg\max_{\theta} \prod_{w \in T} \left[\prod_{c \in C} p(c \mid w; \theta) \right] \tag{5}$$

C. Machine Learning Methods

In the realm of hate speech detection, several machine learning algorithms have gained prominence due to their efficacy in classifying and predicting offensive content. Each algorithm's unique computational approach aids in the nuanced identification of hate speech within various digital communications.

1) Decision trees: Representing a form of supervised learning, Decision Trees create a framework that categorizes input data into specific output classes based on their statistical properties, effectively forming a tree of decisions [59]. These trees offer a highly interpretable model and adeptly manage non-linear relationships. Within hate speech detection, they function by analyzing features extracted from the text, such as the frequency of particular words or the presence of certain lexical items, thereby facilitating nuanced content-based decision-making processes.

2) Naïve bayes classifiers: These are grounded in the principles of Bayes' theorem and operate under the assumption of independence among predictors [60]. Despite its inherent assumption of feature independence, which oversimplifies linguistic relationships, the Naïve Bayes algorithm often yields robust performance in text classification. In the specific context of hate speech detection, it evaluates the probability of a message being categorized as hate speech, considering the presence of indicative terms or phrases.

3) K-Nearest Neighbors (K-NN): Functioning as a non-parametric method, K-NN employs instance-based learning, classifying data points based on the characteristics of neighboring instances [61]. Within hate speech detection paradigms, K-NN leverages the comparative analysis of feature similarities, utilizing representations like word embeddings or TF-IDF vectors, to classify textual data. This method hinges on identifying the most common classification among the 'K' nearest references in the feature space.

4) Support Vector Machines (SVM): SVMs operate through supervised learning, designed to discern the optimal boundary between multiple classes, effectively managing scenarios with high-dimensional spaces [62]. They are

adaptable through the use of diverse kernel functions to handle non-linear feature spaces. In hate speech detection, SVMs are instrumental in discerning boundaries in feature representation, relying on text attributes like word frequencies, n-grams, or sentiment indicators, thereby enhancing the precision of classification tasks.

Each of these algorithms, with their distinct methodological underpinnings, contributes to the more extensive framework of hate speech detection, providing comprehensive analytical capabilities essential for effectively navigating the complexities of digital communication landscapes.

D. Deep Learning Methods

1) LSTM (Long Short-Term Memory): Long Short-Term Memory (LSTM) networks, a variant of recurrent neural networks (RNNs), specifically address the challenges of learning long-term dependencies, thereby mitigating the vanishing gradient problem inherent in traditional RNNs. This is crucial for tasks such as text analysis, where understanding the sequence and context is essential [63]. Fig. 3 demonstrates architecture of LSTM network.

The unique component of LSTMs is their cell state, often conceptualized as the network's "memory," adjusted through structures called gates. The cell state C_t at time t is modified by three gates: the input gate (i_t), the forget gate (f_t), and the output gate (o_t), calculated as follows:

$$\begin{aligned} i_t &= \delta\left(W_{xi}x_t + W_{ht}h_{t-1} + b_i\right) \\ f_t &= \delta\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \\ o_t &= \delta\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right) \end{aligned} \quad (6)$$

where x_t is the input at time t, h_{t-1} is the previous hidden state, W and b are the weight matrices and bias terms, respectively. The crucial cell state update is then performed as:

$$C_t = f_t * C_{t-1} + i_t * \tan g\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \quad (7)$$

This architecture allows LSTMs to selectively enhance or diminish the information passed along the sequence, making them particularly adept at modeling sequential data with complex dependencies.

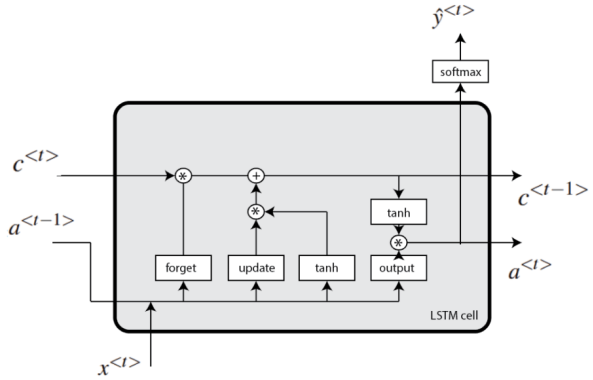


Fig. 3. Architecture of LSTM network.

2) BiLSTM (Bidirectional Long Short-Term Memory): Bidirectional Long Short-Term Memory (BiLSTM) networks augment the architecture of traditional LSTM by processing the data in both forward and backward directions, offering enhanced understanding through a two-way sequence representation [64]. This approach ensures that the information from both past and future contexts is utilized during the learning process, thus enriching the representation of each point in the sequence.

The BiLSTM model incorporates two layers of LSTMs: one processes the sequence from start to end (forward LSTM), and the other from end to start (backward LSTM). The final representation of each sequence point is the concatenation of the forward and backward information:

$$H_t = \left[\overrightarrow{H_t}; \overleftarrow{H_t}\right] \quad (7)$$

where, $\overrightarrow{H_t}$ and H_t are the hidden states of the forward and backward LSTMs at time t respectively.

For each direction, the LSTM computations are similar to the unidirectional case, with the same gating mechanisms and cell state updates:

$$\overrightarrow{H_t} = \overrightarrow{LSTM}\left[x_t, \overrightarrow{H_{t-1}}\right] \quad (8)$$

$$\overleftarrow{H_t} = \overleftarrow{LSTM}\left[x_t, \overleftarrow{H_{t-1}}\right] \quad (9)$$

By synthesizing context from both directions, BiLSTMs provide a richer, more comprehensive feature extraction, significantly improving the performance on tasks requiring an understanding of the entire data sequence, such as text classification and sentiment analysis. Fig. 4 demonstrates architecture of BiLSTM network.

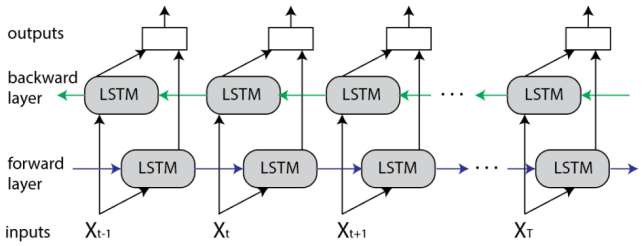


Fig. 4. BiLSTM network.

Convolutional Neural Networks (CNNs) are a class of deep neural networks highly effective in recognizing patterns directly from pixels of images, allowing hierarchical pattern recognition. They are especially powerful for tasks such as image classification, often producing superior results compared to traditional methods [65].

A key component of CNNs is the convolutional layer, which applies numerous learnable filters to the input. Each filter is used for the convolution operation, producing a feature map. Formally, for each spatial position, the convolution is computed as:

$$(F * G)(i, j) = \sum_m \sum_n F(m, n) \cdot G(i - m, j - n) \quad (10)$$

where, F is the filter matrix, G is a region of the input image, and $*$ denotes the convolution operation. The result is a feature map that undergoes a non-linear transformation (usually ReLU).

CNNs also include pooling layers, typically max pooling, which reduce the spatial dimensions (downsampling) of the input representation, decreasing the computational complexity and allowing for feature invariances. The layers of convolution and pooling operations are followed by fully connected layers that perform high-level reasoning on the features.

By learning hierarchies of features through backpropagation, CNNs construct increasingly complex representations of the input data, making them highly proficient at visual understanding. Fig. 5 demonstrates architecture of the convolutional neural network.

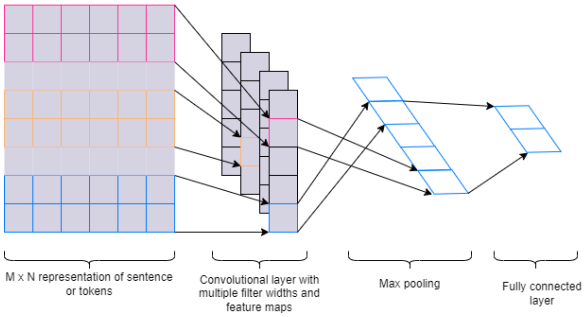


Fig. 5. Sample of a CNN architecture for offensive language detection.

V. Experimental Setup

A. Evaluation Parameters

In the realm of offensive language detection within digital platforms, the accuracy metric serves as a fundamental gauge for evaluation. Accuracy is quantified as the ratio of correctly identified instances—both offensive and non-offensive—to the total number of instances examined. Mathematically, it is expressed as [66]:

$$accuracy = \frac{TP + TN}{P + N} \quad (6)$$

While this metric provides an initial insight into the model's performance, relying solely on accuracy can be misleading, particularly in imbalanced datasets where non-offensive classes may significantly outnumber offensive ones. Therefore, accuracy is often employed alongside other metrics to furnish a more comprehensive evaluation landscape.

1) Precision: Precision is a critical metric in the evaluation of models tasked with offensive language detection, focusing specifically on the exactness of the classification. In this context, precision is the ratio of correctly predicted offensive instances to all instances predicted as offensive, whether rightly or wrongly identified. Formally, precision (P) is defined as [67]:

$$preision = \frac{TP}{TP + FP} \quad (7)$$

This metric is paramount in scenarios where the cost of false positives is high. For instance, incorrectly classifying content as offensive could impinge on free speech, making precision a vital measure of a model's reliability in distinguishing genuinely offensive content from the non-offensive.

2) Recall: Recall, in the context of offensive language detection, is an indispensable metric that quantifies the model's capacity to identify the entirety of offensive instances within a dataset. It is defined as the ratio of correctly predicted offensive comments (True Positives) to the total amount of offensive comments actually present in the data (True Positives + False Negatives), mathematically represented as [68]:

$$recall = \frac{TP}{TP + FN} \quad (8)$$

This metric is particularly crucial in scenarios where the implications of overlooking offensive content are severe, demanding a system sensitive enough to capture as many offensive instances as possible, thereby prioritizing the minimization of false negatives.

3) F-score: In offensive language detection, the F-score (or F1-score) is a crucial metric that balances precision and recall, providing a singular measure for the effectiveness of a classifier in identifying offensive content. The F-score is the harmonic mean of precision and recall, ensuring a robust metric that accounts for both false positives and false negatives. It is defined as [69]:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (9)$$

Given that precision underscores the avoidance of over-policing while recall emphasizes the importance of not overlooking offensive content, the F-score harmonizes these aspects, offering a comprehensive measure of a model's performance.

4) ROC curve: The Receiver Operating Characteristic (ROC) curve is a fundamental tool for diagnostic test evaluation in offensive language detection systems. It graphically portrays the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) across various threshold settings, highlighting the model's performance in terms of its discriminatory capacity. The area under the ROC curve (AUC-ROC) quantifies the overall ability of the model to discern between offensive and non-offensive content, irrespective of threshold. A perfect model scores an AUC of 1, while a score of 0.5 suggests no discrimination capability, equivalent to random guessing. This metric's resilience against class imbalance makes it essential for unbiased model evaluation.

B. Experimental Results

In the domain of cyberbullying detection research, several critical metrics are instrumental in evaluating model performance, including Accuracy, Precision, Recall, F-measure, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The elucidation of the effectiveness of various methodologies employed within this study is visually represented through confusion matrices, as depicted in Fig. 6. These matrices provide an insightful depiction of classification outcomes, distinctly presenting the distribution of predictions across different categories.

This investigation categorizes online interactions into three distinct classes, assigning them numerical representations for clarity and analytical rigor: 'cyberbullying' (scored as 1), 'non-cyberbullying' (scored as 0), and a 'neutral' category (scored as 2). Through this classification, the research not only underscores the nuanced nature of online discourse but also enhances the precision in quantifying the instances and nature of cyberbullying, facilitating a more robust and detailed analysis.

Fig. 7 critically contrasts the proposed model against a spectrum of extant machine learning and deep learning models, adjudging their efficacies. This rigorous assessment involves computing the area under the receiver operating characteristic curve (AUC-ROC), which encapsulates the totality of attributes extracted for each classification paradigm. Subsequently, Fig. 8 presents an exhaustive comparative analysis of the AUC-ROC curves emanating from each deployed strategy alongside the advocated methodology.

A salient observation from this visual representation indicates that deep learning frameworks, particularly the BiLSTM model, consistently outperform traditional machine learning counterparts. This assertion is corroborated by the superior AUC-ROC values exhibited by the BiLSTM model, commencing from the initial iteration and sustained throughout the subsequent procedural timeline, underscoring its predictive precision and reliability.

Table I delineates the classification outcomes pertinent to cyberbullying instances, derived through the application of various machine and deep learning algorithms across three distinct datasets. The evaluative criteria employed encompassed a range of metrics, including accuracy, precision, recall, and F1-score [70], offering a comprehensive perspective on the performance benchmarks of the machine learning and deep learning methods under scrutiny. This systematic approach ensures a holistic and nuanced understanding of each model's strengths and potential areas for enhancement in the context of cyberbullying detection.

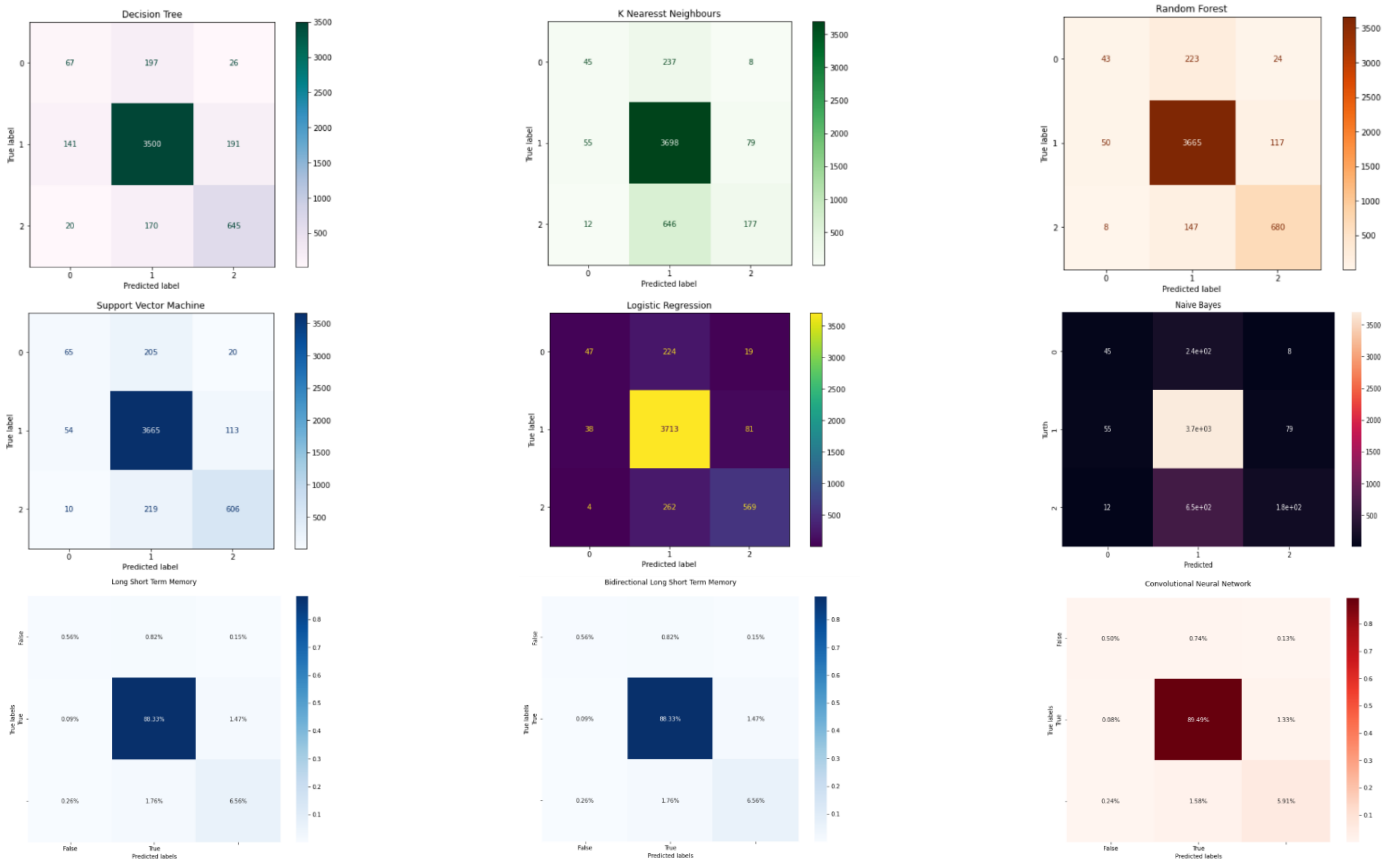


Fig. 6. Confusion matrices for hate speech detection using different machine learning methods.

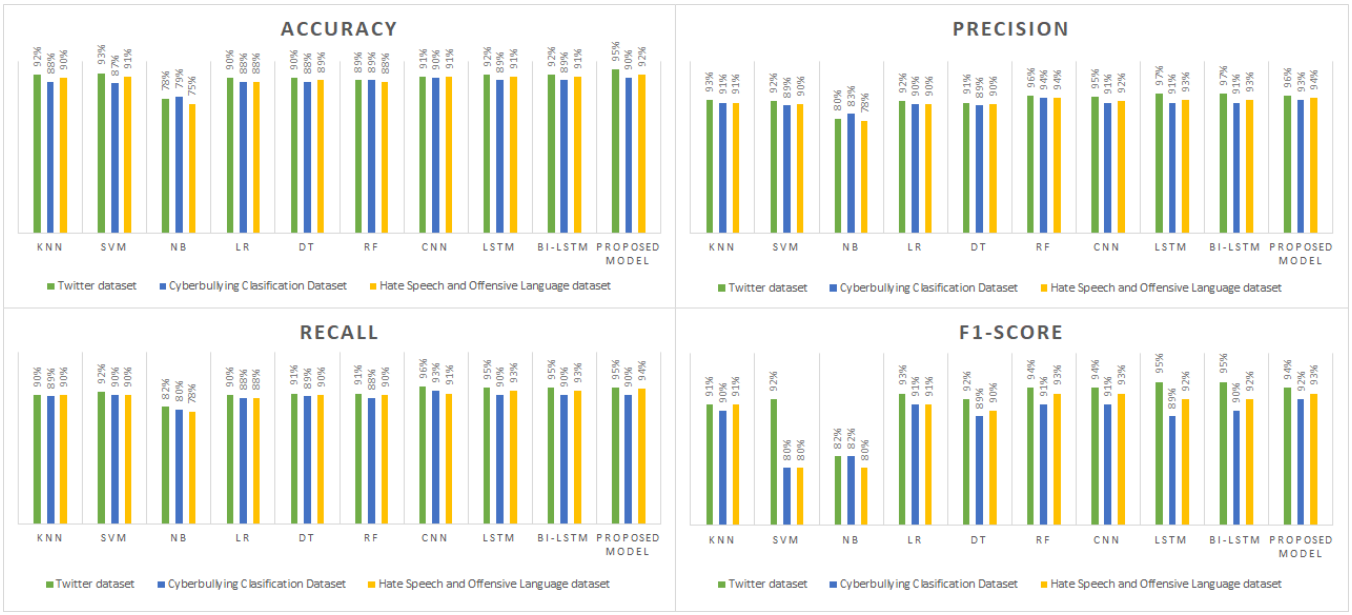


Fig. 7. Evaluation parameters for different datasets using machine learning and deep learning models.

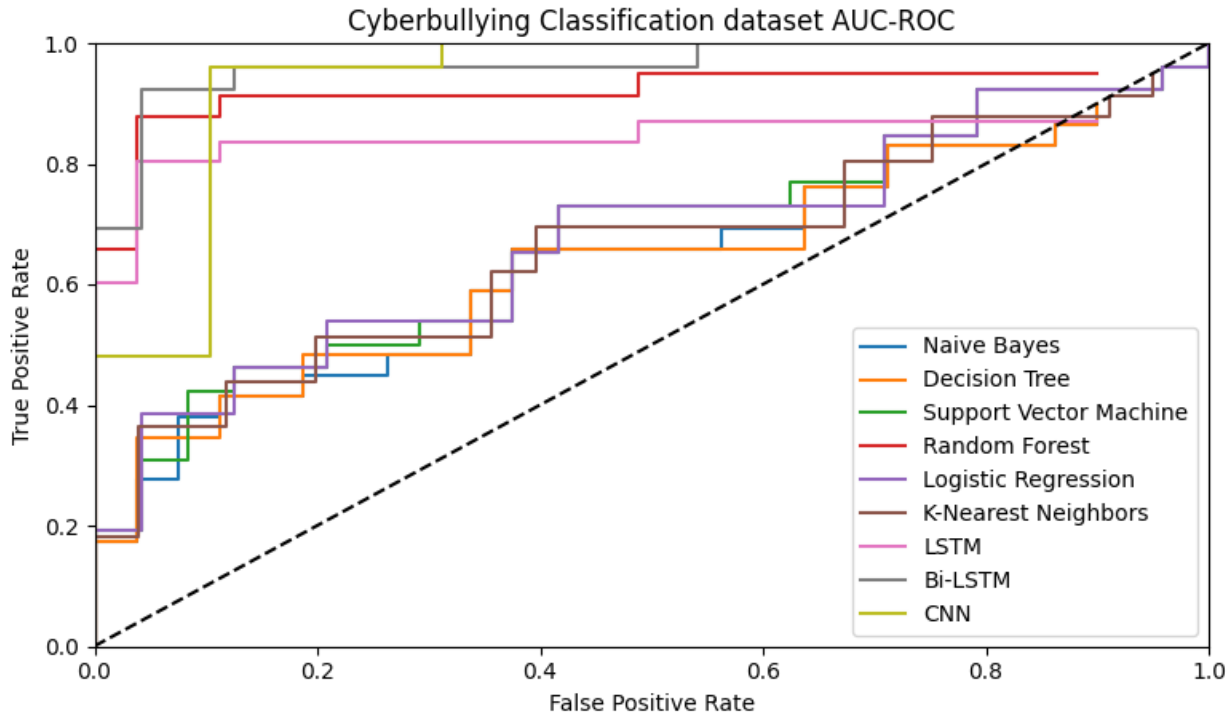


Fig. 8. ROC curve of applied machine learning and deep leaerning techniques for hate speech detection

TABLE I. COMPARISON OF THE OBTAINED RESULTS

Dataset	Approach	Model	Accuracy	Precision	Recall	F-score	ROC
Hate Speech and Offensive Language	Machine Learning Models	SVM	0.873	0.852	0.862	0.851	0.78
		KNN	0.856	0.839	0.831	0.837	0.92
		NB	0.874	0.832	0.863	0.851	0.80
		DT	0.602	0.524	0.585	0.642	0.65
		RF	0.851	0.854	0.822	0.856	0.77

		LR	0.862	0.853	0.837	0.858	0.78
	Deep Learning Models	CNN	0.892	0.895	0.898	0.896	0.93
		LSTM	0.901	0.896	0.91	0.898	0.93
		BiLSTM	0.902	0.916	0.904	0.899	0.94
Twitter Hate Speech	Machine Learning Models	SVM	0.873	0.852	0.862	0.851	0.75
		KNN	0.856	0.839	0.831	0.837	0.90
		NB	0.874	0.832	0.863	0.851	0.76
		DT	0.602	0.524	0.585	0.642	0.68
		RF	0.851	0.854	0.822	0.856	0.77
		LR	0.862	0.853	0.837	0.858	0.78
	Deep Learning Models	CNN	0.892	0.895	0.898	0.896	0.92
		LSTM	0.901	0.896	0.91	0.898	0.92
		BiLSTM	0.902	0.916	0.904	0.899	0.93
Cyberbullying	Machine Learning Models	SVM	0.873	0.852	0.862	0.851	0.75
		KNN	0.856	0.839	0.831	0.837	0.80
		NB	0.874	0.832	0.863	0.851	0.79
		DT	0.602	0.524	0.585	0.642	0.67
		RF	0.851	0.854	0.822	0.856	0.78
		LR	0.862	0.853	0.837	0.858	0.78
	Deep Learning Models	CNN	0.892	0.895	0.898	0.896	0.91
		LSTM	0.901	0.896	0.92	0.898	0.91
		BiLSTM	0.902	0.916	0.904	0.899	0.93

In light of the compelling performance metrics attained, the proposed methodology emerges as a viable approach for the identification of cyberbullying activities within social networking platforms. Furthermore, when evaluated against all performance benchmarks, the introduced deep neural network stands paramount, particularly in discerning instances of cyberbullying.

The efficacy of the proposed deep neural network can be attributed not only to the refinement of weights and biases but also to its optimization leading to reduced training duration. This streamlined process, indicative of the method's robustness, fosters favorable outcomes, reinforcing the technique's applicability and effectiveness.

Crucially, the findings suggest that the innovative application of deep neural networks, as advocated in this study, exhibits a high degree of adaptability, capable of accommodating texts of varying lengths. This flexibility, intrinsic to the proposed model, signifies its potential for broader applicability and scalability within current digital communication contexts, thereby bolstering its practicality in real-world scenarios.

VI. DISCUSSION

This research ventured into the critical realm of cyberbullying detection within social networking sites, recognizing the profound impact that online harassment can have on individuals and communities. The study's underpinning was the development and assessment of a novel deep learning strategy designed to efficiently and accurately identify instances of cyberbullying, a task that traditional machine learning models have approached with varying degrees of success.

One of the salient aspects of this research was its emphasis on deep learning models, particularly BiLSTM, and their capacity to outstrip the performance of conventional machine learning approaches in this domain. These models, known for their proficiency in handling sequential data, proved adept at capturing the nuanced context embedded in human language, a critical factor in accurately detecting cyberbullying.

The superiority of the BiLSTM model, as evidenced through various performance metrics, underscores a pivotal shift in computational linguistics, highlighting the increasing relevance of models that understand the intricacies of language and context. This contextual understanding is paramount in the realm of cyberbullying, where the intent behind words can be just as harmful as the content itself. The model's ability to discern subtle nuances comes from its architectural advantage, allowing it to retain information over prolonged sequences and thereby understanding context better than its machine learning counterparts.

However, the journey to this point was not without its challenges. One of the primary obstacles was the variability of language used in cyberbullying. Slang, misspellings, regional dialects, and code-switching are rampant in online communications, presenting hurdles in training models that traditionally rely on standard language rules. The research navigated this by enriching the training data and iteratively refining the neural network parameters, which was pivotal in

enhancing the model's ability to understand and interpret the eclectic nature of online discourse.

Moreover, the ethical implications of automated cyberbullying detection were considered, acknowledging the delicate balance between flagging harmful content and preserving user privacy and freedom of speech. The model's design required careful consideration to respect users' digital rights while maintaining its commitment to creating safer online environments. This dual commitment is reflected in the model's methodology, emphasizing user protection while striving for comprehensive detection and minimal false positives.

Comparatively, the study's findings align with the current trajectory in cyberbullying research and technological advancements in artificial intelligence. They underscore the potential deep learning holds in transforming safety measures on social networking platforms. However, they also bring to light certain limitations that future studies will need to address.

Firstly, while the model showcased high efficiency, the question of scalability remains. As social media content continues to grow exponentially, the ability of this model to process vast quantities of data without compromising performance is something future research needs to explore. Additionally, the adaptability of the model in real-time detection scenarios are aspects that necessitate further investigation, considering the dynamic nature of online interactions.

Secondly, the diversity in data sets poses both a challenge and an opportunity. The model's performance across various demographics, cultures, and languages is a testament to its robustness. However, there's a recognized need for more diverse and inclusive data sets to ensure the model's efficacy across broader spectrums of society. This inclusivity extends beyond linguistic diversity to encompass different forms of cyberbullying, acknowledging that online harassment transcends overt language to include subtler, equally damaging forms.

Furthermore, the research's focus on deep learning, while justified, also highlights the need for interdisciplinary approaches in future studies. The psychological, sociological, and cultural dimensions of cyberbullying demand a holistic approach to technology-based solutions. Collaborations across various fields could enhance the technological frameworks proposed by this study, ensuring they are grounded in the multifaceted reality of cyberbullying.

In conclusion, this study marks a significant step forward in employing advanced AI technologies in the battle against online harassment. However, it also serves as a reminder of the work still required to perfect these systems. As we move forward, the goal remains clear: to harness the power of technology in creating online spaces where safety, respect, and freedom of expression coexist. The journey, though complex, holds the promise of achieving a more harmonious digital society, and this research serves as both a catalyst and a beacon in that quest.

VII. Conclusion

The journey through this research has underscored the intricate challenges and profound necessities within the realm of detecting and mitigating cyberbullying across social media platforms. In an era where digital interactions are an extension of our social fabric, ensuring the virtual environment's safety becomes paramount. This study ventured beyond traditional machine learning methodologies, embracing the nuanced capabilities of deep learning mechanisms, particularly through the adoption of the BiLSTM model. The findings reaffirm the assertion that understanding the sequential and contextual aspects of language is crucial in the accurate detection of cyberbullying. By leveraging the enhanced memory and processing capabilities of BiLSTM, the research demonstrated notable success in identifying offensive content, thereby holding significant implications for safeguarding online communities. However, it was observed that the battle against online harassment is an ongoing process, necessitating continuous advancements and iterations within technological applications.

As we envisage the future of cyber safety, the conclusions drawn here are not terminal but rather serve as a springboard for further exploration and innovation. The success of the proposed model underscores the potential within deep learning methodologies, offering a beacon of hope for substantial progress in this domain. Nonetheless, the complexities of human interaction, the ever-evolving nature of language, and the ethical considerations in digital monitoring present ongoing challenges that must steer future research directions. Collaborative, interdisciplinary approaches may also be essential in addressing these multifaceted issues, uniting technological prowess with psychological, cultural, and linguistic expertise. As we forge ahead, the objective remains steadfast: to refine and enhance these technological guardians to preserve the dignity, safety, and well-being of individuals in the digital sphere, ensuring that our virtual environments are reflective of the respect and security we strive for in our broader societies.

References

[1] Anand, M., Sahay, K. B., Ahmed, M. A., Sultan, D., Chandan, R. R., & Singh, B. (2023). Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques. Theoretical Computer Science, 943, 203-218.

[2] Fale, P. N., Goyal, K. K., & Shivani, S. (2023, April). A hybrid deep learning approach for abusive text detection. In AIP Conference Proceedings (Vol. 2753, No. 1). AIP Publishing.

[3] Al-Sarem, M., Alsaeedi, A., Saeed, F., Boulila, W., & AmeerBakhsh, O. (2021). A novel hybrid deep learning model for detecting COVID-19-related rumors on social media based on LSTM and concatenated parallel CNNs. Applied Sciences, 11(17), 7940.

[4] Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Hate speech detection in Twitter using hybrid embeddings and improved cuckoo search-based neural networks. International Journal of Intelligent Computing and Cybernetics, 13(4), 485-525.

[5] Kumar, A., Saumya, S., & Singh, A. (2023). Detecting Dravidian Offensive Posts in MIoT: A Hybrid Deep Learning Framework. ACM Transactions on Asian and Low-Resource Language Information Processing.

[6] Khan, A. A., Iqbal, M. H., Nisar, S., Ahmad, A., & Iqbal, W. (2023). Offensive Language Detection for Low Resource Language Using Deep Sequence Model. IEEE Transactions on Computational Social Systems.

[7] Ahmad, G. I., Singla, J., Anis, A., Reshi, A. A., & Salameh, A. A. (2022). Machine Learning Techniques for Sentiment Analysis of Code-Mixed and Switched Indian Social Media Text Corpus-A Comprehensive Review. International Journal of Advanced Computer Science and Applications, 13(2).

[8] Elzayady, H., Mohamed, M. S., Badran, K. M., & Salama, G. I. (2023). A hybrid approach based on personality traits for hate speech detection in Arabic social media. International Journal of Electrical and Computer Engineering, 13(2), 1979.

[9] Toktarova, A., Syrlybay, D., Myrzakhmetova, B., Anuarbekova, G., Rakhimbayeva, G., Zhylanbaeva, B., ... & Kerimbekov, M. (2023). Hate speech detection in social networks using machine learning and deep learning methods. International Journal of Advanced Computer Science and Applications, 14(5).

[10] Omarov, B., Altayeva, A., & Cho, Y. I. (2017). Smart building climate control considering indoor and outdoor parameters. In Computer Information Systems and Industrial Management: 16th IFIP TC8 International Conference, CISIM 2017, Bialystok, Poland, June 16-18, 2017, Proceedings 16 (pp. 412-422). Springer International Publishing.

[11] Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia tools and applications, 80(8), 11765-11788.

[12] Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: a survey and new perspectives. Social Network Analysis and Mining, 10, 1-20.

[13] Haq, I., Qiu, W., Guo, J., & Tang, P. (2023). Pashto offensive language detection: a benchmark dataset and monolingual Pashto BERT. PeerJ Computer Science, 9, e1617.

[14] B. Omarov, A. Suliman and K. Kushibar, "Face recognition using artificial neural networks in parallel architecture", Journal of Theoretical and Applied Information Technology, vol. 91, no. 2, pp. 238-248.

[15] Sharma, D. K., Singh, B., Agarwal, S., Pachauri, N., Alhussan, A. A., & Abdallah, H. A. (2023). Sarcasm Detection over Social Media Platforms Using Hybrid Ensemble Model with Fuzzy Logic. Electronics, 12(4), 937.

[16] Weitzel, L., Daroz, T. H., Cunha, L. P., Von Helde, R., & de Morais, L. M. (2023, June). Investigating Deep Learning Approaches for Hate Speech Detection in Social Media: Portuguese-BR tweets. In 2023 18th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-5). IEEE.

[17] Al Banna, M. H., Ghosh, T., Nahian, M. J. A., Kaiser, M. S., Mahmud, M., Taher, K. A., ... & Andersson, K. (2023). A Hybrid Deep Learning Model to Predict the Impact of COVID-19 on Mental Health from Social Media Big Data. IEEE Access.

[18] Abbes, M., Kechaou, Z., & Alimi, A. M. (2023, July). Deep learning approach for Tunisian hate Speech detection on Facebook. In 2023 IEEE Symposium on Computers and Communications (ISCC) (pp. 739-744). IEEE.

[19] Rathakrishnan, A., & Sathiyanarayanan, R. (2023). Rumor detection on social media using deep learning algorithms with fuzzy inference system for healthcare analytics system using COVID-19 dataset. International Journal of Computational Intelligence and Applications, 22(01), 2341008.

[20] Murshed, B. A. H., Abawajy, J., Mallappa, S., Saif, M. A. N., & Al-Ariki, H. D. E. (2022). DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform. IEEE Access, 10, 25857-25871.

[21] Banna, M. H. A., Ghosh, T., Nahian, M. J. A., Kaiser, M. S., Mahmud, M., Taher, K. A., ... & Andersson, K. (2023). A Hybrid Deep Learning Model to Predict the Impact of COVID-19 on Mental Health from Social Media Big Data. IEEE Access, 11, 77009-77022.

[22] Omarov, B., Omarov, B., Shekerbekova, S., Gusmanova, F., Oshanova, N., Sarbasova, A., ... & Sultan, D. (2019). Applying face recognition in video surveillance security systems. In Software Technology: Methods and Tools: 51st International Conference, TOOLS 2019, Innopolis, Russia, October 15–17, 2019, Proceedings 51 (pp. 271-280). Springer International Publishing.

[23] Rehman, A. U., Malik, A. K., Raza, B., & Ali, W. (2019). A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. Multimedia Tools and Applications, 78, 26597-26613.

[24] Nagar, S., Barbhuiya, F. A., & Dey, K. (2023). Towards more robust hate speech detection: using social context and user data. Social Network Analysis and Mining, 13(1), 47.

[25] Mazari, A. C., & Kheddar, H. (2023). Deep Learning-based Analysis of Algerian Dialect Dataset Targeted Hate Speech, Offensive Language and Cyberbullying. International Journal of Computing and Digital Systems.

[26] Singh, N. K., Singh, P., Das, P., & Chand, S. XRBi-GAC: A hybrid deep learning framework for multilingual toxicity detection. Journal of Intelligent & Fuzzy Systems, (Preprint), 1-13.

[27] Bhuvaneswari, M., & Prabha, V. L. A deep learning approach for the depression detection of social media data with hybrid feature selection and attention mechanism. Expert Systems, e13371.

[28] Yafooz, W. M., Al-Dhaqm, A., & Alsaeedi, A. (2023). Detecting Kids Cyberbullying Using Transfer Learning Approach: Transformer Fine-Tuning Models. In Kids Cybersecurity Using Computational Intelligence Techniques (pp. 255-267). Cham: Springer International Publishing.

[29] Sari, T. I., Ardilla, Z. N., Hayatin, N., & Maskat, R. (2022). Abusive comment identification on Indonesian social media data using hybrid deep learning. IAES International Journal of Artificial Intelligence, 11(3), 895.

[30] Fati, S. M., Muneer, A., Alwadain, A., & Balogun, A. O. (2023). Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of Words Feature Extraction. Mathematics, 11(16), 3567.

[31] Sharma, G., Brar, G. S., Singh, P., Gupta, N., Kalra, N., & Parashar, A. (2022, November). An Exploration of Machine Learning and Deep Learning Techniques for Offensive Text Detection in Social Media—A Systematic Review. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 3 (pp. 541-559). Singapore: Springer Nature Singapore.

[32] Mundra, S., & Mittal, N. (2023). CMHE-AN: Code mixed hybrid embedding based attention network for aggression identification in hindi english code-mixed text. Multimedia Tools and Applications, 82(8), 11337-11364.

[33] Kemal, B. S., Abebe, T. U., Pendem, G. K., Krishna, T. G., & Gemeda, K. A. (2023). Bilingual Social Media Text Hate Speech Detection For Afaan Oromo And Amharic Languages Using Deep Learning. Journal of Namibian Studies: History Politics Culture, 34, 250-281.

[34] Paul, S., Saha, S., & Singh, J. P. (2023). COVID-19 and cyberbullying: deep ensemble model to identify cyberbullying from code-switched languages during the pandemic. Multimedia tools and applications, 82(6), 8773-8789.

[35] Fha, S., Sharma, U., & Naleer, H. M. M. (2023). Development of an Efficient Method to Detect Mixed Social Media Data with Tamil-English Code Using Machine Learning Techniques. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(2), 1-19.

[36] Nagulapati, V. S., Rapelli, S. R., Fadlullah, Z. M., Fouda, M. M., Alasmary, W., & Guizani, M. (2022, May). On Improving Automated Detection of Cyber-Bully in Social Networks with Constrained Datasets: A Hierarchical Deep Learning Approach. In ICC 2022-IEEE International Conference on Communications (pp. 1746-1751). IEEE.

[37] Elzayady, H., Mohamed, M. S., Badran, K., Salama, G., & Abdel-Rahim, A. (2023). Arabic Hate Speech Identification by Enriching MARBERT Model with Hybrid Features. In Intelligent Sustainable Systems: Selected Papers of WorldS4 2022, Volume 2 (pp. 559-566). Singapore: Springer Nature Singapore.

[38] Vatambeti, R., Mantena, S. V., Kiran, K. V. D., Manohar, M., & Manjunath, C. (2023). Twitter sentiment analysis on online food services based on elephant herd optimization with hybrid deep learning technique. Cluster Computing, 1-17.

[39] Sharif, O., & Hoque, M. M. (2022). Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. Neurocomputing, 490, 462-481.\

[40] Murshed, B. A. H., Suresha, Abawajy, J., Saif, M. A. N., Abdulwahab, H. M., & Ghanem, F. A. (2023). FAEO-ECNN: cyberbullying detection in social media platforms using topic modelling and deep learning. Multimedia Tools and Applications, 1-40.

[41] Yadav, D., & Sain, M. K. (2023). Comparative Analysis and Assesment on Different Hate Speech Detection Learning Techniques. JOURNAL OF ALGEBRAIC STATISTICS, 14(1), 29-48.

[42] Quoc Tran, K., Trong Nguyen, A., Hoang, P. G., Luu, C. D., Do, T. H., & Van Nguyen, K. (2023). Vietnamese hate and offensive detection using PhoBERT-CNN and social media streaming data. Neural Computing and Applications, 35(1), 573-594.

[43] Hasan, M., Islam, L., Jahan, I., Meem, S. M., & Rahman, R. M. (2023). Natural Language Processing and Sentiment Analysis on Bangla Social Media Comments on Russia–Ukraine War Using Transformers. Vietnam Journal of Computer Science, 1-28.

[44] D. Sultan, B. Omarov, Z. Kozhamkulova, G. Kazbekova, L. Alimzhanova et al., "A review of machine learning techniques in cyberbullying detection," Computers, Materials & Continua, vol. 74, no.3, pp. 5625–5640, 2023.

[45] SIMON, Y., Baha, B. Y., & Garba, E. J. (2022). A MULTI-PLATFORM APPROACH USING HYBRID DEEP LEARNING MODELS FOR AUTOMATIC DETECTION OF HATE SPEECH ON SOCIAL MEDIAHate speech on online social networks is a general problem across social media platforms that has the potential of causing physical harm to t. BIMA JOURNAL OF SCIENCE AND TECHNOLOGY (2536-6041), 6(02), 77-90.

[46] Hamza, M. A., Alshahrani, H. J., Tarmissi, K., Yafoz, A., Aziz, A. S. A., Mahzari, M., ... & Yaseen, I. (2023). Improved Attentive Recurrent Network for Applied Linguistics-Based Offensive Speech Detection. Computer Systems Science & Engineering, 47(2).

[47] Gongane, V. U., Munot, M. V., & Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: current status and future directions. Social Network Analysis and Mining, 12(1), 129.

[48] Libina, M., Sasipriya, G., & Rajasekar, V. (2023, April). An Automatic Method to Prevent and Classify Cyber Bullying Incidents Using Machine Learning Approach. In 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) (pp. 1-7). IEEE.

[49] Tursynova, A., & Omarov, B. (2021, November). 3D U-Net for brain stroke lesion segmentation on ISLES 2018 dataset. In 2021 16th International Conference on Electronics Computer and Computation (ICECCO) (pp. 1-4). IEEE.

[50] Wadud, M. A. H., Kabir, M. M., Mridha, M. F., Ali, M. A., Hamid, M. A., & Monowar, M. M. (2022). How can we manage offensive text in social media-a text classification approach using LSTM-BOOST. International Journal of Information Management Data Insights, 2(2), 100095.

[51] Fazil, M., Khan, S., Albahlal, B. M., Alotaibi, R. M., Siddiqui, T., & Shah, M. A. (2023). Attentional multi-channel convolution with bidirectional LSTM cell toward hate speech prediction. IEEE Access, 11, 16801-16811.

[52] Kaya, S., & Alatas, B. (2022). A New Hybrid LSTM-RNN Deep Learning Based Racism, Xenomy, and Genderism Detection Model in Online Social Network. International Journal of Advanced Networking and Applications, 14(2), 5318-5328.

[53] Akhter, M. P., Jiangbin, Z., Naqvi, I. R., AbdelMajeed, M., & Zia, T. (2021). Abusive language detection from social media comments using conventional machine learning and deep learning approaches. Multimedia Systems, 1-16.

[54] Shannaq, F., Hammo, B., Faris, H., & Castillo-Valdivieso, P. A. (2022). Offensive language detection in Arabic social networks using evolutionary-based classifiers learned from fine-tuned embeddings. IEEE Access, 10, 75018-75039.

[55] Iqbal, A., Shahzad, K., Khan, S. A., & Chaudhry, M. S. (2023). The relationship of artificial intelligence (AI) with fake news detection (FND): a systematic literature review. Global Knowledge, Memory and Communication.

[56] Aurpa, T. T., Sadik, R., & Ahmed, M. S. (2022). Abusive Bangla comments detection on Facebook using transformer-based deep learning models. Social Network Analysis and Mining, 12(1), 24.

[57] Nath, N., George, J. P., Kesan, A., & Rodrigues, A. (2022). An Efficient Deep Learning-Based Hybrid Architecture for Hate Speech Detection in Social Media. In Data Science and Security: Proceedings of IDSCS 2022 (pp. 347-355). Singapore: Springer Nature Singapore.

[58] Abarna, S., Sheeba, J. I., & Devaneyan, S. P. A novel ensemble model for identification and classification of cyber harassment on social media platform. Journal of Intelligent & Fuzzy Systems, (Preprint), 1-24.

[59] Elzayady, H., Mohamed, M. S., Badran, K., & Salama, G. (2022, July). Improving Arabic hate speech identification using online machine learning and deep learning models. In Proceedings of Seventh International Congress on Information and Communication Technology: ICICT 2022, London, Volume 2 (pp. 533-541). Singapore: Springer Nature Singapore.

[60] Ghosh, T., Al Banna, M. H., Al Nahian, M. J., Uddin, M. N., Kaiser, M. S., & Mahmud, M. (2023). An attention-based hybrid architecture with explainability for depressive social media text detection in Bangla. Expert Systems with Applications, 213, 119007.

[61] Kumar, R., & Bhat, A. (2022). A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media. International Journal of Information Security, 21(6), 1409-1431.

[62] Lin, H., Siarry, P., Gururaj, H. L., Rodrigues, J., & Jain, D. K. (2022). Special issue on deep learning methods for cyberbullying detection in multimodal social data. Multimedia Systems, 28(6), 1873-1875.

[63] Karwa, R. R., & Gupta, S. R. (2022). Automated hybrid Deep Neural Network model for fake news identification and classification in social networks. Journal of Integrated Science and Technology, 10(2), 110-119.

[64] Kumar, A., & Sachdeva, N. (2022). Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. Multimedia systems, 28(6), 2027-2041.

[65] Abdelhakim, M., Liu, B., & Sun, C. (2023). Ar-PuFi: A Short-Text Dataset to Identify the Offensive Messages Towards Public Figures in the Arabian Community. Expert Systems with Applications, 120888.

[66] de Pablo, Á., Araque, O., & Iglesias, C. A. (2022). Transfer Learning with Social Media Content in the Ride-Hailing Domain by Using a Hybrid Machine Learning Architecture. Electronics, 11(2), 189.

[67] Shanmugavadivel, K., Sathishkumar, V. E., Raja, S., Lingaiah, T. B., Neelakandan, S., & Subramanian, M. (2022). Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data. Scientific Reports, 12(1), 21557.

[68] Althobaiti, M. J. (2022). Bert-based approach to arabic hate speech and offensive language detection in twitter: Exploiting emojis and sentiment analysis. International Journal of Advanced Computer Science and Applications, 13(5).

[69] Nascimento, F. R., Cavalcanti, G. D., & Da Costa-Abreu, M. (2023). Exploring automatic hate speech detection on social media: a focus on content-based analysis. SAGE Open, 13(2), 21582440231181311.

[70] Del Valle-Cano, G., Quijano-Sánchez, L., Liberatore, F., & Gómez, J. (2023). SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles. Expert Systems with Applications, 216, 119446.

Automated Detection of Driver and Passenger Without Seat Belt using YOLOv8

Sutikno, Aris Sugiharto, Retno Kusumaningrum
Department of Informatics, Diponegoro University, Semarang, Indonesia

***Abstract*—The issue of traffic accident fatalities is a serious concern on a global scale, and one of the contributing factors is the failure of drivers to adhere to seat belt usage. A notable challenge arises from the limited availability of law enforcement personnel monitoring this particular issue. In this context, there is a compelling need to implement an automated detection system. The development of this system using YOLOv5 has been done. However, there are weaknesses related to the length of training and detection time. Therefore, this paper proposed a new system using the YOLOv8 method to detect drivers and passengers who violate seat belt regulations. The proposed system is divided into three subsystems: windshield detection, passenger classification, and seat belt classification. YOLOv8 is the latest version of the YOLO (You Only Look Once) method and has been proven to provide better performance than previous versions. Furthermore, this paper also compared five YOLOv8 models, namely YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. The proposed model is trained and tested using image data collected from several roads in Indonesia. The experiment results show that the YOLOv8s model produced the best mean Average Precision (mAP) of 0.960 for windshield detection. YOLOv8s-cls and YOLOv8l-cls models achieved the same accuracy of 0.8923 for passenger classification. The YOLOv8l-cls model produced the best accuracy of 0.8846 for seat belt classification. In addition, the proposed method can increase mAP and training time for windshield detection compared to YOLOv5.**

Keywords*—*Windshield detection; passenger classification; seat belt classification; YOLOv8

I. Introduction

Traffic accident fatalities are pressing global issues, ranking among the top 10 causes of mortality in low-income countries [1]. A key factor related to this issue is drivers' and passengers' non-adherence to seat belt usage. Even though the compulsory use of seat belts has been mandated, instances of non-compliance are alarmingly frequent. Using seat belts can significantly reduce the risk of severe injury or death in a traffic accident. Therefore, it is essential to raise public awareness among vehicle operators of the importance of wearing seat belts. Efforts are also needed to improve surveillance and enforcement of the regulations on the road.

In this context, monitoring of drivers and passengers in four-wheeled or larger vehicles is accomplished through direct observation. Several weaknesses have been reported in this approach since many violations remain undetected. This issue can be addressed by using Closed-Circuit Television (CCTV) cameras for surveillance on the road combined with automated detection based on computer vision. This automatic detection requires a method that produces high accuracy and speed. This application can help the police detect drivers and passengers not wearing seat belts. So their work becomes more efficient.

Research into computer vision-based detection of drivers' compliance with seat belt regulations commenced approximately a decade ago. Generally, this research can be divided into two types based on handcrafted and non-handcrafted features. The use of handcrafted features was shown by Qin et al., who combined Haar-like and Histogram of Oriented Gradient (HOG) descriptors to report efficiency and robustness [2]. Additionally, Elihos et al. used Fisher Vector [3]. Most research used non-handcrafted features, such as BN-AlexNet [4], Convolutional Neural Network (CNN) [5], A Nimble Architecture for Driver and Seat Belt Detection through Convolutional Neural Networks (NADS-Net) [6], YOLOv3 [7], and YOLOv5 [8], [9]. YOLOv5 achieved the best mean Average Precision (mAP) compared to other methods for detecting vehicle windshields, passengers, and drivers without seat belts [9]. However, there are weaknesses related to training and detection time. In addition, this method incorrectly detects low-quality data [10]. For this reason, a method is needed to increase detection accuracy and reduce detection time so that it can be applied in real-time.

YOLOv3 has also been used for other detections, for example, to detect pedestrians [11], vehicles [12], and road objects [13]. Another investigation compared YOLOv5, YOLOv6, and YOLOv7 for detecting small objects [14], while YOLOv7 was used for real-time weed detection [15]. Yung et al. compared YOLOv5, YOLOv6, and YOLOv7 to detect safety helmets workers wear [16]. The test results showed that YOLOv7 achieved better mAP than YOLOv5 and YOLOv6. Furthermore, YOLOv7 exhibited superior accuracy and speed compared to YOLOR, PP-YOLOE, YOLOX, Scaled-YOLOv4, and YOLOv5 algorithms.

The latest version of YOLO (You Only Look Once), YOLOv8, was more advanced than its predecessors. YOLOv8 produced high average accuracy. YOLOv8 scores much better than YOLOv5 [17]. Therefore, this paper developed a new system using the YOLOv8 to detect drivers and passengers without seat belts. The proposed system was divided into three subsystems: windshield detection, passenger classification, and seat belt classification. The main contribution of this paper is that the proposed method can produce relatively high accuracy so that it can potentially be applied to actual conditions.

The remainder of this paper is organized as follows: Section II is about related work. Section III describes the dataset, the proposed method, and the parameters used to

evaluate the performance of the proposed method. Section IV explains the test results and analysis of the windshield detection, passenger classification, and seat belt classification subsystems. Additionally, this section compares the proposed method with previous research. Finally, Section V explains the conclusions and research that will be carried out.

II. Related Work

Research on driver detection without seat belts based on computer vision started about a decade ago. Generally, this research can be divided into two types based on handcrafted and non-handcrafted features. Guo et al. used license plate and edge detection to determine the area of the driver's area and the position of the seat belt [18]. Li et al. also used Canny edge detection, with the driver's area determined by cutting the left half of the windshield [19]. Meanwhile, the windshield area's position was detected using the cascade Adaboost classifier.

Zhou et al. proposed Canny edge detection and salient gradient map for feature extraction and learning-based algorithms for binary classification [20]. Yang et al. determined the driver's area using face and seat belt detection through connected area methods [21]. Wu et al. introduced a methodology that includes ascertaining the driver's area through semantic segmentation, streamlined by a pruning process, and conducting classification using connected techniques [22]. Yongquan et al. focused on reducing computation time for driver detection without seat belts by designing a Graphics Processing Unit (GPU) acceleration method [23]. The driver's area was obtained using Squeeze-YOLO, while seat belt usage was determined using semantic segmentation algorithms and full convolution network pruning. Wang et al. used semantic segmentation but lightweight feature extraction and the Squeeze-YOLO algorithm to determine the driver's area [24].

In pursuit of enhanced accuracy, specific research endeavors have embraced a fusion of multiple descriptors. For instance, Qin et al. integrated Haar-like features and HOG [25]. Madake et al. combined feature extraction techniques such as Canny, FAST (Features from Accelerated Segment Test), and BRIEF (Binary Robust Independent Elementary Features) [26]. The test results showed that the proposed feature extraction combinations improved accuracy.

In recent years, most research has used CNN. Furthermore, Sajja et al. used this method and compared the concept to Support Vector Machine (SVM) [27]. Test results showed that the CNN method achieved better accuracy than SVM. Kapdi et al. adopted the MobileNetV2 model [28], which had the advantage of robustness in different weather conditions. Meanwhile, Chen et al. combined CNN and SVM for feature extraction and classification. This method was based on multi-scale feature extraction and applied to images with complex road backgrounds. In this context, the proposed method achieved better average detection than the Adaboost algorithm and CNN [5]. Kannadaguli proposed a detection method using Fully Connected One Shot (FCOS) and added prediction elimination with Non-Maximum Suppression (NMS) [29]. Additionally, Elihos et al. compared the Single Shot MultiBox Object Detector (SSD), VGG16 model, shallower CNN model, and Fisher vector model [3]. The test results showed that the SSD model achieved the highest accuracy.

Some versions of YOLO have also been used in this field. Luo et al. used YOLOv3 to determine the driver's area and CNN for the classification process [30]. The testing showed that the proposed method achieved high accuracy and robustness in complex environments. Wang and Ma also used YOLOv3 and a lightweight network structure [7]. In this context, increasing the number of lightweight templates improved accuracy but reduced speed.

Furthermore, Khalid and Hazela used YOLOv4 to determine the driver's area and the AlexNet model for classification [31]. Feng et al. proposed YOLOv5 to locate the driver's area and used the AlexNet deep convolutional network for classification [8]. This method saved memory and computational time compared to SVM. Hosseini and Fathi used YOLOv5 to detect car windshields and the ResNet32 model for classification [9].

Maduri et al. proposed a method for real-time driver detection without seat belts [32]. This method used deep learning and was embedded in a Raspberry Pi. Upadhyay et al. also focused on real-time conditions and implemented the concept in real-world scenarios to install a camera in a car cabin using YOLOv5 [33]. Zang et al. also proposed the SlimSSDMV2 and Line Segment Detector (LSD) models, which were applied to mobile devices [34].

Some researchers aimed to improve accuracy and speed with the development of the CNN model. Zhou et al. developed the Alexnet model by adding Batch Normalization (BN), known as BN-Alexnet [4]. The proposed method increased the average accurate detection and reduced training time compared to Alexnet, VGGNet-16, and GoogLeNet. Chun et al. developed the CNN model known as NADS-Net using the feature pyramid network (FPN) backbone and multiple detection heads method [6]. Additionally, Yang et al. proposed a method focused on the Central Processing Unit (CPU) and real-time implementation. The proposed method combined traditional operators (texture extraction), SSD MobileNet V2, and particle filter tracking algorithms [35].

III. Material and Method

A. Dataset

This research used three datasets, namely Dataset1, Dataset2, and Dataset3. The datasets were captured from video frame recordings on several roads, using CCTV and cameras in Bandung and Semarang City, Indonesia. Fig. 1 shows samples of images from Dataset1. This dataset contains video frame images with annotations of windshields used in the windshield detection subsystem.

Fig. 2 shows samples of images from Dataset2 used in the passenger classification. This dataset consists of a set of seating area images comprising two classes, namely 488 passenger images and 484 no-passenger images. These images are obtained from the right half of the windshield image.

Fig. 3 shows samples of images from Dataset3 used in the seat belt classification. This dataset contains images of drivers and passengers with and without seat belts. The number of

images of drivers or passengers wearing seat belts is 957, and the number of images of drivers or passengers not wearing seat belts is 727. Whereas Table I shows the number of training and testing data in each dataset. The distribution of training and testing data is carried out randomly, with a ratio of 80% for training data and 20% for testing data.

Fig. 1. Samples of dataset1.

Fig. 2. Samples of dataset2.

Fig. 3. Samples of dataset3.

TABLE I. NUMBER OF TRAINING AND TESTING DATA

Dataset	Image Number		
	Training	***Testing***	***Total***
Dataset1	2631	658	3289
Dataset2	777	195	972
Dataset3	1346	338	1684

B. Proposed System

This research built a system for detecting drivers or passengers without seat belts. The proposed system generally consists of three subsystems: windshield detection, passenger classification, and seat belt classification. Fig. 4 displays this proposed system.

1) Windshield detection: A car's driver and front passengers are visible through the car's windshield. Therefore, the initial step includes detecting each car's windshield position using the YOLOv8 method. YOLO predicts bounding boxes (Bbox) and class probabilities (Cls) through a single network. Furthermore, it excels in high accuracy even when using small model sizes and can be trained on a single GPU [17]. The method is cost-effective for machine learning practitioners with limited hardware resources or cloud computing. Ultralytics developed YOLOv8 as the latest model for object detection, image classification, and image segmentation. This method is an anchor-free model that directly predicts the object's center and augments images during online training. The model analyzes variations of the provided images in each epoch through mosaic augmentation. This augmentation has empirically been found to decrease performance when applied throughout the entire training routine. Fig. 5 shows the architecture of the YOLOv8 model divided into two main parts: the backbone and the head [36]. The backbone architecture consists of 53 convolutional layers and uses partial cross-stage connections to enhance information flow between different layers, promoting better feature representation and extraction based on CSPDarknet53. Meanwhile, the head of YOLOv8 consists of several convolutional and fully connected layers. These layers are used for object detection, including bounding box prediction, objectivity scores prediction, and class probabilities for image objects. The model uses a feature pyramid network to detect large and small objects accurately.

2) Passenger classification: The windshield detection subsystem results in the windshield area. This area has two seats: the driver (left) and the passenger (right). The driver's seat always has a driver, but the passenger's seat may or may not have a passenger. Therefore, the passenger classification subsystem classifies the presence of a passenger in the right seat. This subsystem comprises windshield cropping, division of the driver and passenger areas, and the classification process. The windshield cropping process is used to capture the windshield area of each car. Division of driver and passenger areas separates the area into the left and right sides for the driver and passenger areas. The classification process is conducted in two classes (seat with passenger and seat without passenger) using the YOLOv8 method.

3) Seat belt classification: The final subsystem is the seat belt classification, which determines whether the driver and passenger are wearing a seat belt. We used YOLOv8 method.

C. Performance Evaluation

Performance evaluation for windshield detection used the parameters of precision (P), recall (R), and mAP. Precision and recall are calculated using as in Eq. (1) and Eq. (2), respectively [37]. TP_{dec} represents correct detections of the ground truth bounding box, FP_{dec} is the false detections of objects, and FN_{dec} represents the undetected ground truth.

$$P = \frac{TP_{dec}}{TP_{dec}+FP_{dec}} \quad (1)$$

$$R = \frac{TP_{dec}}{TP_{dec}+FN_{dec}} \quad (2)$$

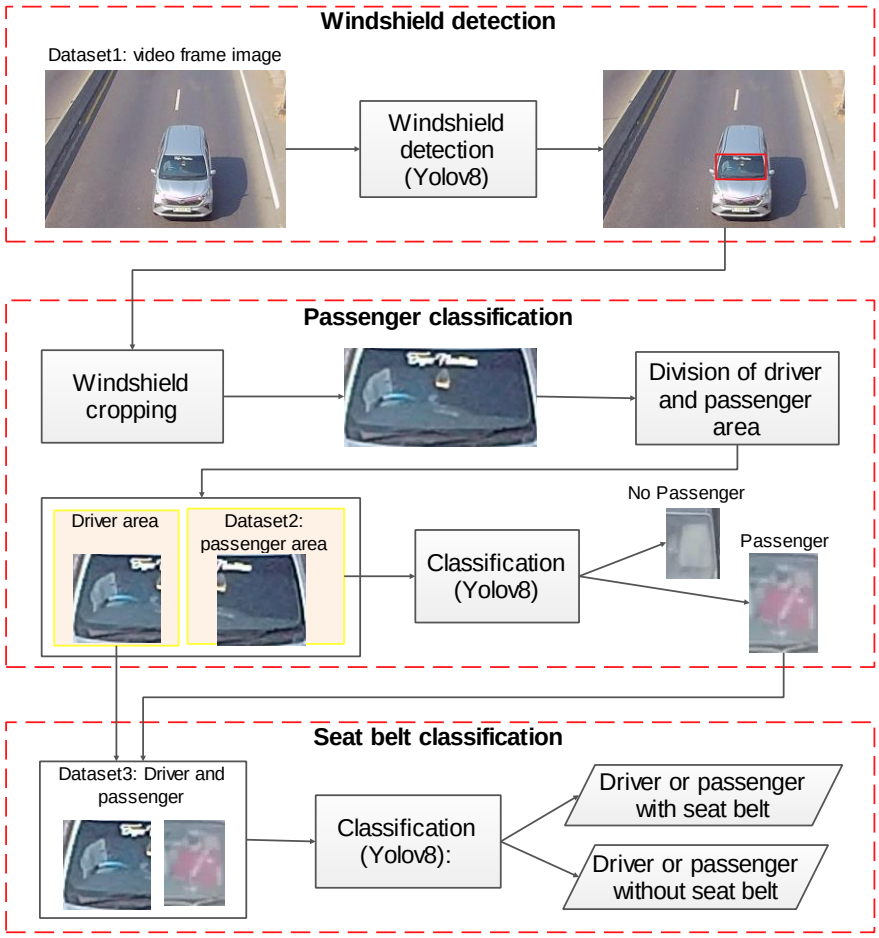


Fig. 4. Proposed system.

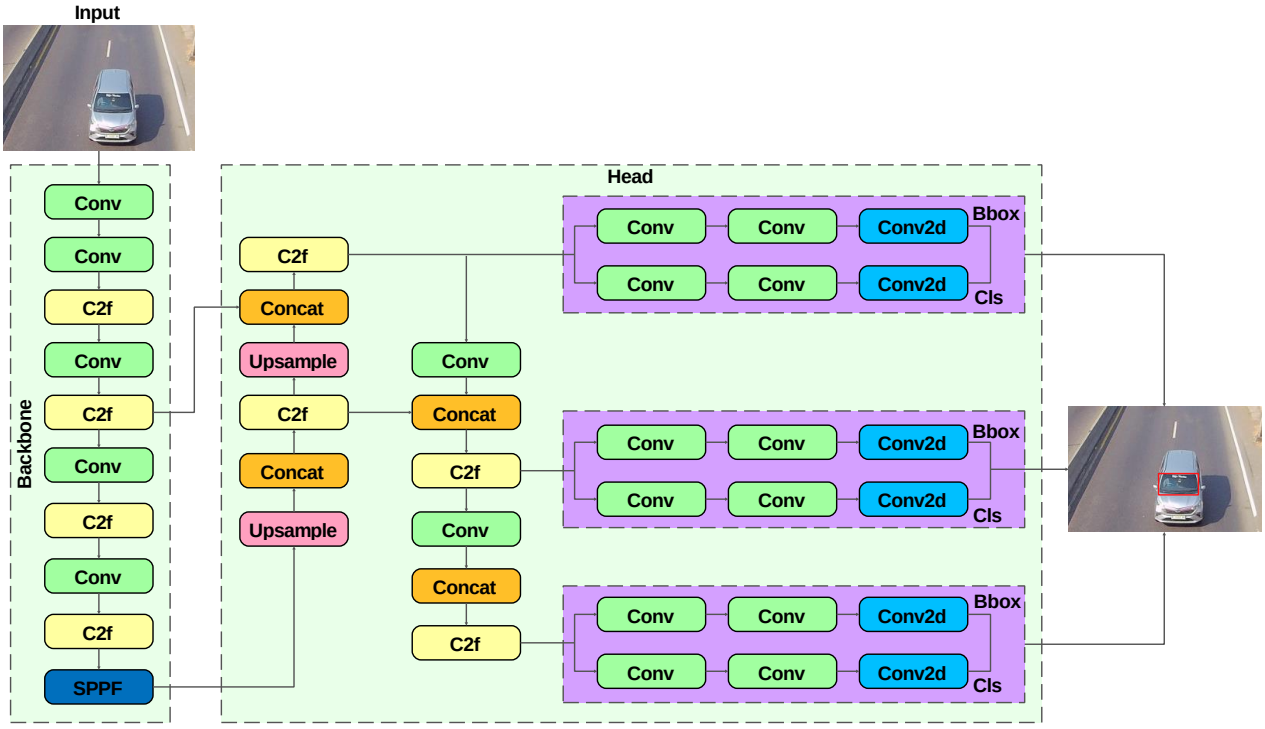


Fig. 5. The architecture of the YOLOv8 model.

Meanwhile, mAP is calculated as in Eq. (3) [37].

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (3)$$

AP_i is the average precision for each class, and N is the number of classes. AP for each class is calculated as in Eq. (4).

$$AP = \sum_n (R_{n+1} - R_n) P_i(R_{n+1}) \quad (4)$$

$P_i(R_{n+1})$ is calculated as in (5).

$$P_i(R_{n+1}) = \max_{\tilde{R}:\tilde{R}\geq R_{n+1}} P(\tilde{R}) \quad (5)$$

$P(\tilde{R})$ is the precision measured at the time of recall $\tilde{R}$.

The performance evaluation for passenger and seat belt classification used an accuracy parameter calculated as in Eq. (6). Table II displays a confusion matrix containing True Positive (TP_{cls}), False Positive (FP_{cls}), False Negative (FN_{cls}), and True Negative (TN_{cls}).

$$Accuracy = \frac{TP_{cls}+TN_{cls}}{TP_{cls}+TN_{cls}+FP_{cls}+FN_{cls}} \quad (6)$$

TABLE II. CONFUSION MATRIX

		Actual Value	
		Positive	**Negative**
Predicted	***Positive***	TP_{cls}	FP_{cls}
	Negative	FN_{cls}	TN_{cls}

IV. EXPERIMENTS AND RESULTS

The experiments were conducted on each subsystem using different datasets. The windshield detection, passenger classification, and seat belt classification subsystems used Dataset1, Dataset2, and Dataset3, respectively. The experiments were performed using YOLOv8, and the model was trained and tested on a Tesla T4 GPU.

A. Windshield Detection

The first experiment was conducted for windshield detection using five YOLOv8 models, including YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. All experiments used 640×640 pixels for input size, 16 for batch size, 100 for epochs, and Adam for optimizer. Table III shows the experiment result of windshield detection. The best precision and recall were achieved using the YOLOv8s and YOLOv8n models, respectively, while the fastest training time was obtained with YOLOv8s. Additionally, the best mAP was achieved using the YOLOv8m model at 0.960. Fig. 6, Fig. 7, and Fig. 8 display the precision-recall, precision-confidence, and recall-confidence curves of this experiment, respectively. Meanwhile, Fig. 9 displays examples of windshield detection results.

TABLE III. EXPERIMENT RESULT OF WINDSHIELD DETECTION

Model	P	R	mAP	Time (hours)
Yolov8n	0.878	**0.933**	0.959	1.659
Yolov8s	**0.882**	0.921	0.950	**1.628**
Yolov8m	0.879	0.913	**0.960**	2.391
Yolov8l	0.860	0.895	0.945	3.606
Yolov8x	0.847	0.893	0.943	5.718

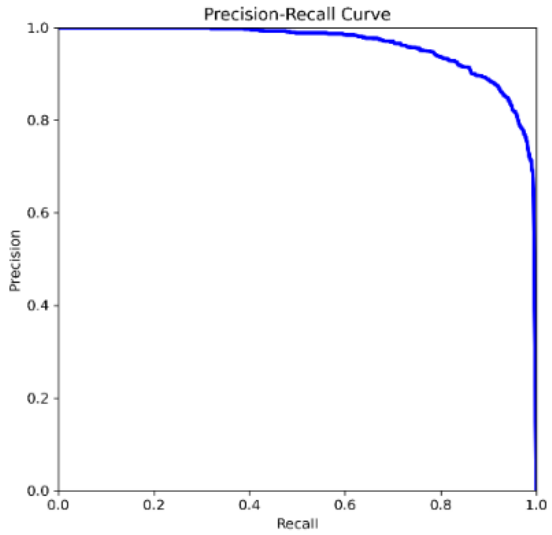


Fig. 6. Precision-recall curve of windshield detection.

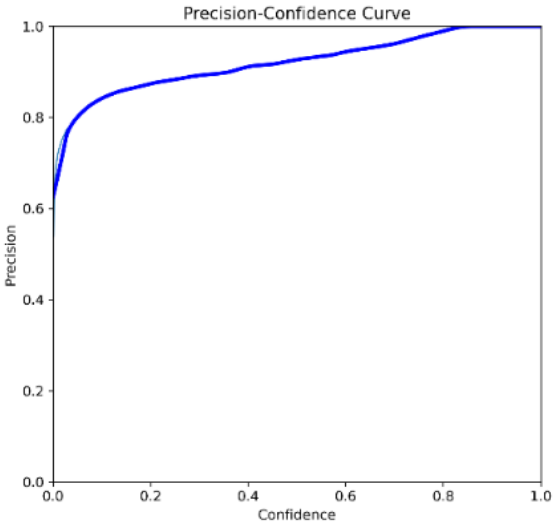


Fig. 7. Precision-confidence curve of windshield detection.

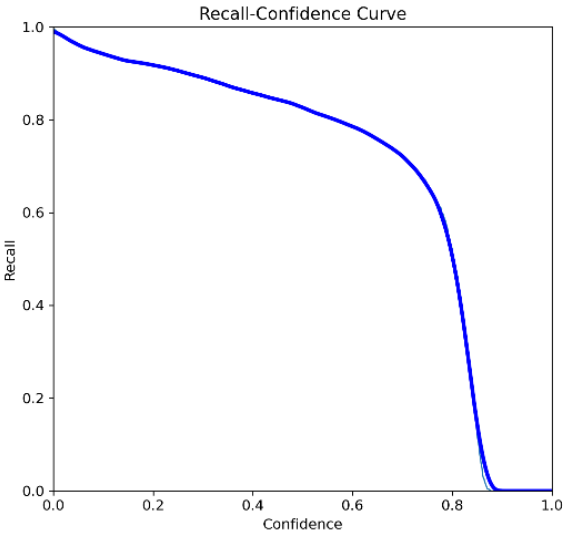


Fig. 8. Recall-confidence curve of windshield detection.

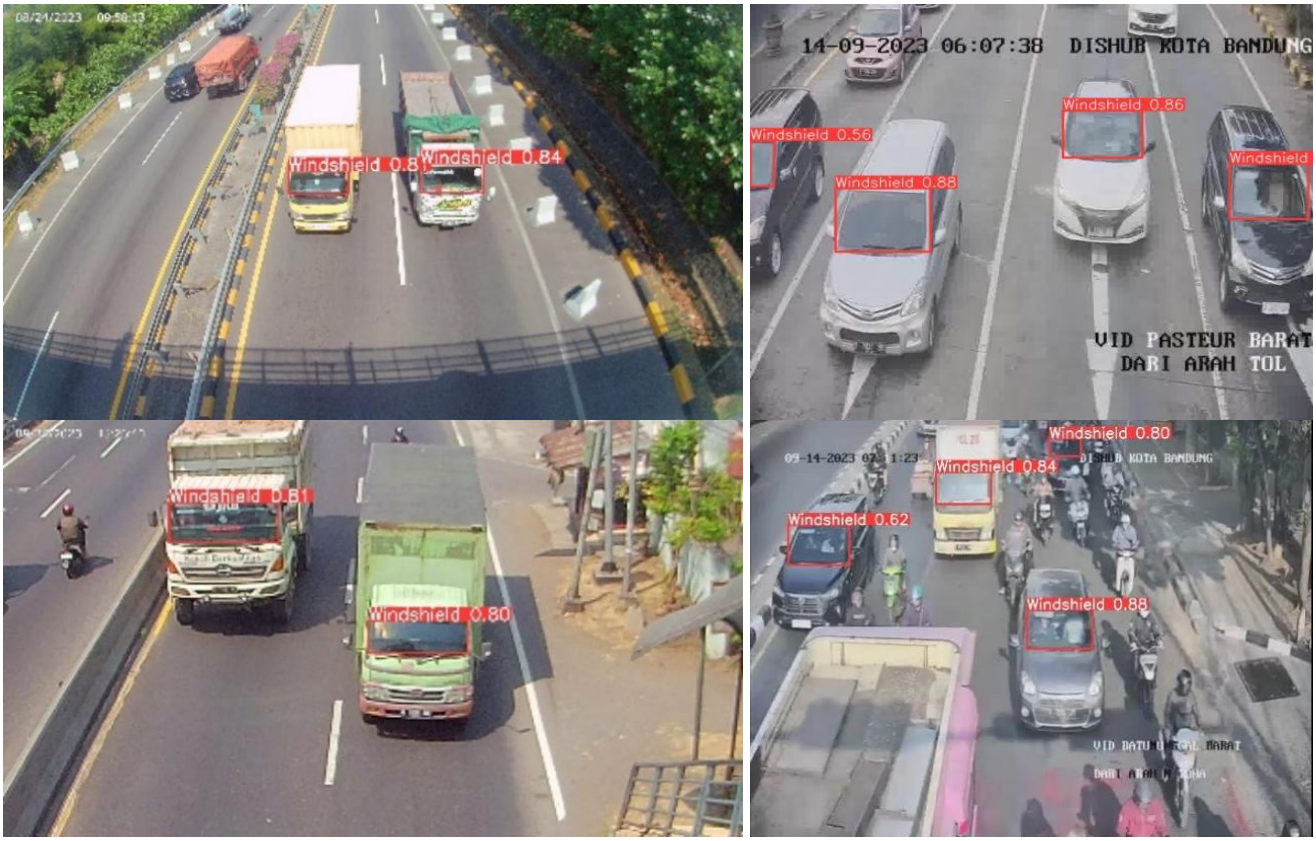


Fig. 9. Example of windshield detection from video frame images.

The proposed method was compared to the previous research, as shown in Table IV. Here, the study in [9] used the YOLOv5s model with 48 for batch size, 0.0001 for learning rate, and Adam for the optimizer. The table shows that the proposed method outperformed precision, recall, and mAP. The proposed method increases mAP by 2% and decreases training time by 0.54 times compared to the YOLOv5s model. Therefore, the proposed method is more suitable for real-time windshield detection.

TABLE IV. COMPARISON BETWEEN THE PROPOSED METHOD AND PREVIOUS RESEARCH METHOD (YOLOV5)

Model	P	R	mAP	Time (hours)
[9]	0.824	9.907	0.940	4.427
Proposed	**0.879**	**0.913**	**0.960**	**2.391**

B. *Passenger classification*

A passenger classification experiment was conducted using five models and three batch sizes. Those models are YOLOv8n-cls, YOLOv8s-cls, YOLOv8m-cls, YOLOv8l-cls, and YOLOv8x-cls and those batch sizes are 4, 8, and 16. All training used 224×224 pixels for image size input, 100 for epochs, Adam for optimizer, and 20 for early stopping. Table V shows the results of this experiment. The highest accuracy reaches 0.8923 using YOLOv8s-cls and YOLOv8l-cls. Table VI is a confusion matrix of the best result of passenger classification. Furthermore, the fastest epochs and training time are achieved using the YOLOv8n-cls model with 16 for the batch size.

TABLE V. EXPERIMENT RESULT OF PASSENGER CLASSIFICATION

Model	Batch size	Accuracy	Epoch	Time (hours)
Yolov8n-cls	4	0.8667	38	0.095
	8	0.8718	43	0.076
	16	0.8564	**33**	**0.051**
Yolov8s-cls	4	0.8871	100	0.235
	8	0.8820	88	0.148
	16	**0.8923**	88	0.125
Yolov8m-cls	4	0.8769	100	0.302
	8	0.8719	65	0.143
	16	0.8718	100	0.165
Yolov8l-cls	4	**0.8923**	76	0.295
	8	0.8821	83	0.229
	16	0.8667	91	0.216
Yolov8x-cls	4	0.8872	100	0.533
	8	0.8821	100	0.447
	16	0.8769	87	0.369

TABLE VI. CONFUSION MATRIX OF PASSENGER CLASSIFICATION

		Actual value	
		Passenger	***No Passenger***
Predicted	***Passenger***	88	11
	No Passenger	10	86

C. Seat Belt Classification

The seat belt classification experiment used five models and three batch sizes, the same as in the passenger classification experiment. Table VII shows the results of this experiment. In this context, the highest accuracy reaches 0.8846 using the YOLOv8l-cls model and 16 for batch size. Table VIII is a confusion matrix of the best experiment result. The fastest training time reaches 0.126 hours using the YOLOv8n-cls model and 16 for batch size.

TABLE VII. EXPERIMENT RESULT OF SEAT BELT CLASSIFICATION

Model	Batch size	Accuracy	Epoch	Time (hours)
Yolov8n-cls	4	0.8669	100	0.432
	8	0.8787	61	0.176
	16	0.8609	**52**	**0.126**
Yolov8s-cls	4	0.8669	100	0.410
	8	0.8698	100	0.281
	16	0.8550	86	0.201
Yolov8m-cls	4	0.8669	81	0.400
	8	0.8640	100	0.365
	16	0.8698	75	0.205
Yolov8l-cls	4	0.8728	82	0.483
	8	0.8787	81	0.351
	16	**0.8846**	100	0.339
Yolov8x-cls	4	0.8639	100	0.725
	8	0.8727	100	0.527
	16	0.8639	65	0.276

TABLE VIII. CONFUSION MATRIX OF SEAT BELT CLASSIFICATION

		Actual Value	
		No seat belt	***Seat belt***
Predicted	***No seat belt***	127	20
	Seat belt	19	172

V. CONCLUSION

This research developed a system for detecting drivers and passengers who violate seat belt regulations. The proposed system consists of three subsystems: windshield detection, passenger classification, and seat belt classification. The proposed methods in each subsystem used YOLOv8. The experiment was conducted by comparing five models. Furthermore, three batch sizes were compared in passenger and seat belt classification. The results show that the proposed method achieved an mAP of 0.960 for windshield detection. An accuracy of 0.8923 and 0.8846 was reported for passenger and seat belt classifications. Moreover, the proposed method excelled in accuracy and training time compared to YOLOv5 for windshield detection. Therefore, the proposed method is more suitable for detecting drivers or passengers who violate seat belts in real-time. A license plate detection subsystem could be adopted for future research to identify vehicles.

ACKNOWLEDGMENT

The authors are grateful to the Ministry of Research and Higher Education of Indonesia through Diponegoro University, which has funded this research with contract numbers 017/E5/PG.02.00.PL/2023.

REFERENCES

[1] W. H. Organization, “The top 10 causes of death,” World Health Organization, 2020. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

[2] X.-H. Qin, C. Cheng, G. Li, and X. Zhou, “Efficient seat belt detection in a vehicle surveillance application,” in 9th IEEE Conference on Industrial Electronics and Applications, ICIEA 2014, 2014, pp. 1247–1250, doi: 10.1109/ICIEA.2014.6931358.

[3] A. Elihos, B. Alkan, B. Balci, and Y. Artan, “Comparison of image classification and object detection for passenger seat belt violation detection using NIR RGB surveillance camera images,” 2019, doi: 10.1109/AVSS.2018.8639447.

[4] B. Zhou, D. Chen, and X. Wang, “Seat belt detection using convolutional neural network BN-Alexnet,” 13th International Conference on Intelligent Computing, ICIC 2017, vol. 10361 LNCS. Springer Verlag, College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China, pp. 384–395, 2017, doi: 10.1007/978-3-319-63309-1_36.

[5] Y. Chen, G. Tao, H. Ren, X. Lin, and L. Zhang, “Accurate seat belt detection in road surveillance images based on CNN and SVM,” Neurocomputing, vol. 274, pp. 80–87, 2018, doi: 10.1016/j.neucom.2016.06.098.

[6] S. Chun et al., “NADS-Net: A nimble architecture for driver and seat belt detection via convolutional neural networks,” in 17th IEEE/CVF International Conference on Computer Vision Workshop, ICCVW 2019, 2019, pp. 2413–2421, doi: 10.1109/ICCVW.2019.00295.

[7] Z. Wang and Y. Ma, “Detection and recognition of stationary vehicles and seat belts in intelligent internet of things traffic management system,” Neural Computing and Applications, vol. 34, no. 5, pp. 3513–3522, 2022, doi: 10.1007/s00521-021-05870-6.

[8] W. Feng, W. Yu, and R. Nan, “Deep learning based vehicle seat belt detection algorithm for driver and passenger seat occupants,” in 7th International Conference on Intelligent Informatics and Biomedical Sciences, ICIIBMS 2022, 2022, pp. 306–310, doi: 10.1109/ICIIBMS55689.2022.9971531.

[9] S. Hosseini and A. Fathi, “Automatic detection of vehicle occupancy and driver’s seat belt status using deep learning,” Signal, Image and Video Processing, vol. 17, no. 2, pp. 491–499, 2023, doi: 10.1007/s11760-022-02244-w.

[10] A. Upadhyay, B. Sutrave, and A. Singh, “Real time seatbelt detection using YOLO deep learning model,” in 2023 IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS), 2023, pp. 1–6, doi: 10.1109/SCEECS57921.2023.10063114.

[11] K. C. Saranya and A. Thangavelu, “Vulnerable road user detection using YOLOv3,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 12, pp. 576–582, 2019, doi: 10.14569/ijacsa.2019.0101275.

[12] O. Bourja, H. Derrouz, H. A. Abdelali, A. Maach, R. O. H. Thami, and F. Bourzeix, “Real time vehicle detection, tracking, and inter-vehicle distance estimation based on stereovision and deep learning using YOLOv3,” International Journal of Advanced Computer Science and Applications, vol. 12, no. 8, pp. 915–923, 2021, doi: 10.14569/IJACSA.2021.01208101.

[13] G. Al-refai and M. Al-refai, “Road object detection using YOLOv3 and kitti dataset,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 8, pp. 48–53, 2020, doi: 10.14569/IJACSA.2020.0110807.

[14] L. Yang, “Investigation of you only look once networks for vision-based small object detection,” International Journal of Advanced Computer Science and Applications, vol. 14, no. 4, pp. 69–82, 2023, doi: 10.14569/IJACSA.2023.0140410.

[15] C. L. Narayana and K. V. Ramana, "An efficient real-time weed detection technique using YOLOv7," International Journal of Advanced Computer Science and Applications, vol. 14, no. 2, pp. 550–556, 2023, doi: 10.14569/IJACSA.2023.0140265.

[16] N. D. T. Yung, W. K. Wong, F. H. Juwono, and Z. A. Sim, "Safety helmet detection using deep learning: implementation and comparative study using YOLOv5, YOLOv6, and YOLOv7," in 2022 International Conference on Green Energy, Computing and Sustainable Technology, GECOST 2022, 2022, pp. 164–170, doi: 10.1109/GECOST55694.2022.10010490.

[17] J. Solawetz and Francesco, "What is YOLOv8? the ultimate guide," 2023. https://blog.roboflow.com/whats-new-in-yolov8/#the-yolov8-annotation-format.

[18] H. Guo, H. Lin, S. Zhang, and S. Li, "Image-based seat belt detection," in Proceedings of 2011 IEEE International Conference on Vehicular Electronics and Safety, ICVES 2011, 2011, pp. 161–164, doi: 10.1109/ICVES.2011.5983807.

[19] W. Li, J. Lu, Y. Li, Y. Zhang, J. Wang, and H. Li, "Seatbelt detection based on cascade adaboost classifier," in Proceedings of the 2013 6th International Congress on Image and Signal Processing (CISP 2013), 2013, vol. 2, pp. 783–787, doi: 10.1109/CISP.2013.6745271.

[20] B. Zhou, L. Chen, J. Tian, and Z. Peng, "Learning-based seat belt detection in image using salient gradient," in 12th IEEE Conference on Industrial Electronics and Applications, ICIEA 2017, 2018, vol. 2018-Febru, pp. 547–550, doi: 10.1109/ICIEA.2017.8282904.

[21] Z. Yang, H. Xiong, Z. Cai, and Y. Peng, "A new method of vision-based seat belt detection," International Journal of Embedded Systems, vol. 11, no. 6, pp. 755–763, 2019, doi: 10.1504/IJES.2019.103992.

[22] T. Wu, Z. Zhang, Y. Liu, W. Guo, and Z. Wang, "Driver seat belt detection based on YOLO detection and semantic segmentation," Jisuanji Fuzhu Sheji Yu Tuxingxue Xuebao/Journal of Computer-Aided Design and Computer Graphics, vol. 31, no. 1, pp. 126–131, 2019, doi: 10.3724/SP.J.1089.2019.17244.

[23] J. Yongquan, W. Tianshu, L. Jin, Z. Zhijia, and G. Chao, "GPU acceleration design method for driver's seatbelt detection," in 2019 14th IEEE International Conference on Electronic Measurement and Instruments, ICEMI 2019, 2019, pp. 949–953, doi: 10.1109/ICEMI46757.2019.9101821.

[24] D. Wang, "Intelligent detection of vehicle driving safety based on deep learning," Wireless Communications and Mobile Computing, vol. 2022, pp. 1–11, 2022, doi: 10.1155/2022/1095524.

[25] X. H. Qin, C. Cheng, G. Li, and X. Zhou, "Efficient seat belt detection in a vehicle surveillance application," in Proceedings of the 2014 9th IEEE Conference on Industrial Electronics and Applications, ICIEA 2014, 2014, pp. 1247–1250, doi: 10.1109/ICIEA.2014.6931358.

[26] J. Madake, S. Yadav, S. Singh, S. Bhatlawande, and S. Shilaskar, "Vision-based driver's seat belt detection," in 2023 International Conference for Advancement in Technology, 2023, pp. 1–5, doi: 10.1109/ICONAT57137.2023.10080147.

[27] D. S. B. Naik, G. S. Lakshmi, V. R. Sajja, D. Venkatesulu, and J. N. Rao, "Driver's seat belt detection using CNN," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 5, pp. 776–785, 2021, doi: 10.17762/turcomat.v12i5.1483.

[28] R. A. Kapdi, P. Khanpara, R. Modi, and M. Gupta, "Image-based seat belt fastness detection using deep learning," Scalable Computing, vol. 23, no. 4, pp. 441–455, 2022, doi: 10.12694/scpe.v23i4.2027.

[29] P. Kannadaguli, "FCOS based seatbelt detection system using thermal imaging for monitoring traffic rule violations," in 2020 4th International Conference on Electronics, Materials Engineering and Nano-Technology, IEMENTech 2020, 2020, pp. 1–6, doi: 10.1109/IEMENTech51367.2020.9270058.

[30] J. Luo, J. Lu, and G. Yue, "Seatbelt detection in road surveillance images based on improved dense residual network with two-level attention mechanism," Journal of Electronic Imaging, vol. 30, no. 3, 2021.

[31] S. Bin Khalid and B. Hazela, "Employing real-time object detection for traffic monitoring," in Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021, 2021, pp. 1–11, doi: 10.2139/ssrn.3834176.

[32] P. K. Maduri, G. Singh, S. Sharma, R. K. Mishra, and N. K. Mishra, "Seat belt and helmet detection using deep learning," in 3rd International Conference on Advances in Computing, Communication Control and Networking, ICAC3N 2021, 2021, pp. 476–480, doi: 10.1109/ICAC3N53548.2021.9725574.

[33] A. Upadhyay, B. Sutrave, and A. Singh, "Real time seatbelt detection using YOLO deep learning model," in 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 2023, pp. 1–6, doi: 10.1109/SCEECS57921.2023.10063114.

[34] Y. Zang, B. Yu, and S. Zhao, "Lightweight seatbelt detection algorithm for mobile device," Multimedia Tools and Applications, vol. 82, pp. 24505–24519, 2023, doi: 10.1007/s11042-023-14555-2.

[35] D. Yang, Y. Zang, and Q. Liu, "Study of detection method on real-time and high precision driver seatbelt," in Proceedings of the 32nd Chinese Control and Decision Conference (CCDC 2020), 2020, pp. 79–86, doi: 10.1109/CCDC49329.2020.9164214.

[36] H. T. Ngoc, K. H. Nguyen, H. K. Hua, H. V. N. Nguyen, and L. Da Quach, "Optimizing YOLO performance for traffic light detection and end-to-end steering control for autonomous vehicles in Gazebo-ROS2," International Journal of Advanced Computer Science and Applications, vol. 14, no. 7, pp. 475–484, 2023, doi: 10.14569/IJACSA.2023.0140752.

[37] R. Padilla, S. L. Netto, and E. A. B. Da Silva, "A survey on performance metrics for object-detection algorithms," in International Conference on Systems, Signals, and Image Processing, pp. 237–242, doi: 10.1109/IWSSIP48289.2020.9145130.

Enhancing Alzheimer's Disease Diagnosis: The Efficacy of the YOLO Algorithm Model

Tran Quang Vinh, Haewon Byeon*
Department of Digital Anti-Aging Healthcare (BK21), Inje University, Gimhae 50834, South Korea

***Abstract*—The diagnosis and early detection of Alzheimer's Disease (AD) and other forms of dementia have become increasingly crucial as our aging population grows. In recent years, deep learning, particularly the You Only Look Once (YOLO) architecture, has emerged as a promising tool in the field of neuroimaging and machine learning for AD diagnosis. This comprehensive review investigates the recent advances in the application of YOLO for AD diagnosis and classification. We scrutinized five research papers that have explored the potential of YOLO, delving into the methodologies, datasets, and results presented. Our review reveals the remarkable strides made in AD diagnosis using YOLO, while also highlighting challenges, such as data scarcity and research lacking. The paper provides insights into the growing role of YOLO in the early detection of AD and its potential to transform clinical practices in the field. This review aims to inspire further research and innovation to enhance AD diagnosis and, ultimately, patient care.**

***Keywords*—Machine learning; deep learning; YOLO; alzheimer's disease; dementia**

I. Introduction

There is a surging interest in the application of Artificial Intelligence (AI) within the realm of healthcare. Health care-related AI research has seen a rapid acceleration in publication growth since 2012, with a 45.1% increase in the past five years, driven by technological breakthroughs and is expected to continue doubling approximately every two years based on this growth trend [1]. AI has solidified its position as a transformative power in the healthcare sector, completely reshaping the approaches to diagnosis, treatment, and medical condition management. In recent years, AI has emerged as an indispensable asset in the healthcare industry, offering groundbreaking solutions to some of the most formidable challenges in medicine, particularly when addressing neurological diseases. Neurological diseases, encompassing a diverse spectrum of conditions, such as Alzheimer's disease (AD), stroke and Parkinson's disease, pose intricate challenges in terms of diagnosis and treatment [2,3]. AI has decisively altered the landscape in this context.

AI applications in the realm of neurological diseases are both diverse and promising. AI, particular machine learning (ML) and deep learning (DL) architecture have the capability to scrutinize extensive volumes of brain imaging data, encompassing magnetic resonance imaging (MRI), positron emission tomography (PET), and computed tomography (CT) scans, in order to unearth subtle anomalies that might elude human perception [4,5,6]. In contrast to conventional diagnostic and treatment methodologies, these AI-driven approaches address several limitations inherent in traditional methods, such as subjectivity, delayed diagnoses often resulting from inconspicuous early-stage symptoms, or findings imperceptible to human observers. This proficiency in early detection of neurological disorders offers the potential for swifter and more precise diagnoses.

In particular, the deep learning object detection algorithm known as You Only Look Once (YOLO) shows great promise in enhancing the accuracy, efficiency, and automation of diagnosing neurological diseases, with a special emphasis on Alzheimer's disease. The primary aim of this brief review is to investigate the present applications of YOLO in the classification of neurological diseases with a particular focus on Alzheimer's disease. Additionally, we will delve into the methods used and the challenges faced when applying AI to the diagnosis and treatment of neurological diseases.

II. Materials and Methods

A. Artificial Intelligence in AD Diagnosis

AD is a formidable and complex neurological condition that has captured the attention of scientists, healthcare professionals, and society at large. Named after Dr. Alois Alzheimer, who first described the disease in the early 20th century [7], Alzheimer's is a progressive and degenerative brain disorder that predominantly affects memory, cognitive function, and daily life activities. The impact of AD extends far beyond the affected individuals themselves, as it profoundly affects their families and caregivers, often placing an immense emotional and practical burden on them. It is the most common cause of dementia, a term that encompasses a range of cognitive impairments that interfere with an individual's ability to think, reason, remember, and communicate. AD is a devastating and relentless neurological disorder that presents a profound challenge to both the medical community and society as a whole [8]. It is estimated that over 50 million people worldwide are currently affected by AD [9]. As the global population ages, this number is projected to escalate significantly in the coming decades. This ailment has grown into one of the most prevalent and impactful health concerns of our time [10].

As is the case with numerous other neurological disorders [11], early diagnosis holds a crucial position in the care and strategic planning for Alzheimer's disease (AD). The classification of AD is based on different levels, which include Alzheimer's disease (AD), mild cognitive impairment (MCI), and cognitively normal (CN). Early identification in MCI level empowers individuals and their families to take proactive steps in addressing critical aspects of their future, encompassing healthcare preferences, support requirements, and financial and

*Corresponding Author.

legal considerations [12, 13]. Additionally, early detection allows for proactive safety measures to reduce the risk of wandering or disorientation-related incidents. Moreover, it opens up the possibility of participating in clinical trials for innovative treatments during the disease's early stages, contributing to advancements in research.

Despite recent advancements in clinical trials related to Alzheimer's disease, several challenges have emerged. These challenges include the difficulty of distinguishing AD from normal age-related cognitive changes, limited access to specialized diagnostic tools in certain geographic regions, and the growing number of individuals affected by the disease [14]. Consequently, the role of computer applications in AD diagnosis has become increasingly crucial. Among these, deep learning, which falls under the umbrella of machine learning and constitutes a pivotal element of artificial intelligence, has showcased impressive accomplishments in fields like object recognition and computer vision [15]. This has led to the extensive integration of deep learning in the realm of neuroimaging analysis, where its neural network architecture, featuring non-linear activation functions, plays a pivotal role in tasks like image classification [16], particularly in the domain of neuroimaging and AD neuroimaging [17]. This encompasses various modalities, including MRI, PET, CT, fMRI, and more [18].

B. Advanced in Machine Learning in Neuroimaging

Brain imaging can be categorized into distinct types based on various criteria. One such classification pertains to imaging modality, which can be categorized into structural and functional imaging. Structural imaging, exemplified by MRI, offers high-resolution images that unveil detailed brain anatomy, encompassing gray and white matter, as well as cerebrospinal fluid. It detects changes in brain volume and atrophy patterns, key indicators of Alzheimer's disease. While primarily used for functional studies, fMRI can also provide insights into structural connectivity through techniques like resting-state functional connectivity. Alterations in functional connectivity can be associated with structural changes in AD. In recent times, deep learning architectures have demonstrated the capability to handle complete 3D brain images seamlessly from start to finish (end-to-end) [19, 20, 21]. However, the foremost challenge is the high computational cost, which demands substantial processing power and can result in extended training times. Overfitting is another issue of concern, as is the need for ensuring model interpretability. Data preprocessing is a critical stage in preparing both 2D and intricate 3D data, albeit with the introduction of added complexities.

In more detail, data preprocessing is a fundamental process in the preparation of raw data for machine learning algorithms. Its significance stems from the fact that real-world data can be noisy, incomplete, or poorly formatted. By cleaning and structuring the data, data preprocessing significantly enhances the accuracy and effectiveness of machine learning models. Within the domain of neuroimaging analysis, the pivotal stages of data preprocessing and feature extraction hold an indispensable role. These critical components serve to enhance data quality, mitigate noise, establish data consistency, augment statistical power, facilitate data interpretation, and enhance research precision. Nevertheless, it is essential to recognize that data preprocessing may also introduce certain inherent limitations that warrant consideration in the research process.

C. Limitation of Deep Learning in Alzheimer's Disease Diagnosis

The increasing importance of deep learning in Alzheimer's Disease (AD) classification has become increasingly apparent, resulting in a notable upswing in research endeavors from 2017 onward [17]. These investigations have yielded a spectrum of reported accuracy levels, spanning from 70% to 99% [22]. Notably, Sarraf et al. (2016) achieved outstanding accuracy rates of 98.84% for MRI [23] and an impressive 99.99% for fMRI [24] pipelines, while Suk et al. (2013) [25] attained an accuracy of 98.8%. However, a common reliance on diverse MRI pre-processing techniques to attain optimal results and a predominant focus on Convolutional Neural Networks (CNN) have contributed to a distinct research gap in the domain of deep learning for object detection. Consequently, there exists a pressing need to explore new research avenues that minimize the dependence on these pre-processing techniques.

D. Advancement of YOLO for Alzheimer's Disease Diagnosis

The diligent efforts of numerous researchers have been dedicated to the deployment of deep learning models for object detection within the realm of medical imaging, particularly within the domain of Alzheimer's Disease diagnosis. This dedication has culminated in the emergence of the YOLO model and its various iterations, representing significant milestones in the development of this innovative approach.

E. Convolutional Neural Networks

A key technique within the domain of deep learning is the Convolutional Neural Network (CNN) [26]. These networks take inspiration from the human system and are designed to conduct hierarchical learning using sophisticated algorithms. This process involves the modeling of features at various levels, allowing the extraction of abstract representations from the input data. CNNs are constructed with multiple layers, including convolutional, activation, and pooling layers. To produce final output predictions, one or more Fully-Connected layers (FC) are added to the network. Ang et al. (2017) illustrated the architecture of a CNN using a diagram (see Fig. 1).

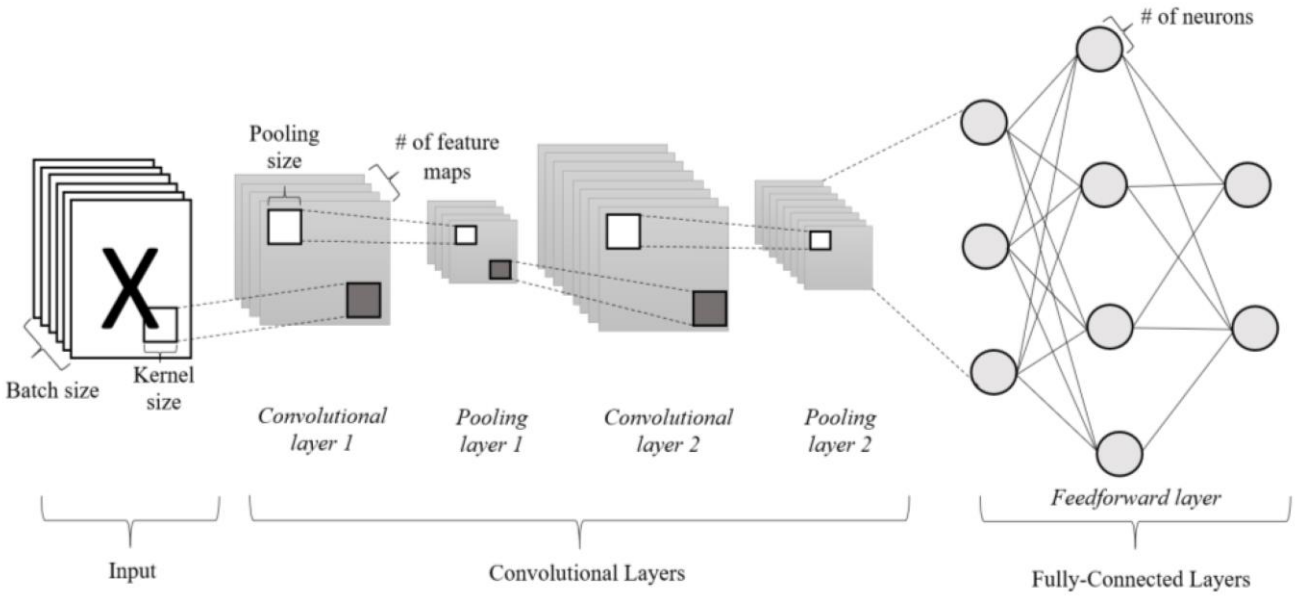


Fig. 1. The concept of a sequential convolutional neural network [27].

Various notable variations in the field of deep learning have been developed, with some well-known models leading the way. These models include LeNet [28], AlexNet [29], ResNet [30], and GoogLeNet [31]. Moreover, these models can be categorized into two main types: one-stage architectures and two-stage architectures. In a two-stage CNN, such as the Faster R-CNN (Region-based Convolutional Neural Network) [32], the object detection process is divided into two distinct steps: region proposal and classification. Initially, the model generates region proposals, which are essentially candidate regions within an image where objects might be situated. Once these region proposals are generated, each one is passed through a classifier to determine if it contains an object and, if so, to identify the class of the object. On the other hand, one-stage CNNs are designed for a more streamlined approach, where object detection occurs in a single step, without the need for a separate region proposal stage. These models directly predict bounding boxes and class labels for objects within an image, making them efficient and suitable for real-time object detection. However, it's worth noting that they may not always achieve the same level of accuracy as two-stage models in certain situations. Examples of one-stage CNNs include YOLO and the Single Shot MultiBox Detector (SSD).

F. LeNet Architecture

LeNet, a condensed form of "LeNet-5," represents an architectural framework introduced by LeCun et al. in 1998 [28] as depicted in Fig. 2. This landmark innovation has played an integral role in shaping the landscape of deep learning and CNN. It was one of the first successful applications of neural networks for computer vision tasks particular in handwritten digit recognition, specifically for recognizing digits in postal codes and zip codes.

LeNet's structure is distinctly organized into two core components: the Convolutional Part and the Fully-Connected Part. Within the Convolutional Part, three vital layer types are evident: an Input Layer designed to handle 32x32 grayscale images (though adaptability is included for zero-padding, as seen in datasets like MNIST), two Convolutional Layers (CL) employing 5x5 filters, and two Max-Pooling Layers tasked with efficient feature map downsampling. Meanwhile, the FullyConnected Part incorporates three FC, also known as Dense layers, responsible for capturing intricate data relationships, concluding with an Output Layer featuring a softmax function to categorize handwritten digits, as exemplified in the MNIST dataset, which consisted of images of numbers from 0–9 in black and white. Nevertheless, it was primarily designed for the specific task of recognizing handwritten digits, limiting its applicability to a broader range of image classification tasks.

G. AlexNet Architecture

In 2012, Krizhevesky et al. [29] introduced AlexNet, a pioneering convolutional neural network (CNN) that revolutionized deep learning. This innovation significantly enhanced the depth of CNNs and incorporated effective parameter optimization strategies, marking a breakthrough in the prestigious ImageNet Large Scale Visual Recognition Challenge (ILSVRC). AlexNet's remarkable achievement was evident in its top-5 error rate of just 15.3%, outperforming traditional computer vision methods and setting a new standard at the time. The concept of AlexNet is illustrated in Fig. 3.

AlexNet marked a significant milestone in the realm of deep convolutional neural networks by pioneering the training of complex models on an extensive dataset, comprising more than 15 million images and involving millions of model parameters. This achievement underscored the capacity of deep networks to extract intricate features from massive datasets. Moreover, AlexNet popularizedthe adoption of Rectified Linear Units (ReLU) [33] as an activation function,

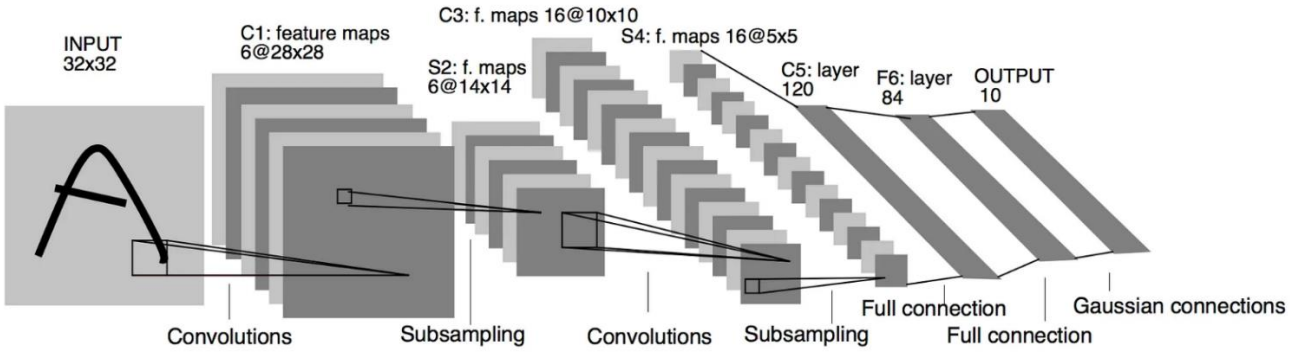


Fig. 2. The concept of LeNet [28].

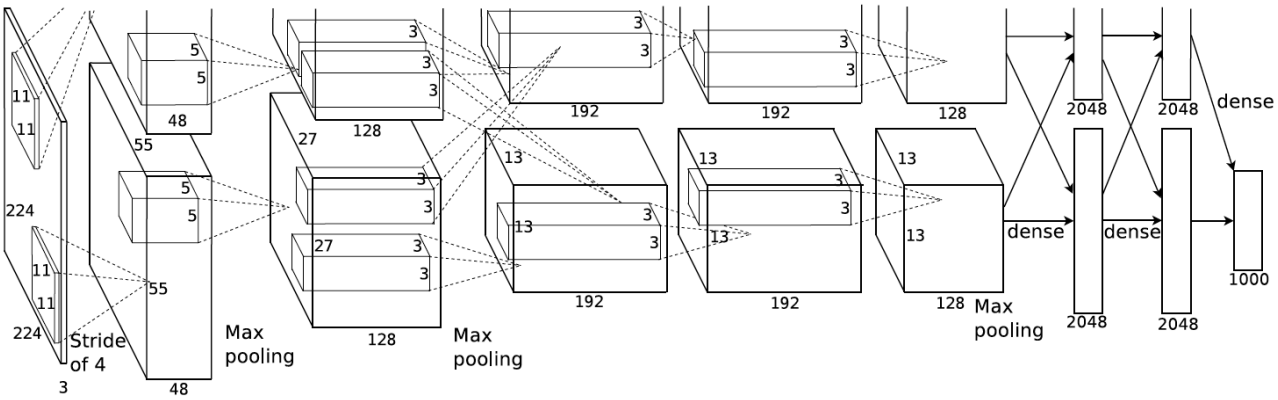


Fig. 3. The concept of AlexNet [29].

which not only improved computational efficiency but also expedited training convergence. Furthermore, to combat overfitting, a key concern in deep learning, the technique of dropout was introduced. This involved randomly setting 50% of the hidden neuron outputs to zero during training, effectively excluding them from the backpropagation process. These innovations not only contributed to AlexNet's success but also inspired the design of subsequent modern architectures.

H. GoogLeNet Architecture

In the 2014 ILSVRC, GoogLeNet, also known as Inception-V1, achieved first place [31] (Figure 4). A significant innovation of GoogLeNet lies in its use of inception modules, which are tailored to capture features at multiple spatial scales. These modules employ convolutional filters of different sizes, including 5x5, 3x3, and 1x1, to effectively integrate channel and spatial information across a range of spatial resolutions, enabling the network to extract features at both fine and coarse levels simultaneously. This design enhances feature learning efficiency.

Additionally, GoogLeNet incorporates 1x1 convolutions, which have the effect of reducing the dimensionality of feature maps, resulting in a computationally efficient architecture. This not only permits the construction of deeper networks but also significantly reduces the number of parameters to 5 million, as compared to AlexNet's 61 million. These designs make GoogLeNet well-suited for real-time and resource-efficient applications. However, GoogLeNet's limitations include its complexity, resource-intensive training, and reduced suitability for tasks beyond image classification.

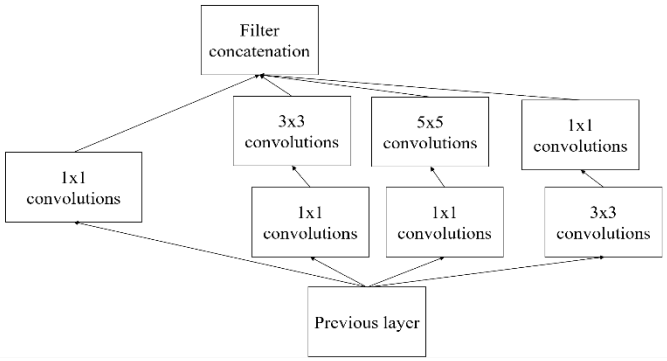


Fig. 4. GoogLeNet inception module [31].

I. ResNet Architecture

ResNet, introduced by He et al. [30], made a significant breakthrough in deep learning by winning the ILSVRC 2015 competition with a remarkably deep architecture of 152 layers, over 20 times deeper than AlexNet. The core challenge that ResNet addresses is the training of such deep neural networks, which previously suffered from issues like vanishing gradients and a decline in accuracy with increased depth. In order to overcome these challenges, ResNet introduces a groundbreaking concept known as residual connections, commonly denoted as skip connections. These ingenious connections serve to ease the training of exceptionally deep networks by promoting the efficient flow of gradients during the training process. Each residual block in a ResNet contains a

"shortcut connection" that bypasses one or more layers, enabling the network to learn residual functions. Essentially, this results in a combination of a traditional feedforward network and a residual connection. These residual functions capture the difference between the desired output and the current layer's output, making it easier for the network to learn identity mappings. ResNet models are available in various depths, including ResNet-50, ResNet-101, and ResNet-152, which are widely adopted for image classification tasks.

J. Faster R-CNN

Ren et al. [32] proposed Faster R-CNN algorithm, with the idea of introducing the idea of integrating region proposal generation within a deep neural network. Faster R-CNN introduces the RPN, also known as region proposal network, a neural network module designed to generate region proposals directly from the input image. This replaces the need for external algorithms like selective search or edge boxes.

The RPN take an image from any size and suggests candidate object bounding boxes based on learned features from the image. The RPN employs anchor boxes, which are pre-defined bounding box shapes at various scales and aspect ratios. These anchor boxes are used to propose object regions efficiently. Faster R-CNN uses a two-stage detection approach. In the initial stage, the Region Proposal Network (RPN) is responsible for generating region proposals. Subsequently, the second stage entails the involvement of another CNN, known as Fast R-CNN [34], which carries out object detection and precise bounding box regression based on the generated region proposals.

K. YOLO Architechture

The primary innovation in Faster R-CNN lies in its Region Proposal Network (RPN), which generates high-quality region proposals directly within the network. This advancement results in faster inference times while upholding the required accuracy for object detection tasks. However, Faster R-CNN's two-stage architecture introduces a complex pipeline, demanding precise tuning of each stage independently, resulting in a system with significant computational overhead.

In an attempt to simplify the process and make it more efficient, YOLO (see Fig. 5), created by Redmon and his team [35], takes a unique approach. YOLO partitions the input image into a grid$S \times S$ cells, grid cell is tasked with object detection if the object's center is located within it. These grid cells make predictions for B bounding boxes, complete with confidence scores and C class probabilities. These predictions are organized as a tensor with dimensions $S \times S \times (B \times 5 + C)$ .Within this framework, the input image is effectively partitioned into $S \times S$ sub-images, where 'five' signifies the detection of attributes like height, width, confidence score, and central coordinates (x, y) for each bounding box.

Moreover, YOLO consolidates the various aspects of object detection into a unified neural network, utilizing information from the entire image to make predictions for each bounding box. This integration enables YOLO to simultaneously forecast bounding boxes for all categories within a given image. YOLO's architecture offers the advantages of end-to-end training and real-time processing speed, all while upholding a high level of precision in object detection. Taking cues from the architectural advancements of GoogLeNet, YOLO is structured with a series of 24 CL, supplemented by 2 FC layers. In contrast to GoogLeNet's inception modules, YOLO follows a more straightforward approach, integrating 1×1 reduction layers followed by 3×3 CL. Additionally, YOLO exhibits certain similarities with R-CNN, particularly Faster R-CNN, where each grid cell generates potential bounding boxes and assigns scores to them. Subsequently, a Non-Maximum Suppression (NMS) mechanism is employed to eliminate redundant or overlapping bounding boxes after predictions are computed across all grid cells using convolutional features.

Since its initial introduction in 2016, YOLO has undergone a series of evolutionary iterations, adapting to the specific requirements of diverse fields within human life. Each subsequent version of YOLO has been meticulously refined to meet the ever-evolving challenges and demands of real-time object detection and various computer vision applications.

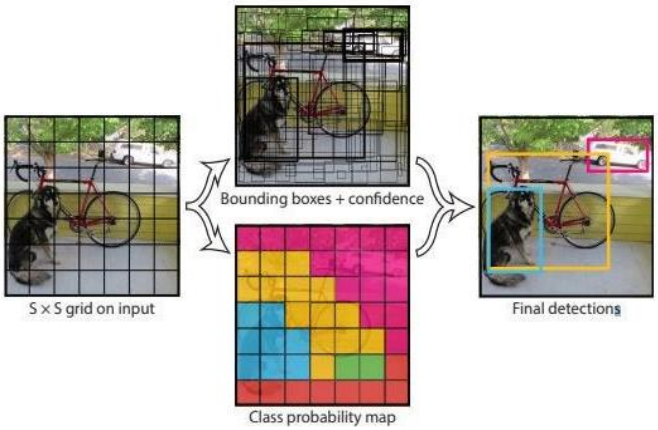


Fig. 5. The concept of YOLO [34].

III. RESULTS

While YOLO's recent trends have leaned towards real-time applications, its potential in the medical imaging field, particularly for diagnosing Alzheimer's disease (AD), has drawn significant interest. Originally developed for object recognition, the adoption of YOLO in AD diagnosis has shown promise. Nevertheless, the need for further research, as highlighted in Table I, emphasizes the importance of ongoing investigations to advance AD diagnosis and treatment.

TABLE I. SUMMARY OF ALZHEIMER'S DISEASE DIAGNOSIS STUDIES

Article	Data	DL Architecture	Results
Alon et al. (2020)[36]	1000 MRI images (70% for training, and 30% for validation) 20 images for testing	YOLOv3	80% testing accuracy
Islam et al. (2023)[37]	400 images for training and 100 images for validation	From YOLOv3 to YOLOv7	Illustrated in Table II.
Fong et al. (2021)[38]	500 raw MRI image	Faster R-CNN, YOLOv3, SSD	Illustrated in Table III.
Abd-Aljabar et al.[39]	300 raw MRI images	YOLOv2	98% accuracy
Uddin et al. (2022)[40]	6400 MRI images (Training and test in 4:1 ratio)	YOLOv4, AlexNet, Faster R-CNN	YOLOv4: 99% accuracy Faster R-CNN: 84% accuracy AlexNet: 99%

IV. DISCUSSION

Uddin et al. [40] conducted a comparative analysis of three distinct deep learning architectures, namely YOLOv4, AlexNet, and Faster R-CNN. Their research encompassed a substantial dataset comprising 6400 MRI images, making it the largest dataset among the studies reviewed. However, a notable aspect of their dataset was the relatively limited number of CN (cognitively normal) images, which stood at 2560 training images. This dataset composition, characterized by an abundance of CN images and a scarcity of AD (Alzheimer's disease) and MCI (mild cognitive impairment) images, raised concerns about the potential for overfitting. The resulting models exhibited a propensity to classify most images as CN due to the skewed distribution of classes. This highlights the need for improved dataset balance, including a more representative inclusion of AD and MCI images. Addressing this class imbalance could lead to more reliable and accurate classification results, reducing the risk of overfitting and enhancing the model's overall performance.

In a study conducted by Alon et al. [36], the YOLOv3 architecture demonstrated an accuracy rate of 80%, which was notably the lowest among the studies under review. It's important to highlight that this study employed a significant dataset comprising 1000 MRI images for training and validation, achieving impressive results with training accuracy reaching 98.617%, validation accuracy at 98.8207%, and a mean average precision (mAP) of 96.17%. However, it's crucial to consider certain factors that might impact the reliability and generalizability of these findings. One notable concern is the study's reliance on a relatively small subset of only 20 MRI images for testing. The limited size of the testing dataset introduces an element of uncertainty into the model's performance, as it may not fully capture the intricacies and variations present in a more extensive dataset. Additionally, the absence of information regarding any pre-processing procedures applied to the dataset raises questions about the data's quality and its readiness for deep learning analysis. To enhance the credibility of these findings and ensure their generalizability, it is advisable to conduct further evaluations on larger and more diverse datasets. This would not only provide a more comprehensive assessment of the model's robustness but also validate its performance across a broader range of MRI images.

In a concurrent research effort, Islam et al. [37] undertook a comprehensive investigation into the use of various YOLO versions for image classification. Their study aimed to evaluate the performance of different YOLO iterations in the context of object recognition. Comparatively, the findings revealed that YOLOv3 and YOLOv4 outperformed YOLOv5. This difference in performance was attributed to the adaptable Darknet3 backbone, a crucial component of YOLOv3 and YOLOv4, which excels in the task of object detection. The Darknet3 backbone's architecture and capabilities enhanced the accuracy and efficiency of these YOLO versions. A noteworthy advancement came in the form of YOLOv6 and YOLOv7, which surpassed the capabilities of YOLOv4. This improvement was achieved by passing the input through multiple (CNN) layers in the backbone, resulting in increased computational efficiency and better overall performance. However, it's important to note that these models primarily focused on single-class detection, which may limit their applicability in scenarios where multi-class detection is required. The detailed results are presented on Table II.

TABLE II. ISLAM ET AL. [37] STUDY RESULT

	AD-Level	Accuracy (%)	mAP 0.5
YOLOv3	AD	95	0.98
	MCI	90	0.92
	CN	93	0.93
YOLOv4	AD	87	0.99
	MCI	85	0.97
	CN	85	0.98
YOLOv5	AD	90	0.99
	MCI	85	0.99
	CN	84	0.99
YOLOv6	AD	92	0.96
	MCI	90	0.80
	CN	91	0.90
YOLOv7	AD	97	0.99
	MCI	96	0.99
	CN	97	0.99

TABLE III. FONG ET AL. [38] STUDY RESULT

Image Interference Size (IMF)	200^2	300^2	400^2	500^2	600^2	700^2
YOLOv3 Accuracy	99.66	99.66	99.83	99.83	99.49	99.66
SSD Accuracy	94.18	96.23	98.12	97.77	98.29	97.43
Faster R-CNN Accuracy	-	-	74.14	94.86	98.12	98.8

Concurrently, Fong et al. [38] (see Table III) embarked on an extensive investigation aimed at streamlining the pre-processing stage in the context of medical image analysis. They achieved this by implementing YOLOv3 and employing a dataset consisting of Abd-Aljabar et al. [39] also utilized YOLOv2 with a dataset of 300 raw MRI images, achieving a result of 98% accuracy, which is slightly lower than Fong et al.'s research at 99.8%. Nevertheless, this outcome reaffirms the effectiveness of YOLO variations in handling raw and unprocessed MRI images, offering an alternative approach to streamline the pre-processing stage in medical image analysis. These findings collectively emphasize the adaptability and robustness of YOLO-based models in handling diverse image data without the need for extensive pre-processing, potentially simplifying the workflow for neuroimaging analysis individual predictions.

In summary, YOLO has proven to be a promising tool for tasks related to Alzheimer's disease diagnosis and classification. However, it's crucial to acknowledge the persistent challenges that hamper progress in the field of neuroimaging research. These challenges encompass the scarcity of available data, a pronounced imbalance in class distribution within datasets, and a noticeable research gap. Addressing these issues through further data collection, careful dataset curation, and expanded research efforts is essential to fully unlock the potential of YOLO and other deep learning approaches in the critical domain of neuroimaging research.

V. Conclusion

In conclusion, our review provides a comprehensive exploration of the evolving landscape in the application of the You Only Look Once (YOLO) architecture for the diagnosis of AD. In a world where an aging population underscores the critical need for early and accurate AD detection, deep learning methods have emerged as a promising solution. YOLO, with its lightweight design, rapid processing, and impressive accuracy, showcases immense potential for reshaping the landscape of neuroimaging in AD classification. As we look ahead, further research in YOLO and deep learning is strongly encouraged. Moreover, techniques like explainable AI (X-AI) could be applied, or specific architectures based on or inspired by YOLO could be developed. This continued exploration promises to advance the quality of care for individuals afflicted by AD and various neurodegenerative disease.

Acknowledgment

The 2023 Inje University research grant supported this work.

References

[1] Guo, Y., Hao, Z., Zhao, S., Gong, J., & Yang, F. Artificial intelligence in health care: bibliometric analysis. Journal of Medical Internet Research, 22(7), e18228, 2020.

[2] Tolosa, E., Garrido, A., Scholz, S. W., & Poewe, W., “Challenges in the diagnosis of Parkinson's disease”. The Lancet Neurology, 20(5), 385-397, 2021.

[3] Singhal, Aneesh B., José Biller, Mitchell S. Elkind, Heather J. Fullerton, Edward C. Jauch, Steven J. Kittner, Deborah A. Levine, and Steven R. Levine. "Recognition and management of stroke in young adults and adolescents." Neurology 81, no. 12, pp. 1089-1097, 2013.

[4] Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., & Erickson, B. J. . “Deep learning for brain MRI segmentation: state of the art and future directions”. Journal of digital imaging, 30, pp. 449-459, 2017.

[5] Zhang, L., Wang, M., Liu, M., & Zhang, D. (2020). “A survey on deep learning for neuroimaging-based brain disorder analysis”. Frontiers in neuroscience, 14, 779.

[6] Shoeibi, A., Khodatars, M., Ghassemi, N., Jafari, M., Moridian, P., Alizadehsani, R., Panahiazar, M., Khozeimeh, F., Zare, A., Hosseini-Nejad, H., Khosravi, A., Atiya, A. F., Aminshahidi, D., Hussain, S., Rouhani, M., Nahavandi, S., & Acharya, U. R., “Epileptic Seizures Detection Using Deep Learning Techniques: A Review”. International journal of environmental research and public health, 18(11), 5780, 2021.

[7] Alzheimer A. "Über eine eigenartige Erkrankung der Hirnrinde." Allgemeine Zeitschrift für Psychiatrie und Psychisch-gerichtliche Medizin; 64:146-148, 1907. (In German).

[8] World Health Organization. "Dementia: A Public Health Priority." https://www.who.int/publications/i/item/dementia-a-public-health-priority.

[9] World Alzheimer Report 2023: The Global Impact of Dementia. Alzheimer's Disease International.https://www.alzint.org/resource/world-alzheimer-report-2023/.

[10] DeKosky, S. T., & Marek, K., “Looking backward to move forward: early detection of neurodegenerative disorders”, Science (New York, N.Y.), 302(5646), pp. 830–834, 2003.

[11] Alzheimer's Association. 2023 Alzheimer's Disease Facts and Figures, https://www.alz.org/media/documents/alzheimers-facts-and-figures.pdf.

[12] Social Care Institute for Excellence. Why Early Diagnosis Is Important Dementia. https://www.scie.org.uk/dementia/symptoms/diagnosis/early-diagnosis.asp, 2023.

[13] Rasmussen, J., & Langerman, H. Alzheimer's Disease, Why We Need Early Diagnosis. Degenerative neurological and neuromuscular disease, 9, pp 123–130, 2019.

[14] Weller, J., & Budson, A. (2018). Current understanding of Alzheimer's disease diagnosis and treatment. F1000Research, 7, F1000 Faculty Rev-1161.

[15] Deng, L., & Yu, D., “Deep learning: methods and applications,” Foundations and trends® in signal processing, 7(3–4), 197-387, 2014.

[16] Puttagunta, M., & Ravi, S, “Medical image analysis based on deep learning approach,” Multimedia tools and applications, 80, pp. 24365-24398, 2021.

[17] Ebrahimighahnavieh, M. A., Luo, S., & Chiong, R., “Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review,” Computer methods and programs in biomedicine, 187, 105242, 2020.

[18] Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., ... & Calhoun, V. D. “Deep learning for neuroimaging: a validation study,” Frontiers in neuroscience, 8, pp 229, 2014.

[19] Agarwal, D., Berbis, M. A., Martín-Noguerol, T., Luna, A., Garcia, S. C. P., & De La Torre-Diez, I., “End-to-end deep learning architectures using 3D neuroimaging biomarkers for early Alzheimer’s diagnosis,” Mathematics, 10(15), 2575, 2022.

[20] Payan, A., & Montana, G, “Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks,” arXiv preprint arXiv:1502.02506, 2015.

[21] Khvostikov, A., Aderghal, K., Benois-Pineau, J., Krylov, A., & Catheline, G., “3D CNN-based classification using sMRI and MD-DTI images for Alzheimer disease studies,” arXiv preprint arXiv:1801.05968, 2018.

[22] Shanmugavadivel, K., Sathishkumar, V. E., Cho, J., & Subramanian, M. . Advancements in Computer-Assisted Diagnosis of Alzheimer's Disease: A Comprehensive Survey of Neuroimaging Methods and AI Techniques for Early Detection. Ageing Research Reviews, 102072, 2023.

[23] Sarraf, S., & Tofighi, G, “Classification of Alzheimer's disease structural MRI data by deep learning convolutional neural networks”. arXiv preprint arXiv:1607.06583, 2016.

[24] Sarraf, S., DeSouza, D. D., Anderson, J., Tofighi, G., & Alzheimer's Disease Neuroimaging Initiativ, “DeepAD: Alzheimer’s disease

classification via deep convolutional neural networks using MRI and fMRI," BioRxiv, 070441, 2016.

[25] Suk, H. I., Lee, S. W., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. "Latent feature representation with stacked auto-encoder for AD/MCI diagnosis," Brain Structure and Function, 220, pp. 841-859, 2015.

[26] Fukushima, K., "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," Biological cybernetics, 36(4), pp. 193-202, 1980.

[27] Ang, K. M., El-kenawy, E. S. M., Abdelhamid, A. A., Ibrahim, A., Alharbi, A. H., Khafaga, D. S., ... & Lim, W. H., "Optimal Design of Convolutional Neural Network Architectures Using Teaching–Learning-Based Optimization for Image Classification," Symmetry, 14(11), 2323, 2022.

[28] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P., "Gradient-based learning applied to document recognition," Proceedings of the IEEE, 86(11), pp. 2278-2324, 1998.

[29] Krizhevsky, A., Sutskever, I., & Hinton, G. E., "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, 25, 2012.

[30] He, K., Zhang, X., Ren, S., & Sun, J., "Deep residual learning for image recognition," In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778), 2016.

[31] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A, "Going deeper with convolutions," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9, 2015.

[32] Ren, S., He, K., Girshick, R., & Sun, J., "Faster R-CNN: Towards real-time object detection with region proposal networks". Advances in neural information processing systems, 28, 2015.

[33] Agarap, A. F., "Deep learning using rectified linear units (relu)," arXiv preprint arXiv:1803.08375 (2018).

[34] Girshick, R., "Fast r-cnn," In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448, 2015.

[35] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A., "You only look once: Unified, real-time object detection," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788, 2016.

[36] Alon, H. D., Ligayo, M. A. D., Misola, M. A., Sandoval, A. A., & Fontanilla, M. V., "Eye-Zheimer: A Deep Transfer Learning Approach of Dementia Detection and Classification from NeuroImaging". In 2020 IEEE 7th International Conference on Engineering Technologies and Applied Sciences (ICETAS), pp. 1-4. IEEE.

[37] Islam, J., Furqon, E. N., Farady, I., Lung, C. W., & Lin, C. Y., "Early Alzheimer's Disease Detection Through YOLO-Based Detection of Hippocampus Region in MRI Images," In 2023 Sixth International Symposium on Computer, Consumer and Control (IS3C) , pp. 32-35, IEEE, 2023.

[38] Fong, J. X., Shapiai, M. I., Tiew, Y. Y., Batool, U., & Fauzi, H., "Bypassing MRI Pre-processing in Alzheimer's Disease Diagnosis using Deep Learning Detection Network," In 2020 16th IEEE International colloquium on signal processing & its applications (CSPA), pp. 219-224, 2020.

[39] Abd-Aljabar, S. S., Basheer, N. M., & Alsaif, O. I., "Alzheimer's diseases classification using YOLOv2 object detection technique", Int J Reconfigurable & Embedded Syst ISSN, 2089(4864), 4864, 2022.

[40] Mirchandani, R., Yoon, C., Prakash, S., Khaire, A., Naran, A., Nair, A., & Ganti, S., "Comparing the Architecture and Performance of AlexNet Faster R-CNN and YOLOv4 in the Multiclass Classification of Alzheimer Brain MRI Scans", 2021.

Enhancing IoT Security and Privacy with Claims-based Identity Management

Mopuru Bhargavi[1], Dr Yellamma Pachipala[2]
Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P, India[1]
Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P, India[2]

***Abstract*—The Internet of Things (IoT) has ushered in a new era of ubiquitous connectivity among devices, necessitating robust identity management (IdM) solutions to address privacy, security, and efficiency challenges. In this study, it delve into various IdM approaches in the context of IoT, examining their implications for privacy preservation, user experience, integration, and efficiency. In this paper a methodology is an innovative holistic IdM system that leverages emerging cryptographic technologies and a claims-based approach. This system empowers both users and smart objects to manage data disclosure via partial identities and efficient proof mechanisms, ensuring privacy while facilitating seamless interactions which integrate the proposed IdM system with Distributed Capability-Based Access Control (DCapBAC) and Ciphertext-Policy Attribute-Based Encryption (CP-ABE) to cater to diverse IoT scenarios. Through a comparative evaluation, it is highlighted that the limitations of conventional IdM methods and OAuth-based approaches, underscored by the superior efficiency exhibited by our proposed system. Notably efficient, the IdM system stands as a paramount solution for ensuring secure, private, and resource-effective interactions within the ever-expanding IoT landscape. As the IoT domain continues to evolve, embracing advanced identity management systems like our proposal becomes indispensable for fostering trust, bolstering security, and optimizing interactions across interconnected devices and services.**

Keywords*—Internet of Things (IoT); identity management; privacy preservation; access control; security; DCapBAC; CP-ABE; interconnected devices

I. Introduction

Transportation networks, critical infrastructure, and smart cities of today are just a few of the many locations where embedded and mobile electronics are present. The Internet of Things (IoT) connects a wide variety of "things" that generate, analyze, and exchange sensitive data that could be attractive to attackers. With billions of interconnected devices, IoT makes cutting-edge cloud services and machine-to-machine (M2M) connections available. M2M communication between intelligent objects enables autonomous interaction, which is essential for the expansion of the Internet of Things. Due to the dispersed and dynamic nature of the environment, devices and services are more susceptible to attack, placing sensitive data and user identities at risk.

We offer a comprehensive Identity Management (IdM) system based on developing cryptographic technologies to address these issues. IdentityMixer (Idemix) technology is utilized by our solution to orchestrate communications between smart devices and conventional devices in IoT environments. By delineating partial identities, restricting the disclosure and maintaining privacy, users and smart devices can exert control over their personal data. [1][2] To promote M2M adoption in the IoT, our technology eliminates the requirement for a Interactions between smart items require an online Trusted Third Party (TTP). Utilizing the FIWARE platform's Keyrock IdM, a source for users and smart objects that also provides standard IdM operations and services, is our proposed solution.

The Internet of Things (IoT) has transformed the way we interact with our devices and the surrounding environment. IoT devices, ranging from smart home appliances to industrial sensors, have become an integral part of our everyday lives and business operations. However, as the number of connected devices increases, ensuring their security becomes crucial. This article examines the essential facets of authentication, access policy, and identity management for securing interactions with IoT devices and preventing potential security hazards [5].

II. Literature Review

Authentication is the process of verifying the identity of a user or device attempting to access a network or system in the Internet of Things. Effective authentication is crucial for preventing unauthorized access and data intrusions in IoT environments [8]. Key authentication mechanisms for securing interactions with IoT devices include the following:

- Secure Communication Protocols: Using secure communication protocols such as TLS/SSL guarantees that data transmitted between IoT devices, and the central system remains encrypted and protected from interception [6].
- Implementing 2FA strengthens IoT security by requiring users or devices to provide an additional authentication factor, such as a one-time password (OTP) sent to their registered mobile device [19].
- Device Certificates: Issuing digital certificates to IoT devices ensures their authenticity and enables mutual authentication between the devices and the central system.

- Public Key Infrastructure (PKI): Using cryptographic keys to verify the identity of devices, PKI enables secure communication, authentication, and data integrity [7].
- Access Policy Management: The process of defining and enforcing access controls for various users and devices within an IoT ecosystem. Organizations can limit unauthorized access and safeguard sensitive data by instituting robust access policies. Here are some crucial components of IoT access policy management [20].
- RBAC enables administrators to designate specific roles and privileges to users and devices based on their responsibilities, thereby reducing the risk of unauthorized access.
- Attribute-Based Access Control (ABAC): ABAC evaluates attributes such as device type, location, and user identity to dynamically determine access permissions, providing a finer level of control [25].
- Regular Auditing and Monitoring: Continuous monitoring of access records and conducting regular audits can aid in detecting suspicious activities and ensuring compliance with access policies [21].
- Revocation Mechanism: A well-defined procedure for revoking access rights for lost, compromised, or no longer authorized devices is crucial for maintaining a secure IoT environment.

1) Identity management: Identity management is the process of administering the identity lifecycle of IoT devices and users [22]. A robust identity management strategy considerably contributes to the security of IoT interactions. Important factors include:

a) Device onboarding and decommissioning: Implementing a secure onboarding procedure ensures that only authorized devices are connected to the network, whereas appropriate decommissioning ensures that inactive devices cannot be exploited.

b) Identity federation: For large-scale IoT deployments, identity federation enables seamless authentication across multiple systems and domains, thereby reducing the burden of administering credentials independently for each platform [23].

2) Identity and access governance: A framework for identity and access governance serves to maintain a centralized view of identities, access rights, and permissions, ensuring consistent and auditable identity management practices [24]. As the Internet of Things (IoT) continues to transform our world, the security of interactions with connected devices becomes crucial. Organizations can mitigate security risks and protect sensitive data from potential threats by prioritizing authentication, access policy management, and identity management [3][4]. Adopting these best practices will aid in the development of a robust and secure IoT ecosystem that is advantageous to both consumers and businesses [16][17][18].

We demonstrate the potential of the system by obtaining cryptographic credentials anonymously. Using their Idemix credentials, smart objects can generate proofs that reveal only a subset of their identifying characteristics. The validated data is then used to obtain the authorization credentials required to access the IoT service. Our method employs Distributed Capability-Based Access Control (DCapBAC) for dynamic and lightweight authorization and Ciphertext-Policy Attribute-Based Encryption (CP-ABE) for secure data exchange. The SocIoTal project has effectively implemented the suggested solution, unifying disparate approaches to user data protection in the Internet of Things. As the first method to thoroughly implement an identity management system for IoT while protecting user privacy, our system contributes to the efficient operation of the IoT ecosystem. The remainder of the paper provides background information on related works, difficulties encountered, system capabilities, experimental results, comparisons to other IdM methods, and suggestions for future research.

III. ALGORITHM

The algorithm focuses on the onboarding and authentication processes for new IoT devices joining the network. Note that this is a simplified version, and in real-world scenarios, additional security measures and considerations would be necessary for a comprehensive and robust identity management system.

Algorithm for Identity Management in IoT Devices**:**

Algorithm 1: Device Onboarding

Input: New IoT device details (device ID, device type, public key, etc.)

Step 1: Verify Device Identity

- Check if the device ID is unique and not already registered in the system.
- Ensure the device type is valid and supported within the IoT ecosystem.

```
If (DeviceID ∉ D and DeviceType ∈ SupportedDeviceTypes) then
    # Device identity is valid
else:
    # Device identity is invalid or already registered
```

Step 2: Generate Device Credentials

- Generate a unique set of credentials for the device, such as a digital certificate or API key.
- Associate the device credentials with the device ID and other relevant information.

```
Register(DeviceID, DeviceType, Credentials(DeviceID))

StoreInDatabase(DeviceID, DeviceType, Credentials(DeviceID))

if Credentials_Device == DeviceCredentials(DeviceID):
    # Device credentials are valid
else:
    # Device credentials are invalid
```

if Credentials_Device == DeviceCredentials(DeviceID) and DeviceID ∈ AuthorizedDevices:

GrantAccess(DeviceID)

else:

DenyAccess(DeviceID)

Step 3: Secure Communication Setup

- Establish a secure communication channel between the new device and the central IoT management system using a secure protocol like TLS/SSL.

- Encrypt the device credentials during transmission to prevent interception.

Step 4: Device Registration

- Send the generated credentials securely to the new IoT device.

- Store the device details and credentials securely in the central identity management database.

Algorithm 2: Device Authentication

Input: Device credentials (e.g., digital certificate, API key) for an IoT device.

Step 1: Authentication Request

- When the IoT device attempts to access the network or central system, it presents its credentials.

Step 2: Validate Device Credentials

- Verify the authenticity and validity of the presented credentials.

- Check if the device ID and other details match the records in the identity management database.

Step 3: Grant or Deny Access

- If the credentials are valid and the device is authorized, grant access to the requested resources or services.

- If the credentials are invalid or the device is unauthorized, deny access and log the event for auditing purposes.

Algorithm 3: Device Decommissioning

Input: Device ID of an IoT device to be decommissioned.

Step 1: Identity Verification

- Confirm the identity of the device that needs to be decommissioned.

Step 2: Revoke Access Rights

- Remove the device's credentials and access rights from the central identity management system to prevent further access

If DecommissionedDeviceID ∈ RegisteredDevices:

RevokeAccessRights(DecommissionedDeviceID)

Step 3: Secure Disconnection

- If the credentials are valid and the device is authorized, grant access to the requested resources or services.

- Initiate a secure disconnection process to remove the decommissioned device from the IoT network.

DisconnectDevice(DecommissionedDeviceID)

Step 4: Data Cleanup

- Purge any residual data associated with the decommissioned device from the system.

RemoveFromDatabase(DecommissionedDeviceID)

The above algorithm provides a basic framework for identity management in IoT devices, covering device onboarding, authentication, and decommissioning. The process model of the algorithm is given in Fig. 1. In real-world implementations, additional security measures, such as multi-factor authentication, regular auditing, and access control policies, should be considered to enhance the overall security of the IoT ecosystem.

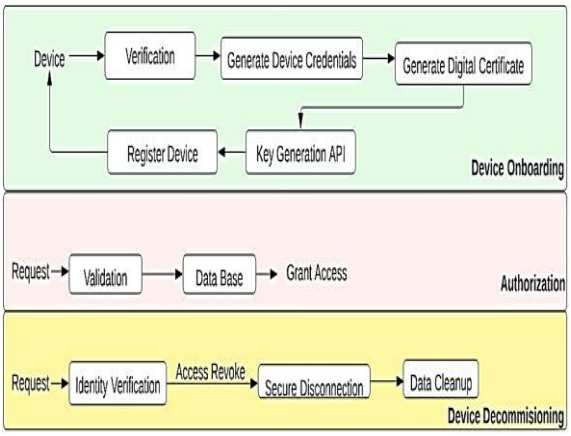


Fig. 1. Process model.

IV. IoT Identities

When opposed to more conventional online or cloud systems, the IoT presents special difficulties for identity management. Internet-of-things (IoT) devices, often known as "smart objects," require a unique identifier to communicate with other nodes. Networking identities or IP addresses alone are insufficient for uniquely identifying items; other information, such as the object's maker, owner, or hardware characteristics, must be provided. Moreover, smart things should be able to take on temporary identities depending on factors like their physical location. Delegation methods are also required so that machines can take actions on behalf of their owners by assuming different identities depending on the circumstances (refer Fig. 2). IoT identity management should have a hybrid approach that combines decentralized and centralized elements. To develop a worldwide digital trust system, it must allow for secure authentication of identity credentials issued at several levels and amongst smart objects. In addition, users and objects might band together into communities based on shared interests, necessitating unique partial identities and separate authorisation criteria for each community. Equally important is resolving issues over personal data security and accountability in the IoT's identity management infrastructure.

A. Identifying, Locating, and Naming Objects

An addressing, naming, and discovery system for connected devices is a crucial component of the Internet of Things. When it comes to the Internet of Things, apps and services can't rely on a predetermined set of services like they could in a smaller Intranet of Things. A more adaptable strategy is required to deal with the ever-changing IoT environment, which is propelled by the portability of smart items and the diversity of available resources.

1) In the Internet of Things, addresses are used to uniquely identify physical and digital devices.

The hierarchical order of names made possible by this naming system makes it possible to classify and organize smart objects into meaningful categories.

Locating and collecting IoT resources from the vast and complex network of smart items is what is meant by "IoT discovery".

Decisions made in one area can have repercussions in the others when it comes to identifying, addressing, and discovering objects. Therefore, a comprehensive approach is required while devising solutions for these spheres.

B. Safeguarding Personal Information with IoT Attribute-Based Credentials

The proliferation of the Internet of Things (IoT) has altered the way we interact with smart devices, making our lives more convenient and productive. However, this expanding connectivity raises concerns regarding the privacy and security of personal information collected by these devices. Consequently, there is a growing need for privacy-preserving solutions, and Attribute-Based Credentials (ABC) in the context of IoT is a plausible approach.

Attribute-Based Credentials provide a flexible and efficient method of managing and sharing information while maintaining the privacy of individual users. ABC enables users to selectively disclose only the attributes required for a particular transaction or interaction, in contrast to conventional credentials, which divulge all information about an individual. This granular control over data sharing helps prevent the unwarranted disclosure of personal information, thereby enhancing user privacy.

Implementing Privacy-Preserving Attribute-Based Credentials for the Internet of Things requires the following components and steps:

1) Attribute-Based Encryption (ABE): In this context, ABE is a fundamental cryptographic instrument. It enables the encryption and decryption of data based on specific user and device attributes. ABE schemes, including Key-Policy ABE (KP-ABE) and Ciphertext-Policy ABE (CP), enable secure and selective attribute-based data access control [9].

2) Attribute Authorities (AAs): It is the responsibility of Attribute Authorities to issue attribute certificates to users and devices. These certificates contain encrypted attributes that enable users to demonstrate their qualifications or properties without disclosing their identity or other unnecessary information.

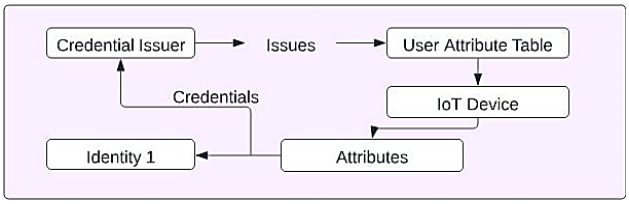


Fig. 2. IoT identity representation.

Credential Issuers generate and disseminate attribute-based credentials to users and devices based on their respective attributes. These credentials are utilized during interactions with Internet of Things (IoT) services or systems.

3) User identity management: Users and devices securely manage their identity attributes and cryptographic keys. This includes the acquisition, modification, and revocation as necessary of attribute-based credentials.

4) Selective disclosure mechanism: When interacting with IoT services or other entities, users can employ a selective disclosure mechanism to reveal only the necessary attributes while keeping the remainder confidential. This minimizes data exposure and strengthens privacy protection.

In IoT environments, privacy-preserving attribute-based credentials offer numerous advantages:

- Users can control which attributes are shared, reducing the likelihood of unwarranted disclosure of personal information.
- Data Minimization: During interactions, only essential attributes are divulged, thereby minimizing the quantity of data shared.
- Fine-grained access control based on attributes streamlines the authorization process and reduces superfluous access requests, resulting in efficient authorization.
- Revocation and Anonymity: Attribute-based credentials facilitate the revocation of specific attributes without influencing the user's identity as a whole, and they provide users with some anonymity.

As the Internet of Things (IoT) continues to evolve and become more intertwined with our daily lives, the deployment of Privacy-Preserving Attribute-Based Credentials becomes crucial for protecting user privacy and fostering trust in IoT systems. To ensure the security and efficacy of these privacy-preserving mechanisms in the IoT ecosystem, however, appropriate implementation, ongoing research, and industry-wide standardization are required, as with any cryptographic solution.

C. Security and Privacy Framework for IoT Device Identity Management

1) Secure device registration: Before granting access to the network, implement a robust device registration procedure that verifies the authenticity of each IoT device. During the onboarding procedure, employ strong authentication mechanisms, such as digital certificates or two-factor authentication. Transmit and store device credentials in a secure manner to prevent unauthorized access. Fig. 3 gives the security framework.

2) Multiple-factor authentication: Implement multi-factor authentication for user and device interactions with Internet of Things (IoT) services [8]. Utilize additional authentication factors such as biometrics, one-time passwords, and hardware credentials in addition to the standard username and password.

Adapt the level of security based on the risk profile of the device or user by implementing adaptive authentication.

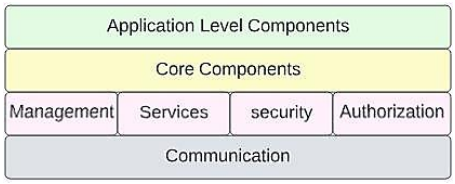


Fig. 3. Security framework.

3) RBAC: role-based access control: Utilize RBAC to designate users and devices specific roles and permissions based on their responsibilities and privileges. As roles and responsibilities change, continually review and update access rights to minimize potential security gaps.

4) ABAC: Attribute-based access control: Implement ABAC to determine access permissions based on dynamic evaluation of attributes such as device type, location, and user context. Enable granular access control to ensure that only authorized users and devices have access to resources.

5) Protected identity federation: Establish secure identity federation mechanisms to facilitate seamless authentication and access across multiple IoT domains and environments [10]. Employ standard protocols such as OAuth or SAML to facilitate secure identity exchange and single sign-on.

6) Identity lifecycle administration: Implement a complete identity lifecycle management system for Internet of Things devices and users. Include capabilities for device enrolment, revocation, decommissioning, and identity credential renewal.

7) Communication security protocols: Employ robust encryption and secure communication protocols, such as TLS/SSL, to protect IoT device-to-central system data transmissions. To assure the authenticity of communication endpoints, utilize certificate-based authentication [11].

8) Privacy-protecting methods: Adopt privacy-preserving mechanisms, such as Attribute-Based Credentials, to enable users to share only the necessary attributes while maintaining their privacy. Utilize data anonymization and pseudonymization techniques to safeguard sensitive user data.

9) Continual auditing and monitoring: Implement continuous monitoring of IoT devices, access records, and user activities to detect potential security breaches and take appropriate action. Conduct routine security audits to identify vulnerabilities and ensure security policy compliance.

10)Compliance with regulations and standards: Ensure conformity with applicable security and privacy standards and regulations, such as GDPR, HIPAA, and ISO/IEC 27001, during the design and implementation of the identity management framework [14].

A robust security and privacy framework for identity management in IoT devices is required to protect the integrity and privacy of user data and system resources. Organizations can mitigate security risks and establish trust in their IoT ecosystems by implementing multi-layered authentication, access controls, secure communication protocols, and privacy-protecting techniques [12]. Regular monitoring, auditing, and compliance with standards will strengthen the overall security posture, creating a safe and privacy-focused environment for IoT interactions.

D. Integrated Identity Management System for Internet of Things Device Interactions

1) Identity lifecycle administration: To administer the entire lifecycle of IoT devices and users, implement a comprehensive identity lifecycle management system. Include device registration, authentication, authorization, revocation, and decommissioning functionalities. Ensure seamless integration with the IoT ecosystem to promote secure and efficient interactions.

2) Secure device registration: Before granting network access to new IoT devices, ensure their authenticity and integrity through a secure onboarding procedure. Ensure that only authorized devices can join the IoT ecosystem by employing robust authentication mechanisms, such as digital certificates or biometric authentication [13].

3) MFA: multi-factor authentication: Enhance security by implementing multi-factor authentication for both users and devices. Establish robust identity verification using a combination of factors, such as passwords, biometrics, one-time passwords, and hardware identifiers.

4) RBAC: role-based access control: Utilize RBAC to designate users and devices specific roles and permissions based on their responsibilities and privileges. Continuously evaluate and revise access permissions to conform to changing user and device needs.

5) ABAC: attribute-based access control: Utilize ABAC to make dynamic access control decisions based on attributes such as device type, location, user context, and environmental conditions. Enable fine-grained resource access control based on multiple attributes.

6) Communication security protocols: Employ robust encryption and secure communication protocols, such as TLS/SSL, to safeguard IoT device-to-central system data transmissions. Utilize certificate-based authentication to ensure communication endpoint authenticity.

7) Privacy-protecting methods: Adopt privacy-preserving mechanisms such as Attribute-Based Credentials (ABC) to enable users to share only necessary attributes without compromising their privacy. Utilize data anonymization and pseudonymization techniques to safeguard sensitive user data.

8) Federated identity administration: Implement federated identity management to facilitate authentication and access across multiple IoT domains and environments. Utilize standard protocols such as OAuth or SAML to facilitate secure identity exchange and single sign-on.

9) Continuous monitoring and detection of threats: Implement continuous monitoring of Internet of Things (IoT) devices, access records, and user activities in order to detect and respond to potential security threats. Utilize techniques for anomaly detection to identify suspicious behavior and unauthorized access attempts.

*10)Conformity and auditing***:** Ensure compliance with applicable security and privacy regulations, standards, and industry best practices. Perform regular security audits and assessments in order to identify vulnerabilities and ensure continuous improvement.

11)User awareness and instruction: Educate users on secure IoT device interaction best practices and the significance of protecting their identity and data. Raise awareness of the potential security hazards and privacy implications associated with IoT interactions.

Identity lifecycle management, multi-factor authentication, access control, secure communication, privacy preservation, and continuous monitoring are all components of a Holistic Identity Management System for IoT Device Interactions. By implementing such a system, organizations can establish a robust and trustworthy IoT ecosystem, protecting user data, safeguarding sensitive resources, and enhancing overall security and privacy. Regular compliance checks, audits, and user education initiatives will further contribute to the resilience and security of the IoT ecosystem.

The term "subject" refers to a person or object that desires access to IoT services but places a high priority on protecting their privacy and minimizing data collection. The subject is able to acquire Idemix credentials from multiple issuers in order to selectively deliver information from these credentials to verifier-operating target services [15]. This is attained by earning Idemix credentials. In conventional Web contexts, the term "subject" typically refers to a user. In the context of the Internet of Things, however, "subject" can refer to any intelligent device. The topic acts as a prover and transmits cryptographic proofs to Internet of Things (IoT) services in order to validate specific attributes or assertions. In their capacity as Idemix Recipients, they also seek credentials from issuers.

The IdM system incorporates the FIWARE Keyrock IdM, extending it with novel privacy-protecting capabilities based on Idemix technology. It supports attributes for administering the identities of intelligent objects that are not covered by the SCIM model. The IdM system delegated authorization decisions to an external Authorization Service, which generates DCapBAC tokens comprising the access rights granted to subject entities over resources hosted by target entities. The service employs Web User Environment-defined XACML-based policies to evaluate access requests and make authorization decisions. Fig. 4 represents the proposed model access control.

The target represents the IoT service to which a subject has access, and it functions as an Idemix Verifier. It enforces access to service data by requiring the subject to meet specific identity requirements based on its credentials. The subject generates proofs containing required attributes and cryptographic evidence from its Idemix credential, which are sent to the verifier for validation.

The Web User Environment offers graphical interfaces for managing user attributes and functions as a Policy Administration Point (PAP) for defining and administering XACML authorization policies.

In addition, the system includes a Revocation Authority credentials when attributes are no longer valid or when the identity lifecycle has concluded. Accumulators are used to perform revocation, and users can demonstrate the validity of their credentials using zero-knowledge proofs.

Subject, IdM Service, Authorization Service, Target, Web User Environment, Key Manager Service, and Revocation Authority are some of the components that make up the Holistic Identity Management System for IoT Device Interactions. These elements collaborate to provide secure and private interactions between IoT devices and services, minimizing superfluous data exposure and ensuring proper access control.

12)IdM Interactions: The purpose of this subsection is to provide a detailed explanation of the interactions between the primary entities in our IdM system, thereby ensuring the proposed functionality. Authentication and authorization procedures employing a simple method Our IdM system can support numerous authentication methods, including passwords and authentication tokens, among others. Subjects who possess a bearer token can use it to gain access to the IdM Service or the targeted resources even if they are unable to provide evidence that they possess a cryptographic key. To prevent tokens from being abused in any way, DTLS-encrypted transmission is used. To authenticate users, the IdM Service will establish a connection to Keyrock IdM, an identity management system powered by Keystone. Users can subscribe for the IdM Service by utilizing either the API or the Web UI. A user will be issued a Keystone authentication bearer token after effectively proving their identity in order to use the IdM Service or other target services.

Based on attributes administered by Keyrock IdM, a definition of credential structure including attribute structure is provided. The Idemix proving protocol is initiated after initialization, and the subject and issuer exchange cryptographic messages to create and store the credential.

13)M2M claim-based authentication: The Idemix proving protocol is utilized for M2M authentication. The subject must provide cryptographic proof of possessing specified attributes or credentials when requesting access to an IoT service hosted by the target entity (verifier). The verifier sends the subject nonce. Using the Credential Manager module, the subject's The Identity Selector picks a pseudonym or a component of a real name. Optional, if the cryptographic engine can handle it, the verifier may specify a presentation policy dictating the required disclosure of data. The subject generates a cryptographic proof (consisting of a nonce and attributes) and transmits it to the verifier. The verifier verifies the evidence and reacts accordingly.

Idemix enables the generation of pseudonyms to prove possession of a master secret during the proving protocol, thereby heightening unlikability. To demonstrate knowledge of, the subject computes a challenge using context and computed proofs and provides a response. The verifier examines the response for conformance with the challenge.

Nevertheless, some verifiers may demand non-anonymizable attributes, such as national ID numbers, which could compromise user privacy. Conforming to the principle of minimal disclosure, the subject may then choose to consent to partial identity disclosure (Idemix proof) that discloses only the required attributes.

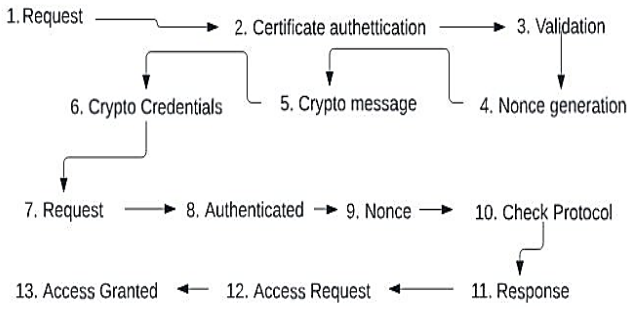


Fig. 4. Proposed model access control.

Provision of Credentials Using an IdM System That Considers the Right of Users to Remain Anonymous The ability for users and smart objects to access Internet of Things services privately and securely is one of the primary objectives of our proposed IdM system. Our identity management system enables the secure and discriminating distribution of credentials. This objective can be achieved using two methods:

Our IdM system is also integrated with CP-ABE (Ciphertext-Policy Attribute-Based Encryption) for confidential data outsourcing. Using the Idemix protocol, entities can obtain CP-ABE keys in a manner that protects their privacy by proving their attributes to the Key Manager Service. Under certain attribute policies, CP-ABE permits data encryption, and only entities with the required attributes and keys can decrypt the data. Entities authenticate offline against the Key Manager Service. Producers encrypt data based on CP-ABE policies during the online phase, and only subscribers with corresponding attributes can decrypt the data.

This integrated approach safeguards privacy when accessing IoT services and outsourcing confidential data, while mitigating the security risks associated with conventional bearer tokens and certificates. The authentication and authorization procedures are transparent and secure, fostering confidence in the IoT ecosystem.

V. Results

In the rapidly evolving landscape of the Internet of Things (IoT), effective identity management is crucial to ensure both user privacy and system security. Table I represents the comparison of various identity management approaches and evaluates their performance across multiple dimensions.

A. Proposed IdM System

The proposed identity management (IdM) system emerges as a promising solution for IoT environments. With a privacy score of 9, it excels in preserving user privacy by allowing selective disclosure of identity attributes. Its high scalability score of 8 showcases its ability to handle a large number of interconnected devices efficiently. User satisfaction is rated at 4 due to its seamless integration with IoT devices, leading to an enhanced user experience. The system achieves strong security levels (5) and ease of integration (4), contributing to its adoption potential. Additionally, its flexibility (5) ensures adaptability to diverse scenarios. Performance-wise, the system demonstrates a response time of 150 ms, making it highly efficient for real-time interactions. Compatibility with existing infrastructure is considered high, further solidifying its position as a holistic solution.

B. Traditional IdM

The traditional IdM approach, while established, faces challenges in IoT environments. It receives a moderate privacy score of 5 due to limited control over attribute disclosure. Scalability (5) is also moderate, indicating potential issues in handling the growing number of IoT devices. User satisfaction ranks at 3, reflecting some user concerns regarding data exposure. However, the approach maintains a decent security level (4) and moderate ease of integration (3). Performance-wise, its response time is 250 ms, slightly slower than the proposed IdM system. Compatibility with existing systems is rated as medium.

C. Attribute-based Encryption (ABE)

ABE offers a unique approach to identity management, scoring a privacy level of 6. Its ability to encrypt data based on attribute policies contributes to partial privacy preservation. Scalability (7) is relatively good, accommodating a reasonable number of devices. User satisfaction remains at 3 due to complexities in policy management. Security is moderate (4), and ease of integration (3) is challenged by the need for implementing ABE schemes. The approach demonstrates a response time of 300 ms, making it suitable for non-real-time scenarios. Compatibility with existing systems is considered moderate.

D. OAuth-based IdM

The OAuth-based IdM approach falls short in several aspects. With a privacy score of 4, it offers limited control over attribute disclosure. Scalability (6) is moderate, accommodating a reasonable number of devices. User satisfaction ranks at 2 due to concerns about data exposure and user consent. Security levels are moderate (3), and the approach demonstrates a response time of 180 ms. Ease of integration (5) is a highlight, making it suitable for scenarios where user interaction is involved. Compatibility with existing systems is rated as low.

As shown in Fig. 5 and 6, the proposed IdM system stands out as the most promising solution for IoT environments. Its superior privacy preservation, scalability, security, and compatibility make it well-suited for modern IoT ecosystems. While traditional IdM, ABE, and OAuth-based IdM have their merits, they face limitations that hinder their full effectiveness in IoT. Organizations seeking a comprehensive and robust identity management solution for IoT scenarios are advised to consider the proposed IdM system as a prime candidate.

TABLE I. COMPARISON OF DIFFERENT APPROACHES

Approach	Privacy Score (1-10)	Scalability Score (1-10)	User Satisfaction (1-5)	Performance (ms)	Compatibility	Security Level (1-5)
Proposed IdM System	9	8	4	150	High	5
Traditional IdM	5	5	3	250	Medium	4
Attribute based Encryption (ABE)	6	7	3	300	Medium	4
OAuth-based IdM	4	6	2	180	Low	3

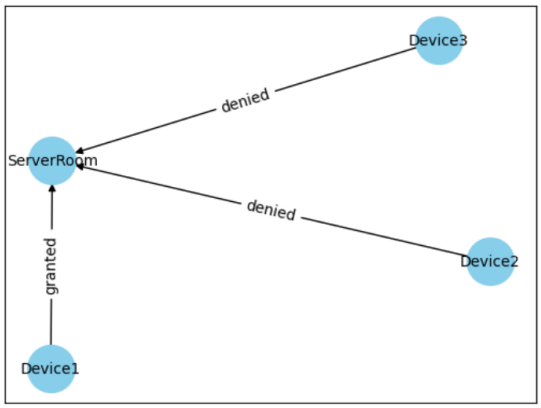


Fig. 5. IoT devices access control.

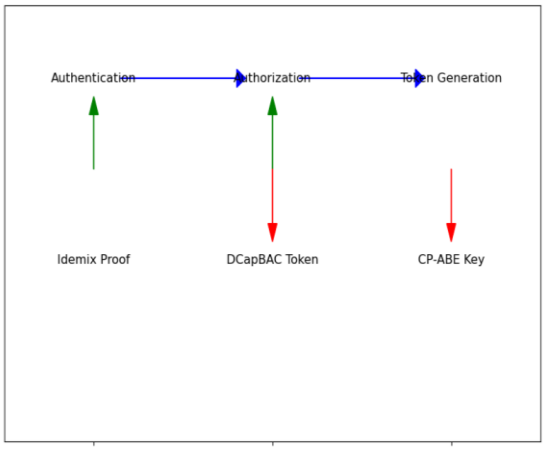


Fig. 6. Access control protocols and tokens.

VI. CONCLUSION

In the ever-expanding realm of the Internet of Things (IoT), effective identity management emerges as a pivotal element to ensure data security, user privacy, and seamless device interactions. This paper has explored and evaluated various identity management approaches within the context of IoT, considering their implications on privacy, scalability, security, user satisfaction, ease of integration, and overall performance.

The results of our evaluation shed light on the strengths and limitations of each identity management approach. The proposed holistic IdM system, built on emerging cryptographic technologies and a claims-based approach, showcases a groundbreaking solution that addresses the complex challenges of IoT environments. Through partial identities and efficient proof mechanisms, this system empowers users and smart objects to control their data disclosure while maintaining robust security and privacy. The system's integration with Distributed Capability-Based Access Control (DCapBAC) and Ciphertext-Policy Attribute-Based Encryption (CP-ABE) brings versatility and adaptability to diverse IoT scenarios.

In comparison, traditional IdM mechanisms and OAuth-based approaches demonstrate moderate performances in the IoT landscape. Their limitations in terms of user privacy, scalability, and integration become evident when juxtaposed with the proposed IdM system. Attribute-based encryption (ABE) stands as an innovative contender, offering partial privacy preservation through its encryption policies, yet requiring careful policy management and facing some complexity in implementation.

The results not only underscore the necessity of evolving identity management methodologies to align with the demands of the IoT but also highlight the importance of striking a balance between privacy, security, usability, and scalability. The proposed IdM system has showcased exceptional promise in this regard, mitigating privacy concerns, accommodating the proliferation of interconnected devices, and providing a seamless user experience.

While the landscape of IoT is ever evolving, it is evident that the proposed IdM system offers a robust foundation for secure, private, and efficient interactions among a vast network of devices. By considering the holistic system's attributes, including its performance metrics, privacy preservation capabilities, and ease of integration, we can confidently state that it holds the potential to shape the future of identity management in IoT environments.

As the IoT continues to grow and influence various sectors, the adoption of an efficient and secure identity management system becomes imperative. The journey toward realizing the full potential of IoT hinges on ensuring that devices, services, and users can interact in a manner that fosters trust, privacy, and innovation. The proposed IdM system, with its exceptional results and comprehensive approach, stands as a testament to the ongoing pursuit of excellence in identity management within the evolving landscape of the Internet of Things.

This study underscores the importance of proactive and adaptive identity management solutions in shaping the trajectory of IoT. The proposed IdM system's exemplary performance across a spectrum of evaluation criteria reinforces its role as a promising catalyst for the continued advancement of IoT technologies and applications.

REFERENCES

[1] N. Yousefnezhad, A. Malhi, T. Keyriläinen, and K. Främling, "A Comprehensive Security Architecture for Information Management throughout the Lifecycle of IoT Products," Sensors, vol. 23, no. 6, p. 3236, Mar. 2023, doi: 10.3390/s23063236.

[2] U. Tariq, I. Ahmed, A. K. Bashir, and K. Shaukat, "A Critical Cybersecurity Analysis and Future Research Directions for the Internet of Things: A Comprehensive Review," Sensors, vol. 23, no. 8, p. 4117, Apr. 2023, doi: 10.3390/s23084117.

[3] M. Amirthavalli, S. Chithra, and R. Yugha, "An Improved Pairing-Free Ciphertext Policy Framework for IoT," Computer Systems Science and Engineering, vol. 45, no. 3, pp. 3079–3095, 2023, doi: 10.32604/csse.2023.032486.

[4] T. Schüppstuhl, K. Tracht, and J. Fleischer, Eds., Annals of Scientific Society for Assembly, Handling, and Industrial Robotics 2022. Cham: Springer International Publishing, 2023. doi: 10.1007/978-3-031-10071-0.

[5] J. Bernal Bernabe, J. L. Hernandez-Ramos, and A. F. Skarmeta Gomez, "Holistic Privacy-Preserving Identity Management System for the Internet of Things," Mobile Information Systems, vol. 2017, pp. 1–20, 2017, doi: 10.1155/2017/6384186.

[6] T. Yousuf, R. Mahmoud, F. Aloul, and I. Zualkernan, "Internet of Things (IoT) Security: Current Status, Challenges and Countermeasures," IJISR, vol. 5, no. 4, pp. 608–616, Dec. 2015, doi: 10.20533/ijisr.2042.4639.2015.0070.

[7] R. Mahmoud, T. Yousuf, F. Aloul, and I. Zualkernan, "Internet of things (IoT) security: Current status, challenges and prospective measures," in 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST), London, United Kingdom: IEEE, Dec. 2015, pp. 336–341. doi: 10.1109/ICITST.2015.7412116.

[8] I. Ali, S. Sabir, and Z. Ullah, "Internet of Things Security, Device Authentication and Access Control: A Review," vol. 14, no. 8, 2016.

[9] M. A. Sahi et al., "Privacy Preservation in e-Healthcare Environments: State of the Art and Future Directions," IEEE Access, vol. 6, pp. 464–478, 2018, doi: 10.1109/ACCESS.2017.2767561.

[10] M. Kumar, M. Sethi, S. Rani, D. K. Sah, S. A. AlQahtani, and M. S. Al-Rakhami, "Secure Data Aggregation Based on End-to-End Homomorphic Encryption in IoT-Based Wireless Sensor Networks," Sensors, vol. 23, no. 13, p. 6181, Jul. 2023, doi: 10.3390/s23136181.

[11] J. Miguel-Alonso, "Securing IoT networks through SDN technologies," Computer Science and Mathematics, preprint, Jul. 2023. doi: 10.20944/preprints202307. 1781.v1.

[12] M. Abomhara and G. M. Koien, "Security and privacy in the Internet of Things: Current status and open issues," in 2014 International Conference on Privacy and Security in Mobile Systems (PRISMS), Aalborg, Denmark: IEEE, May 2014, pp. 1–8. doi: 10.1109/PRISMS.2014.6970594.

[13] R. S. M. Joshitta and L. Arockiam, "Security in IoT Environment: A Survey," Mechanical Engineering, no. 7, 2016.

[14] B. Kishiyama, J. Guerrero, and I. Alsmadi, "Security Policies Automation in Software Defined Networking," SSRN Journal, 2023, doi: 10.2139/ssrn.4384690.

[15] E. Becker, M. Gupta, and K. Aryal, "Using Machine Learning for Detection and Classification of Cyber Attacks in Edge IoT".

[16] Sharma, A.; Sharma, S.; Dave, M. Identity and Access management—A Comprehensive Study. In Proceedings of the 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), Greater Noida, India, 8–10 October 2015; pp. 1481–1485.

[17] Ravidas, S.; Lekidis, A.; Paci, F.; Zannone, N. Access Control in Internet-of-Things: A Survey. J. Netw. Comput. Appl. 2019, 144, 79–101.

[18] Qiu, J.; Tian, Z.; Du, C.; Zuo, Q.; Su, S.; Fang, B. A Survey on Access Control in the Age of Internet of Things. IEEE Internet Things J. 2020, 7, 4682–4696.

[19] Dramé-Maigné, S.; Laurent, M.; Castillo, L.; Ganem, H. Centralized, Distributed, and Everything in between: Reviewing Access Control Solutions for the IoT. ACM Comput. Surv. 2021, 54, 138.

[20] Alnefaie, S.; Alshehri, S.; Cherif, A. A survey on access control in IoT: Models, architectures, and research opportunities. Int. J. Secur. Netw. 2021, 16, 60–76.

[21] M. Mamdouh, A.I. Awad, A.A. Khalaf, H.F. Hamed Authentication and identity management of IoHT devices: achievements, challenges, and future directions Comput. Secur., 111 (2021), Article 102491.

[22] G.D. Putra, V. Dedeoglu, S.S. Kanhere, R. Jurdak Trust management in decentralized IoT access control system 2020 IEEE International C*onference on Blockchain and Cryptocurrency (ICBC)*, IEEE (2020, May), pp. 1-9.

[23] S. Joshi, S. Stalin, P.K. Shukla, P.K. Shukla, R. Bhatt, R.S. Bhadoria, B. Tiwari Unified authentication and access control for future mobile communication based lightweight IoT systems using blockchain Wireless Commun. Mobile Comput., 2021 (2021).

[24] A. Vieira, J.A. Nacif, M. Nogueira Survey on Identity and Access Management for Internet of Things (2020).

[25] Vincent C. et a. NIST Special Publication 800- 162. Guide to Attribute Based Access Control (ABAC) Definition and Considerations http://nvlpubs.nist.gov/nistpubs/specialpublications/ NIST.sp.800-162.

Speech Enhancement using Fully Convolutional UNET and Gated Convolutional Neural Network

Danish Baloch[1], Sidrah Abdullah[2], Asma Qaiser[3], Saad Ahmed[4], Faiza Nasim[5], Mehreen Kanwal[6]
Department of Computer Science, DHA Suffa University, Karachi, Pakistan[1]
Department of Computer Science and Information Technology,
NED University of Engineering & Technology, Karachi, Pakistan[2, 5]
Department of Computer Science, IQRA University, Karachi, Pakistan[3, 4]
MS Fast University, Department of Computer Science, Pakistan[6]

***Abstract*—Speech Enhancement aims to enhance audio intelligibility by reducing background noises that often degrade the quality and intelligibility of speech. This paper brings forward a deep learning approach for suppressing the background noise from the speaker's voice. Noise is a complex nonlinear function, so classical techniques such as Spectral Subtraction and Wiener filter approaches are not the best for non-stationary noise removal. The audio signal was processed in the raw audio waveform to incorporate an end-to-end speech enhancement approach. The proposed model's architecture is a 1-D Fully Convolutional Encoder-to-Decoder Gated Convolutional Neural Network (CNN). The model takes the simulated noisy signal and generates its clean representation. The proposed model is optimized on spectral and time domains. To minimize the error among time and spectral magnitudes, L1 loss is used. The model is generative, denoising English language speakers, and capable of denoising Urdu language speech when provided. In contrast, the model is trained exclusively on the English language. Experimental results show that it can generate a clean representation of a clean signal directly from a noisy signal when trained on samples of the Valentini dataset. On objective measures such as PESQ (Perceptual Evaluation of Speech Quality) and STOI (Short-Time Objective Intelligibility), the performance evaluation of the research outcome has been conducted. This system can be used with recorded videos and as a preprocessor for voice assistants like Alexa, and Siri, sending clear and clean instructions to the device.**

***Keywords*—Speech enhancement; speech denoising; deep neural network; raw waveform; fully convolutional neural network; gated linear unit**

I. Introduction

Speech Enhancement has been a topic of interest for five decades. Speech enhancement aims to improve speech quality (reducing background noise) by various algorithms [1]. The purpose of enhancement is to enhance the intelligibility of the speech signal degraded by the noise using audio signal processing techniques. The conventional methods used for noise reduction are Spectral subtraction and the Wiener filter [2] and [3]. Still, both approaches leave musical artifacts in synthesized speech [4], need multiple sources as noise profile information, and distort the desired output.

Deep Learning approaches can overcome the pitfalls of conventional approaches because these systems can learn to map between complex nonlinear functions [5]. In addition, they have the ability to produce desirable outputs that can be used to decrease the Word Error Rate (WER) of automatic speech recognition (ASR) systems [6], boost the performance of speech-to-text systems [7], and in general, increase the intelligibility of speech which can be beneficial for any system whose performance is dependent on the intelligibility of speech. In Deep Learning, the classical approach to suppress noise through the signal is mask-based signal denoising [8], in which DNN models produce a TF mask that filters out the noise and leaves the speech. Mask-based approaches are mostly done on magnitude spectrograms of audio [9], [10]; this creates a challenge of reconstructing the audio again to the time domain once it is filtered using the predicted spectrogram mask and reconstruction of audio is heavily dependent on the phase of noisy input audio.

Another investigated approach is a mapping-based approach where a representation of a complex nonlinear noisy signal is directly mapped onto a clean signal [11], [12] and [13]. Mapping-based approaches directly map noisy signals to their clean representations. Due to the fast variation of amplitudes in raw audio waveforms, mapping-based approaches are based on STFT (short-time Fourier Transform) of audio.

A. Proposed Approach

Our proposed approach is a mapping-based approach in raw audio waveform (time-domain). The loss function is optimized for time and STFT of audio. This approach eliminates the requirement of reconstruction of the audio from the spectrogram output into raw audible audio waveform as in [11] and [12], rather it generates the audible enhanced speech output directly. The magnitude spectrogram of audio is incorporated inside of the loss function rather than as input to the model as in [9] and [13], which gives us leverage to do speech enhancement on raw audio waveform directly. Given an audio, our system directly generates its clean representation without any additional post-processing on the output of the model. The proposed approach focuses on enhancing the speech and suppressing the noise in audio sampled at 22.05 KHz. To achieve this U-Net architecture is used. The choice of this architecture is due to the fact that it takes audio as raw waveform without any manual feature extraction and provides output also in the raw audio waveform, which can be converted to mp3 file and can be saved on disk directly. It consists of convolutional layers and a middle layer which is a

bottleneck. The middle layer (bottleneck layer) represents the data in encoded representation which is then decoded by the Decoder architecture connected to the bottleneck layer.

The objectives of this research are two-fold:

1) Discusses the common approach in deep learning, specifically mask-based signal denoising, where deep neural network (DNN) models generate a time-frequency (TF) mask to filter out noise from the audio signal.

2) Reviews mapping-based approaches that directly map complex nonlinear noisy signals onto clean representations. It distinguishes these approaches from mask-based methods and notes that mapping-based methods typically rely on the short-time Fourier transform (STFT) of audio due to the fast variation of amplitudes in raw audio waveforms.

3) Introduces a novel mapping-based approach specifically applied to raw audio waveforms in the time domain

B. Research Distribution

Section II of this paper covers related work in speech enhancement, followed by Section III, which discusses the dataset used for this research. Section III also consists of discussion on the U-Net Architecture. Section IV covers model training, and Section V, the last section discusses results and concludes the paper.

II. Related Work

Speech enhancement research has been ongoing for the last half-century. Earlier, classical linear noise filtering approaches were used for reducing noise. Two notable examples are spectral subtraction and wiener filter approaches [2] & [3]. The former needed multiple sources and works average with static noise and below the bar with non-stationary noises. The latter had its pitfalls, such as it required two sources; one of them is a mixed signal, and the other is the background sound signal. With the rise of deep learning, these pitfalls were eliminated as deep learning made it easier to notably reduce the noise in the noisy speech samples. This approach made no assumptions regarding the statistical attributes of the signals and used a wide variety of noise types to provide a variety of noisy speech samples for training [14]. Moreover, these systems can learn to map between complex nonlinear functions. They can produce desirable outputs that can be used to decrease the Word error rate (WER) of automatic speech recognition systems (ASR). In general, it can increase intelligibility, which can benefit any system whose performance depends on speech intelligibility.

In a notable work [15], a causal model was proposed based on auto-encoder architecture. They also proposed effective data augmentation techniques, frequency band masking, and reverberation. Their results suggest that the proposed system is comparable to the SOTA (State-Of-TheArt) model across all performance measures while working directly on the raw waveform. It also discovered that up-sampling the audio before feeding it into the encoder improves accuracy, and then they downsampled the outputs by the same amount.

Another innovative approach, presented in [16] proposed a new deep learning-based framework for real-time speech enhancement on dual-microphone mobiles for close-talk scenarios. They used a masking-based approach using a computationally efficient CRN (Convolutional Recurrent Neural Network), which was trained for intra-channels and inter-channels. Their experimental results showed that their proposed approach outmatched the DNN-based and other traditional methods.

Alternatively, authors in [17] used a hybrid approach using DSP techniques and deep learning for noise suppression. The deep recurrent neural network with four hidden layers was used. The resulting lower complexity made it practical to be used in video-conferencing systems. Their results showed a significant improvement in quality from deep learning, especially for non-stationary noise types.

The authors of [7] also used a hybrid approach consisting of noise estimation and speech-to-text block. This paper's focus was on spontaneous speech in the medical domain. As the medical terms used in the area are complex, and speech recognition systems often fail to recognize those words, the idea here was to propose an algorithm that resolves this issue. Non-linear spectral subtraction for noise reduction and the Hidden Markov Model (HMM) were incorporated for converting the speech to text to reduce the word error rate.

This paper [18] discussed the classical approaches for noise reduction by using filters. It also discusses stationary and non-stationary noise and its subtypes. This approach [19] combined a short-time Fourier transform (STFT) and a learned analysis and synthesis basis in a stacked-network method with less than one million parameters for real-time noise suppression. [20] an improved approach to their previous research was proposed, where the Deep Denoising Autoencoder (DAE) is trained on only clean speeches. In this paper, they trained DAE on pairs of noisy signals and clean output using a mapping-based approach stack AE approach where AE is stacked to form DAE to estimate the noise from the noisy signal. This paper [21] explores a greedy layer-wise pretraining strategy to train a DAE for speech restoration and then applies that restored speech for noisy robust speech recognition.

III. Methodology

A. Dataset

Datasets consist of pairs of corresponding audios sampled at 22.05 KHz stored in WAV format in a Linux environment. The dataset is created using two publicly available datasets. From one dataset noisy environment audio samples are obtained and from another dataset, clean human speech audio samples are obtained later these two samples of datasets are mixed together using simple arithmetic addition in order to create noisy simulated environments and their corresponding clean speech pairs.

Noise samples are from the DEMAND dataset to generate simulated noisy environments [22]. The DEMAND dataset is recorded with an array of sixteen microphones with an original sampling rate of 48kHz. It is publicly available in 48kHz or a downsampled version of 16kHz. In this paper, the

48khz version is downsampled to 22.05kHz utilizing the librosa module of Python and later used in the dataset. Three noise profiles (noise environments) are chosen from the DEMAND dataset namely DKITCHEN, PRESTO, and OMEETING. DKITCHEN includes recordings of kitchen noises while cooking. At the same time, PRESTO consists of a set of noise recordings taken from the university restaurant during lunchtime, and OMEETING, consisting of meeting room sounds during discussions from the microphone array. At first, all 16 channels of DKITCHEN, PRESTO, and OMEETING are mixed together respectively in order to create a single noise profile for each environment.

Next, the first eight channels of PRESTO and OMEETING are mixed together. This was done because it is observed that in the case of PRESTO and OMEETING with all sixteen channels added together, the noise profile was overruling the speech components in raw audio and also in spectrograms. Our proposed model takes 2.97sec windows of inputs, so eight sections of length equal to 2.97sec are used from each noise profile. Eight sections are used because for each speaker eight utterances are chosen. Later these noise sections are mixed with each speaker.

For clean speech representations, eight unique utterances of 47 notable speakers from the Valentini dataset are used [23]. Then mixed the noisy environment samples and the clean samples from the Valentini dataset, to create simulated noisy environment signals, and their correspondence clean speech representation. These pairs of audio are converted to pickle format and saved on the disk.

There are a total of five batches of data, each with 376 utterances of 47 unique speakers. Batch-1 represents the DKITCHEN (sixteen channels mixed) noise condition, Batch-2 represents the PRESTO (sixteen channels mixed) noise condition, Batch-3 represents the OMEETING (sixteen channels mixed) noise condition, Batch-4 represents PRESTO (eight channels mixed) noise condition and Batch-5 represents OMEETING (eight channels mixed) noise condition.

B. Notations and Problem Settings

Targeting speech enhancement in mono-aural signals, where $x \in R^T$ is the given signal composed of additive background noise as $n \in R^T$ and clean speech $y \in R^T$. so that $x = y + n$. The length T is of a fixed duration, which equals *65536* samples when audio is sampled at 22.05 KHz. Our main objective is to find a function *f* through the non-linear architecture of the neural network that reduces the enhancement function to $f(x) \approx y$.

In this problem set, the function *f* is the neural network architecture, producing the clean speech *y* at its output layer.

C. UNET Architecture

As presented in Fig. 1, the adopted neural network architecture is a one-dimensional UNET encoder-to-decoder architecture with skip connections [24] and gated linear units. The input shape of the model is equal to the number of 65536 samples when audio is sampled at 22.05 KHz; the output shape is also the same as this is an end-to-end approach.

Gated Linear Units [25] are incorporated in the encoding and decoding blocks of the model; there are no GLU in the bottleneck section of the model; convolution layers throughout the model are one-dimensional.

A detailed overview of the proposed Fully Convolutional Gated Encoder-to-Decoder architecture is shown in Fig. 2. The proposed model has three sections Encoder Section, Bottleneck section, and Decoder section, which is in the end is connected to the output layer producing clean speech output. Each section of the model and their connection with each other is discussed as follows.

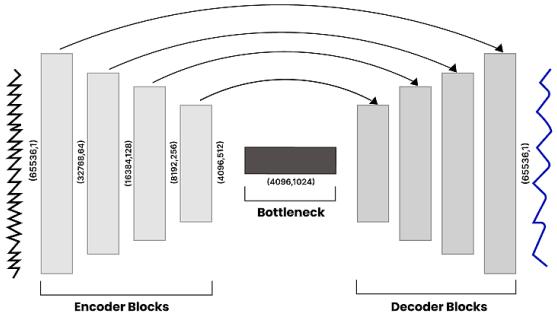


Fig. 1. Proposed fully convolutional gated encoder-to-decoder architecture with bottleneck for speech enhancement. Input and output shapes are the same.

1) Encoder section: The encoder section compresses the data of the input from a higher dimension to a lower dimension while reducing the noise from the data. The input of the encoder is the raw audio, and its output is a compressed data format that represents the input raw audio. Leaky ReLU is used as an activation function so that the model can learn to flow the gradient from most neurons. Most of the input data consists of negative values and the range of input data is [-1,1]. ReLU was also used as an activation function in order to obtain a sparse network but with ReLU, a dead ReLU problem was observed as input data is in the range of [-1, 1]. One-dimensional convolutional layers are used with batch normalization, Leaky Relu activation, and Gated linear Units are applied. The encoder section contains two layers of convolution than the max pooling layer.

2) Bottleneck section: The tanh activation function is used in the bottleneck section to apply a non-linear transformation on the signal at the most encoded layer. The input of the bottleneck section is the output of the encoder section, there are no Gated linear units applied at the bottleneck section, and it is also fully convolutive as the encoder section. In the bottleneck section, the one-dimensional convolutive layer is used with batch normalization, and the tanh activation function is applied. The bottleneck section contains two convolution layers only.

3) Decoder section: Transpose one-dimensional convolution is applied to convert back the original shape of the data. The decode section consists of one-dimensional convolution transpose along with gated linear units and

concatenation layers which serve as skip connections so that the model uses the features learned in the encoder section of the model.

4) Output layer: The output layer consists of a single channel focusing on mono-aural speech enhancement; Tanh activation is applied at the output layer, which provides the final denoised signal. In the output layer, a single one-dimensional convolutional layer is used which acts as a dense layer with a shape of (65536,1). The output layer is connected to the decoder layer. And it gives the same data shape as the input layer.

D. Objective

Mean Absolute Error (L1 Loss) is used as a loss function to minimize the error between the predicted signal $\boldsymbol{y'}$ and the clean signal $\boldsymbol{y}$. L1 loss is incorporated over the time domain of signals to reduce the loss over the sequential time domain and to minimize the loss over the spectral domain, L1 loss is used over on STFT (short-time Fourier transform) of the signals.

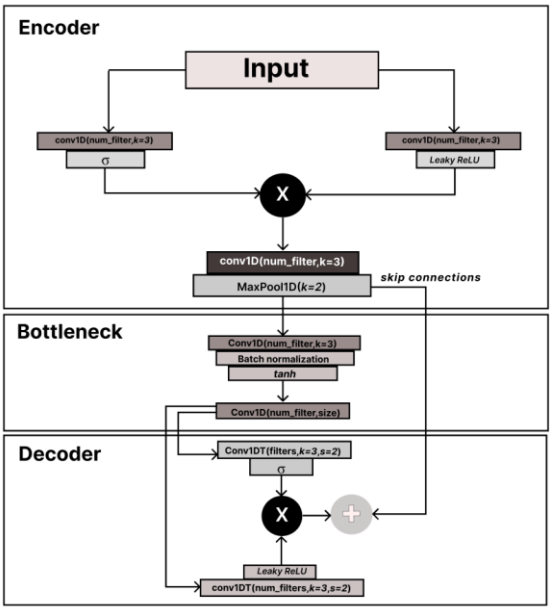


Fig. 2. Detailed overview of each separate block of the model and their connections with each other. The input and output shapes are the same.

Lastly, these L1 losses are added together to optimize the model in both the time domain (on amplitude vectors) and the spectral domain (on STFT).

A minimal epsilon value = $1e^{-10}$ is used to omit the undefined log error.

$$L_{stft}\,(y, y') = L_{mag}\,(y, y')$$

$$L_{mag}\,(y, y') \Rightarrow \frac{1}{T}|| \, log\,|STFT(y)| + \varepsilon - log|STFT(y')| + \varepsilon||_1$$

And,

$$L_{time}\,(y, y') = ||y - y'||_1$$

Overall, we wish to minimize the following:

$$L_1 \Rightarrow L_{mag}\,(y, y') + L_{time}\,(y, y')$$

IV. Training

Proposed model is trained on our custom dataset, 100 epochs for each of the three batches. The best model is saved for each of the three batches while monitoring the lowest L1 loss over the spectral domain, as discussed in Section 3.3.

Table I reports the training parameters, using the Adam optimizer, which has a learning rate of 1^{e-4}, with a momentum of β1=0.9 and a denominator momentum of β2=0.999.

TABLE I. Tuning Hyper-Parameters of Adam Optimizer Used for Training the UNET Model

Optimizer Tuning hyper-parameters	
optimizer	Adam
lr (learning rate)	1^{e-4}
β1 (momentum)	0.9
β2 (denominator)	0.999

The audio samples used for training are sampled at 22.05 KHz. The first 282 samples are used for the training set and the remaining 94 samples are used for the validation set. A batch size of 4 is used, dividing the data into 71 batches, each consisting of four unique utterances of the same speaker and the next batch containing the. The model optimizes the loss function and adjusts the weights. The next batch includes four different unique utterances of the previous speaker. After that, the next batch focuses on different speakers. The random order of data was not utilized. In each training session with varying noise conditions, the utterances' order remains consistent with their corresponding pairs in the noisy environment.

Gated linear units are used in order to control the information flow inside the model. It is observed that Gated liner units act as voice activity detection layers. Where clean speech is present, the neurons have greater weight and lower weight where noise is present. The output layer acts as a normalization layer for the denoised output.

V. Result

The performance of the model is calculated over cumulative loss of Mean Absolute Error of predicted and actual audio samples in both the time domain which is WAV MAE and also is spectral domain which is STFT MAE.

It is observed that the addition of STFT MAE in loss function improved the audible quality of speech in noisy environments and preserved the speech components of audio by making them sound less distorted.

Firstly, performance is evaluated over sixteen channels of respective noise environments that were present in the DEMAND dataset [22], this is reported in Table II. Then performance is evaluated over only eight channels of respective noise environments. It is observed that in the case of 16 channels of PRESTO and OMEETING audile quality of

audio was distorted, the reason behind this is that in the case of PRESTO and OMEETING noise environment babble noise is present and when it gets mixed with human speech it is hard for model to differentiate between noise and speech. We reported the performance of PRESTO and OMEETING with eight channels mixed together in Table III and a significant drop in Loss is observed while preserving the speech quality with less distortion.

TABLE II. PERFORMANCE OVER SIXTEEN CHANNEL MIX

Noise Environment All Sixteen Channel Mixed	Train			Test		
	LOSS	***WAV MAE***	***STFT MAE***	***LOSS***	***WAV MAE***	***STFT MAE***
DKITCHEN	0.755	0.011	0.744	0.777	0.012	**0.764**
PRESTO	0.734	0.017	0.716	0.811	0.018	**0.792**
OMEETING	0.537	0.010	0.527	0.573	0.010	**0.563**

This table represents the cumulative loss, the wav loss over signals, and the STFT loss of the signals. This metric represents the performance of the model where the lowest STFT MAE is observed in each noise condition. All sixteen channels of noise are added together in all of three noise profile cases.

TABLE III. PERFORMANCE OVER EIGHT CHANNEL MIX

Noise Environment First Eight Channel Mixed	Train			Test		
	LOSS	***WAV MAE***	***STFT MAE***	***LOSS***	***WAV MAE***	***STFT MAE***
PRESTO	0.665	0.012	0.653	0.744	0.013	**0.730**
OMEETING	0.494	0.008	0.486	0.510	0.009	**0.500**

This table represents the cumulative loss, the wav loss over signals, and the STFT loss of the signals. This metric represents the performance of the model where the lowest STFT MAE is observed in each noise condition. Eight channels of noise are added together in all of the two noise profile cases.

A significant drop in STFT Loss is observed in Table III. When the first eight channels are mixed, and the output audio quality is better than the previous setting in the case of PRESTO and OMEETING. Reducing the number of channels in DKITCHEN is not tested because the audible intelligibility of the clean speeches was satisfactory, this shows that the model is giving better results on the DKITCHEN environment despite higher STFT MAE as reported in Table II.

On Objective measures, intelligibility, and speech quality is measured as reported in Tables IV and V. The metrics used are PESQ [26] and STOI [27]. Baseline PESQ and STOI are calculated over denoised signals by using the noisereduce library of Python. Then it is compared with PESQ and STOI of denoised signals of the model.

TABLE IV. PERFORMANCE ON OBJECTIVE MEASURES USING PESQ AND STOI (SIXTEEN CHANNELS MIXED)

Noise Environment All Sixteen Channel Mixed	Test Set	
	PESQ wb	***STOI***
Baseline_DKITCHEN	1.16	0.83
DKITCHEN	**1.60**	**0.85**
Baseline_PRESTO	1.30	0.66
PRESTO	**1.31**	**0.73**
Baseline_OMEETING	1.30	0.82
OMEETING	**2.46**	**0.88**

Objective measures of enhanced speech on PESQ and STOI. Sixteen channels of noise are added together in all three noise profile cases.

TABLE V. PERFORMANCE ON OBJECTIVE MEASURES USING PESQ AND STOI (EIGHT CHANNELS MIXED)

Noise Environment First Eight Channel Mixed	Test Set	
	PESQ wb	***STOI***
Baseline_PRESTO	1.24	0.74
PRESTO	**1.38**	**0.81**
Baseline_OMEETING	1.35	0.85
OMEETING	**3.24**	**0.90**

Objective measures of enhanced speech on PESQ and STOI. Eight channels of noise are added together in all of the two noise profile cases.

In Tables IV and V, our performance of the model in objective measures using PESQ and STOI is reported. Improvement in the audible quality of denoised speech is observed when the human voice is more prominent (audible) in signal than in the noise environment. There is a significant improvement in both objective measures over the baseline model, baseline measurement is carried with PESQ and STOI on our test set using spectral gating.

VI. DISCUSSION

In summary, this research addresses the challenges in speech enhancement by examining a mapping-based approach on raw audio waveforms using the U-Net architecture. The study identifies limitations in conventional methods like Spectral subtraction and the Wiener filter, prompting an exploration of deep learning solutions.

The proposed approach optimizes the loss function for both time and short-time Fourier transform (STFT) of audio, enabling the direct generation of clean audio representations without additional post-processing. The research systematically evaluates mask-based and mapping-based deep learning approaches, revealing the effectiveness of the latter in various noise environments through a comprehensive metric.

Objective measures, including PESQ and STOI, indicate notable improvements in audible quality and intelligibility of denoised speech compared to baseline models. The research findings have practical implications for applications such as automatic speech recognition and speech-to-text systems. The mapping-based approach on raw audio waveform emerges as a viable strategy for addressing the inherent challenges in speech enhancement, offering tangible advancements in audio quality assessment.

VII. CONCLUSION

The proposed approach can be scaled by incorporating all the noise profiles into one single dataset and training a model in a single pass over the entire dataset, the hypothesis is that the model may generalize better to each noise condition, and a reduction in Loss is observed. It is observed that the intelligibility of output speech samples is improved when STFT is included with the Loss Function, as discussed in the objective of this paper. In the future, the proposed approach can be scaled with the use of multiple-resolution STFT loss as used in [28, 29].

REFERENCES

[1] Yuliani, A. R., Amri, M. F., Suryawati, E., Ramdan, A., & Pardede, H. F. (2021). Speech enhancement using deep learning methods: A review. Jurnal Elektronika dan Telekomunikasi, 21(1), 19-26.

[2] Boll, Steven. "Suppression of acoustic noise in speech using spectral subtraction." IEEE Transactions on acoustics, speech, and signal processing 27, no. 2 (1979): 113-120.

[3] Lim, J. S. "Enhancement and bandwidth compression of noisy speech." Proc. IEEE 67, no. 12 (1962): 1689-1697.

[4] Scalart, Pascal. "Speech enhancement based on a priori signal to noise estimation." In 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 2, pp. 629-632. IEEE, 1996.

[5] Xu, Yong, Jun Du, Li-Rong Dai, and Chin-Hui Lee. "An experimental study on speech enhancement based on deep neural networks." IEEE Signal processing letters 21, no. 1 (2013): 65-68.

[6] Subramanian, Aswin Shanmugam, Xiaofei Wang, Murali Karthick Baskar, Shinji Watanabe, Toru Taniguchi, Dung Tran, and Yuya Fujita. "Speech enhancement using end-to-end speech recognition objectives." In 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 234-238. IEEE, 2019.

[7] Gnanamanickam, J., Natarajan, Y., & KR, S. P. (2021). A hybrid speech enhancement algorithm for voice assistance application. Sensors, 21(21), 7025.

[8] Soni, M. H., Shah, N., & Patil, H. A. (2018, April). Time-frequency masking-based speech enhancement using generative adversarial network. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5039-5043). IEEE.

[9] Takeuchi, Daiki, Kohei Yatabe, Yuma Koizumi, Yasuhiro Oikawa, and Noboru Harada. "Effect of spectrogram resolution on deep-neural-network-based speech enhancement." Acoustical Science and Technology 41, no. 5 (2020): 769-775.

[10] Liu, Kuan-Yi, Syu-Siang Wang, Yu Tsao, and Jeih-weih Hung. "Speech enhancement based on the integration of fully convolutional network, temporal lowpass filtering and spectrogram masking." In Proceedings of the 31st Conference on Computational Linguistics and Speech Processing (ROCLING 2019), pp. 226-240. 2019.

[11] Park, S. R., & Lee, J. (2016). A fully convolutional neural network for speech enhancement. arXiv preprint arXiv:1609.07132.

[12] Liu, Chang-Le, Sze-Wei Fu, You-Jin Li, Jen-Wei Huang, Hsin-Min Wang, and Yu Tsao. "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 1888-1900.

[13] Tan, Ke, and DeLiang Wang. "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6865-6869. IEEE, 2019.

[14] Wang, Y., Han, J., Zhang, T., & Qing, D. (2021). Speech enhancement from fused features based on deep neural network and gated recurrent unit network. EURASIP Journal on Advances in Signal Processing, 2021, 1-19.

[15] Defossez, A., Synnaeve, G., & Adi, Y. (2020). Real time speech enhancement in the waveform domain. arXiv preprint arXiv:2006.12847.

[16] Tan, K., Zhang, X., & Wang, D. (2019, May). Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5751-5755). IEEE.

[17] Valin, J. M. (2018, August). A hybrid DSP/deep learning approach to real-time full-band speech enhancement. In 2018 IEEE 20th international workshop on multimedia signal processing (MMSP) (pp. 1-5). IEEE.

[18] Kaur, J., Baghla, S., & Kumar, S. (2015). A review: Audio noise reduction and various techniques. Int. J. of Advances in Sci. Engn. and Techn., 3(3), 132-135.

[19] Westhausen, N. L., & Meyer, B. T. (2020). Dual-signal transformation lstm network for real-time noise suppression. arXiv preprint arXiv:2005.07551.

[20] Lu, X., Tsao, Y., Matsuda, S., & Hori, C. (2013, August). Speech enhancement based on deep denoising autoencoder. In Interspeech (Vol. 2013, pp. 436-440).

[21] Lu, X., Matsuda, S., Hori, C., & Kashioka, H. (2012). Speech restoration based on deep learning autoencoder with layer-wised pretraining. In Thirteenth Annual Conference of the International Speech Communication Association.

[22] J. Thiemann, N. Ito and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," 9 June 2013. [Online]. https://zenodo.org/record/1227121#.Y_RhTHZBzIU [Accessed 30 Jan 2022]

[23] Valentini-Botinhao, C. (2017). Noisy speech database for training speech enhancement algorithms and tts models. University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR).

[24] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234-241. Springer International Publishing, 2015.

[25] Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2017, July). Language modeling with gated convolutional networks. In International conference on machine learning (pp. 933-941). PMLR.

[26] Rix, Antony W., John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs." In 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221), vol. 2, pp. 749-752. IEEE, 2001.

[27] Taal, Cees H., Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. "An algorithm for intelligibility prediction of time–frequency weighted noisy speech." IEEE Transactions on Audio, Speech, and Language Processing 19, no. 7 (2011): 2125-2136.

[28] Yamamoto, Ryuichi, Eunwoo Song, and Jae-Min Kim. "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6199-6203. IEEE, 2020.

[29] Yamamoto, Ryuichi, Eunwoo Song, and Jae-Min Kim. "Probability density distillation with generative adversarial networks for high-quality parallel waveform generation." arXiv preprint arXiv:1904.04472 (2019).

Hotspot Identification Through Pick-Up and Drop-Off Analysis of Ride-Hailing Transport Service

Ragil Saputra[1], Suprapto[2*], Agus Sihabudin[3]
Department of Computer Science, Universitas Diponegoro, Jl. Prof. Soedarto SH Tembalang Semarang Indonesia[1]
Department of Computer Science and Electronics, Universitas Gadjah Mada, Bulaksumur Yogyakarta Indonesia[1, 2, 3]

***Abstract*—It is important to extract hotspots in urban traffic networks to improve driver route efficiency. This research aims to identify hotspot pick-up and drop-off (PUDO) areas in ride-hailing transportation services using a clustering approach. However, there are challenges in applying clustering algorithms to trajectory data in the coordinates of the Global Positioning System (GPS). So this research proposes modifications to the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm by considering the radius from the center of the cluster to determine the presence of amenities around the cluster. We used a dataset containing 55,988 trip trajectories of Grab drivers over a two-week period in Jakarta. A preliminary statistical analysis was carried out to understand the distribution of trips. Next, we identify the PUDO point of each trip for use in the clustering analysis. The research explores the various parameters and settings of the clustering method and their impact on the results. The study found that the results obtained from the clustering method are sensitive to parameter selection, including epsilon radius and minimum number of points needed to form a cluster. The optimal cluster with the best parameters (eps: 0.25, minpts: 100) in the pick-up (PU) location analysis produced 17 clusters with the silhouette coefficient of 0.752, while in the drop-off (DO) location there are 18 clusters with a silhouette coefficient of 0.694. Overall, the research highlights the potential of the clustering analysis method for ride-hailing transportation.**

***Keywords*—Hotspot identification; ride-hailing; transportation; PUDO location; clustering analysis**

I. Introduction

A. Background

The rise of application-based transportation services, such as ride-hailing has revolutionized the way people move around in modern cities [1]. In the context of transportation services, "pick-up" often denotes the time when a vehicle (such as a taxi, or ride-hailing service) arrives at a designated location to collect passengers or packages. While “drop off” can refer to the time when a vehicle arrives at the destination location, and passengers or packages are left at that location.

The exponential growth in the use of these services has created enormous opportunities to analyze and understand urban mobility patterns [2]. In the study of Zhang et al [3], mobility-based data analysis has become the main focus in uncovering complex urban movement patterns. There have been many studies that use taxi data to examine urban mobility, such as Veloso et al [4] use of taxi trajectory data from Lisbon to discuss the spatiotemporal variation of taxi services, correlations between pick-up and drop-off (PUDO) sites, and driver behavior, and Liu et al [5] use of taxi trajectory data from Shanghai to explore human movement patterns. The taxi traces were also mined for characteristics and behaviors of the vehicular network, including anomalous and driver behavior, dynamics of mobility patterns, interactions between vehicles, and the relationship between gender and mobility [6]. Additionally, Keler et al [7] examined where automobile routes connect at specific rush hours in urban locations.

By identifying origin and destination (OD) flow clusters in urban travel data, it is possible to determine prospective routes for public transportation service settings [8]. In order to locate the taxi OD hotspot, the available OD pairs from empirical mobility traces are first grouped [9]. A deep understanding of hotspots, namely areas with heavy travel activity, has great potential for shaping more efficient transportation planning and better traffic management [10]. Although there have been previous studies exploring hotspot analysis in urban mobility, Dutta et al [11], use of more sophisticated and comprehensive clustering methods is still an important area of exploration.

Exploring urban mobility patterns has become a main focus in recent years, with previous research analyzing trajectory data [12]. Therefore, in this context, it is essential to define several key terms and assumptions that will form the basis of this study before presenting the main theorem. We define a hotspot as a location with a high concentration of PUDO points are adapted from [13]. We assume that the dataset used in this study is representative of the overall ride-hailing transportation system in the study area. Additionally, we assume that variables like the time of day, day of the week, and the location of well-known destinations have an impact on drivers' PUDO patterns.

B. Motivation

The majority of previous research on ride-hailing has been on the routing method, with little emphasis paid to optimizing PUDO locations that are sensitive to the spatial and temporal need distribution [14]. When taking a shared trip with multiple riders, the PUDO optimization is vital to prevent pointless detours. Whereas in the conventional system, the vehicle is frequently obliged to pick up passengers at certain locations [15]. Detours are made in order to pick up additional riders frequently result in longer travel times and higher costs. The PUDO position, where all potential PUDO points must therefore be optimized urgently.

Shen et al [2], suggests a cluster analysis approach for identifying various metropolitan online ride-hailing operation trends. Based on the suggested intensity and stability indicators

of ride-hailing vehicle operational characteristics, k-means++ clustering technique is applied. The results show that there are three distinct operating patterns for online ride-hailing services. In the study of Zhang et al [3], the goal is to identify the distribution of areas with high travel demand as well as the relationship between travel demand and Point of Interest (POIs). Tang [16], utilizes taxi GPS trajectory data to analyze urban human activity and mobility in Harbin city. The researchers employ the DBSCAN algorithm for PUDO location clustering and develop four spatial interaction models to understand pick-up location searching behavior.

However, with the rise of ride-hailing services, there is a need to develop new methods for analyzing PUDO patterns of drivers. Gunawan and Susilawati [17], addresses limitations in current ride-hailing PUDO location selection practices, which often prioritize spatial distribution and company interests over passenger needs. Research using neural networks was carried out by [18], seeks to examine the integration of clustering models and deep learning techniques. The model, which can concurrently capture the spatial and temporal fluctuations of taxi hotspots, was proposed for taxi hotspot prediction.

There have been several studies interested in discussing the applications for the clustering analysis method in transportation systems. Zhang et al [19], used DBSCAN to cluster PUDO data from a ride-hailing service in China. A pick-up points recommendation model (PPRM) is introduced, utilizing DBSCAN to cluster historical orders. This clustering enables finding contextually relevant candidate PU points. The other research by Wang and Ren [20], introduces a two-level divide approach and enhances the K-means++ algorithm to refine the clustering of taxi passenger hot spots based on GPS location data. The method is validated using a week of New York City's green taxi data, demonstrating superior accuracy and comparable time efficiency when compared to traditional K-means and DBSCAN methods.

Based on previous research, clustering analysis method has become an increasingly popular approach in the field of data mining and machine learning. Rafiq and McNally [21], use clustering these data points, ride-hailing companies can gain insights into traffic patterns and usage trends of their customers, which can help them optimize their operations and improve their overall service quality.

To build on prior research that examined trajectory data to investigate urban mobility patterns, this study employs a clustering approach to identify hotspot PUDO areas for ride-hailing transportation services. The DBSCAN algorithm was chosen due to its ability to identify clusters of varying shapes and sizes. However, it has limitations in handling spatial datasets, therefore, this study proposes modifying the DBSCAN algorithm to take into account the radius from the cluster's center to determine the presence of facilities in the cluster's vicinity. Our method identifies concentrations of PUDO locations that can be used to determine potential hotspots. While this study does not provide a direct comparison with previous methods, the proposed method is superior to previous approaches because it considers the radius from the center of the cluster to determine the presence of amenities around the cluster, uses a large dataset of ride-hailing trajectories.

The paper is structured as follows. Section II describes data processing, and proposed method. The findings and results gleaned from the methodology are discussed in Section III. Section IV concludes by summarizing the work and offering recommendations for the future.

II. METHODOLOGY

A. Data Collection

The dataset is derived from Grab's food delivery and logistics in Jakarta, and it includes 4,000 daily trajectories collected from 2019-04-08 to 2019-04-21 (inclusive, UTC/ Universal Time Coordinated). The trajectories were gathered from drivers' phones while they were on the road. The total number of GPS pings in the collection is 61.549.964; each GPS ping has values for its trajectory ID, latitude, longitude, timestamp, accuracy level, bearing, and speed [22]. The raw data sample is shown in Table I.

The names of the fields in Table I are covered below.

1) Trajectory ID: A number used to identify different GPS mobility trajectories.

2) Latitude: A GPS location's latitude coordinate.

3) Longitude: A GPS location's longitude coordinate.

4) Timestamp: Time the GPS locations were recorded is shown by the timestamp. The UTC standard is used in the format. One second is the quickest sample interval for each GPS point.

5) Accuracy level: Shows the radius of the circle that, with a certain probability, contains the real location.

6) Bearing: The degrees relative to true north.

7) Speed: The immediate speed is expressed in meters per second.

TABLE I. TEST RAW DATA

Trajectory ID	Latitude	Longitude	Timestamp	Accuracy Level	Bearing	Speed
4820	-6.591032	106.834468	09/04/19 10.33	4	194	20.42
46324	-6.247526	106.977663	10/04/19 05.22	10.72	101	4.22
15007	-6.267404	107.036016	19/04/19 02.57	3.149	108	1.99
4239	-6.293342	106.820029	10/04/19 01.31	10	288	1.01

B. Data Preprocessing

The next step is to pre-process the data to create trajectory, the data includes the point of trajectory of captured trips. The trips were divided based on the trajectory ID, and the minimum and maximum trajectory times recorded for each trip were used to determine the PUDO information. The result of dataset contains in total 55,988 trajectories with the distribution of number of trips is show in Fig. 1.

The daily distribution of trips is depicted in Fig. 1. We simply display the relative frequency of travel requests discussed in this article in order to maintain data confidentiality. As seen in Fig. 1(a) pick-up (PU) time, the temporal travel demand is typically distributed over the normal working day. Between 6 P.M and 12 noon, it is relatively high, but after that, the tendency begins to decline until 9 P.M. The trend then went upward once more until 10 P.M. The same trend also occurs in drop-off (DO) time, presented in Fig. 2(b).

C. Proposed Method

The main analysis involves applying the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method to the preprocessed data to identify clusters of PUDO locations. DBSCAN was introduced by Ester et al [23], which is a clustering algorithm commonly used in data analysis to identify clusters of data points based on their spatial density.

It's particularly useful when dealing with data where clusters might have irregular shapes and varying densities. The algorithm categorizes data points into three main types [24]: core points, border points, and noise points. A core point is a data point that has at least a specified number of neighbor points (*minpts*) within a certain distance (ε). These points are at the heart of a cluster. A border point is a point that is within the ε distance of a core point but doesn't have enough neighbors to be considered a core point itself. And finally, noise points are any points that are neither core nor boundary points.

The following steps make up the algorithm [25]:

1) Identify the core points or points visited by more than minpts neighbors by locating all neighbor points within ε.

2) Make a new cluster for each core point if it hasn't already been done so.

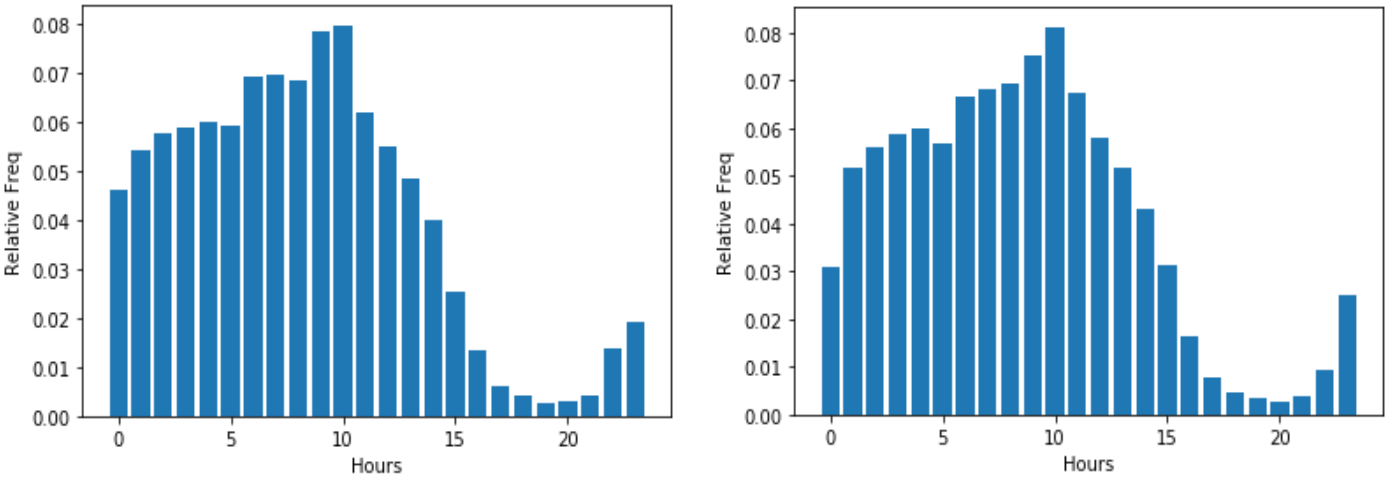
3) Find all points connected to it by density recursively, then group them with the core point in the same cluster.

4) Points a and b are said to be density connected if point c has a significant number of points in its neighbors and both of those points are close to the ε. Chains are used in this process. Inferring that b is a neighbor of a, if b is a neighbor of c, c is a neighbor of d, and d is a neighbor of e, then b is a neighbor of a.

5) Iteratively go over the remaining unexplored points in the dataset. All points that do not form a cluster are considered noise.

In this paper, we modified the DBSCAN algorithm from Kambe and Pe [25] to better suit the ride-hailing context by creating a function called DBSCAN_FIT that takes three parameters: x (dataset), ε (epsilon, the maximum distance between two points to be considered in the same cluster), and *minpts* (the minimum number of samples in a cluster). This function utilizes the DBSCAN algorithm to cluster the data and produces visualizations of the clusters, including those labeled as noise (points not belonging to any cluster). Additionally, the function calculates centroids for each cluster and assesses the amenity around each centroid. The pseudocode for the DBSCAN_FIT is shown in Algorithm 1.

To obtain the amenity information with function `get_amenity`, we used OpenStreetMap (OSM) data to identify the facilities or points of interest around each centroid. Specifically, we used the OSM API to query the database for the amenities within a certain radius of each centroid. We then used the OSM tags to extract the names of the amenities and assigned them as labels to the corresponding clusters. At a given average latitude on Earth, we use the number 111,320 as a conversion factor to translate differences in degrees of latitude or longitude into distances in kilometers.



(a) (b)

Fig. 1. Distribution average of (a) pick up time and (b) drop off time.

Algorithm 1. DBSCAN_FIT

Input : x (dataset), eps, minpts
Output : labels, centroid, amenity

```
Funnction DBSCAN_FIT(x, eps, minpts){
        dbs = DBSCAN(eps,minpts)
        dbs.fit(X)
        labels = dbs.labels_
        label_unique = pd.Series(labels).value_counts()
        if label_unique.shape[0]>2 {
                cent = []
        amen = []
        for i in range(max(labels)+1){
        centroid = (X[labels==i].mean().values/111.320) + np.array([min_lat,min_lon])
        centroid += [centroid]
        amenity = get_amenity(centroid[0], centroid[1], eps*4*1000)
        amenity += [amenity]
                }
        }
        else
                return 0
return labels, centroid, amenity
}
```

D. Cluster Performance Measure

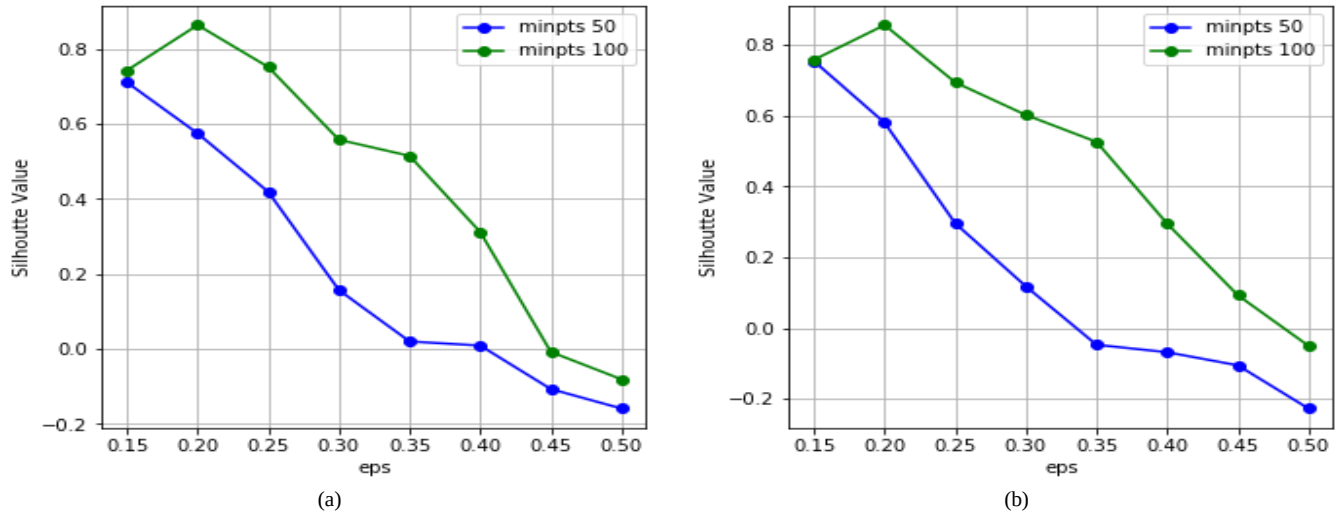
To evaluate the quality and coherence of clusters obtained through clustering algorithms, the Silhouette Coefficient (SC) was used. It quantifies the cohesion and separation between clusters. Ranging from -1 to 1, a higher SC indicates well-defined clusters where data points are closer to their own cluster members than to others. A score close to 0 suggests data points on cluster boundaries, while negative scores indicate potential misassignments [26].

A higher SC would indicate well-defined clusters of PUDO points, highlighting their coherence and separation from other clusters. This metric becomes essential in assessing the accuracy of clustering results and validating the effectiveness of algorithms. The silhouette score guides the determination of how accurately identified hotspots represent distinct patterns, thereby enhancing the credibility of the hotspot identification approach in ride-hailing transport services.

III. Result and Discussion

A. Experimental Result

Our research focuses on the complex interaction of epsilon (ε) values in the DBSCAN_FIT algorithm in order to do a thorough study of clustering results. To achieve this goal, we conducted a preliminary analysis of the data and found that *minpts* values below 50 did not produce enough clusters to be useful for hotspot identification, while values above 100 resulted in too many clusters, making it difficult to identify meaningful patterns. Therefore, we chose to focus on *minpts* values of 50 and 100 in our analysis and visualization, as these values produced the most meaningful results for hotspot identification. In Fig. 2, the outcomes of this investigation are graphically summarized. Fig. 2(a) shows the silhouette value for pick up and Fig. 2(b) shows the silhouette value for drop-off.



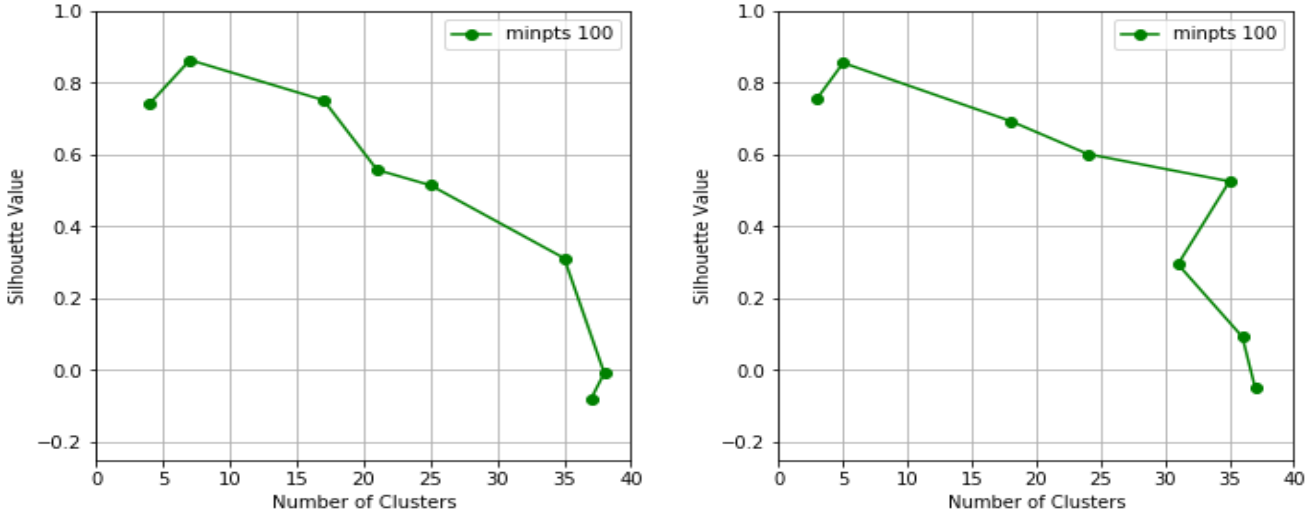
(a) (b)

Fig. 2. Variation epsilon (ε) value for the (a) pick up and (b) drop off.

The graph on Fig. 2, the graph illustrates the variation in the silhouette coefficient with changing values of ε. Two curves are depicted, one for a minimum number of points (*minpts*) set at 50, and the other for *minpts* set at 100. The results indicate that, for *minpts* 50, the silhouette coefficient (SC) tends to decrease as ε increases, suggesting a negative impact on clustering quality. In contrast, for *minpts* 100, the SC exhibits fluctuations, with some ε values resulting in higher coefficients. From the graph, it can be inferred that for *minpts* 100, the SC is more stable and potentially yields better clustering results compared to *minpts* 50 within a specific range of ε values.

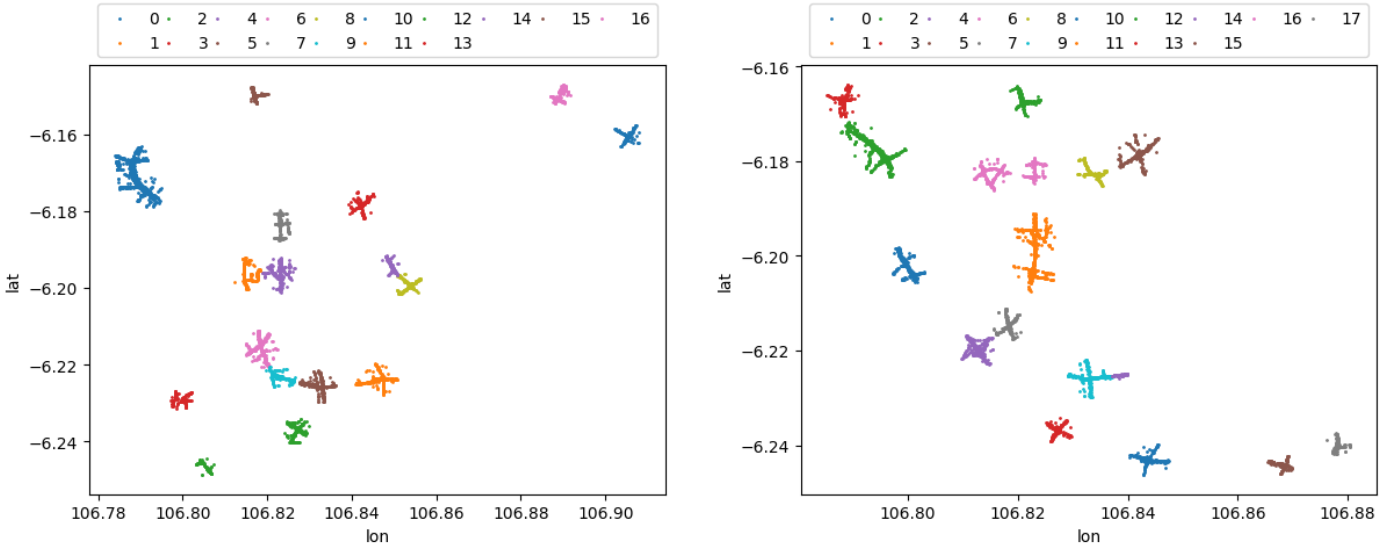
Furthermore, it is evident that the *minpts* value plays a pivotal role. The *minpts* 100 configuration consistently outperforms *minpts* 50, yielding higher SC values. So in this study *minpts* 100 is used as a reference to find the optimal number of clusters. Compare the value of the SC with the number of clusters formed presented in the Fig. 3.

The number of clusters formed ranging from 3 to 37 presented by Fig. 3. In the context of hotspot identification, a higher number of clusters can provide a more detailed picture of variations in the distribution of hotspots. However, on the other hand, the SC value is also important because it determines the quality and coherence between clusters. So the optimal number of clusters was chosen as 17 clusters for the PU location and 18 clusters for the DO location. So that the map distribution of PUDO locations is illustrated in Fig. 4.



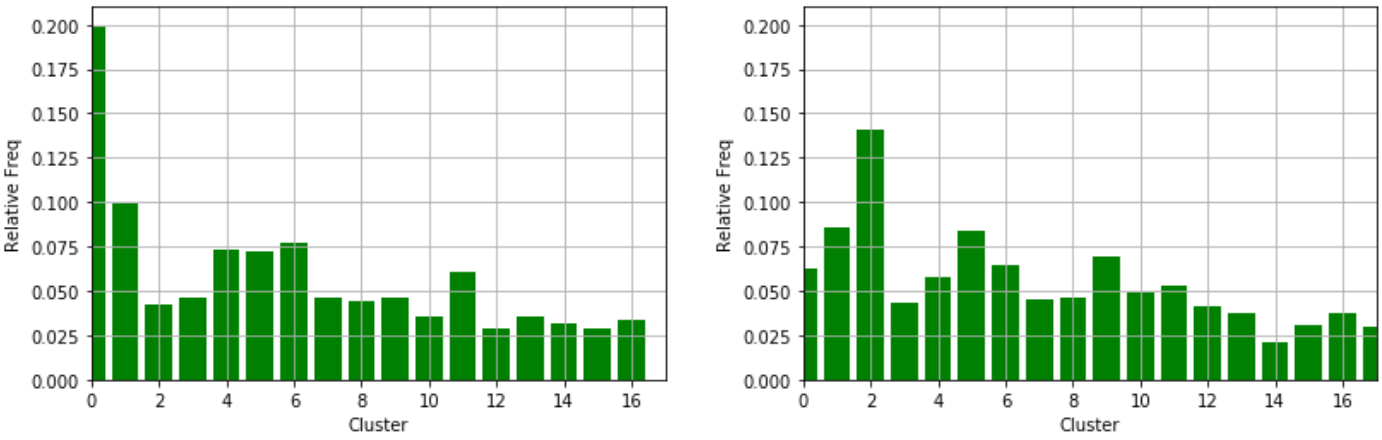
(a) (b)

Fig. 3. Silhouette coefficient and number of clusters of (a) pick up and (b) drop off.



(a) (b)

Fig. 4. The map distribution of (a) pick up location and (b) drop off location.



(a) (b)

Fig. 5. Number of trips (relative frequency) in cluster (a) pick up and (b) drop off.

The distribution of PUDO locations inside each cluster is shown in the Fig. 5. The cluster number is represented on the x-axis, while the number of trips inside each cluster is represented on the y-axis, in relative frequency. This information is crucial in our effort to locate hotspots for application-based transportation services. These findings are directly related to our study's primary objective, which was to develop a special technique for identifying hotspots by closely analysing PUDO patterns. We have learned a lot from the clustering analysis. With the parameters $\boldsymbol{\varepsilon}$: 0.25 and *minpts*: 100, the ideal cluster structure specifically produced 17 clusters for the PU localization analysis, showing an outstanding SC of 0.752. On the other hand, the DO location analysis revealed 18 clusters with a silhouette coefficient of 0.694. These clusters have varied levels of activity, as seen by the bar chart in Fig. 5, indicating the existence of distinct hotspots throughout the transportation network.

Based on Fig. 4 and Fig. 5 show that the highest hotspots for both pick-up and drop-off sites. It follows that these places represent prospective areas that could be used to enhance driver trips. Using the methods we propose, most potential hotspots are presented in Table II along with their amenity labels.

B. Discussion

To assess the performance of clustering, we conducted a comparative analysis with two distinct inpuvts values: 50 and 100, while exploring the influence of the $\boldsymbol{\varepsilon}$ parameter across a range from 0.15 to 0.5. The Silhouette Coefficient (SC) served as our evaluation metric [26]. Our findings are consistent with previous studies that have examined the impact of the $\boldsymbol{\varepsilon}$ parameter on DBSCAN clustering [23] [24]. For example, [23] found that larger values of $\boldsymbol{\varepsilon}$ can lead to the formation of overly large clusters, which can reduce the effectiveness of the clustering algorithm. Similarly, [24] found that larger values of $\boldsymbol{\varepsilon}$ can lead to a decrease in the quality of the clustering results. Our study builds on these findings by examining the impact of both the $\boldsymbol{\varepsilon}$ parameter and the *minpts* value on the performance of clustering in the context of ride-hailing services.

One important aspect of the study is the influence of algorithm parameters on the clustering results. The study found that the results obtained from the clustering method are sensitive to parameter selection, including epsilon radius and minimum number of points needed to form a cluster. The study explored the influence of the $\boldsymbol{\varepsilon}$ parameter across a range from 0.15 to 0.5 and found that for *minpts* 50, the SC tends to decrease as $\boldsymbol{\varepsilon}$ increases, suggesting a negative impact on clustering quality. In contrast, for *minpts* 100, the SC exhibits fluctuations, with some eps values resulting in higher coefficients.

Although our study does not explicitly mention the most surprising results, we found that our approach of analyzing pick-up and drop-off (PUDO) patterns was effective in identifying hotspots in ride-hailing transport services. Our analysis revealed commuting patterns of users and different hotspots in the transportation network. In this study, several limitations were identified that require careful consideration. Firstly, the dataset utilized in this research was limited to a single ride-hailing service, which may limit the generalizability of the analysis results to other services. Secondly, this study only accounted for factors such as time and location in the PUDO analysis, while other factors such as weather or special events in certain areas may influence PUDO patterns and were not considered in this study.

TABLE II. THE POTENTIAL HOTSPOT

Cluster	Type	Center Latitude	Center Longitude	Amenity
0	PU	-6.2210668	106.824997	restaurant: Planet Hollywood Jakarta / cinema: Setiabudi 21 / fuel: SPBU / restaurant: Loewy / restaurant: Bakso Solo
2	DO	-6.1999132	106.824234	nightclub; restaurant: DanceSignal Spot - Immigrant / parking_entrance: restaurant: Lanna Thai / cafe: nan / restaurant: Skye

IV. CONCLUSIONS

Our study reveals that DBSCAN_FIT clustering with ε: 0.25 *minpts*: 100 and yields 17 clusters for PU locations and 18 clusters for DO locations, demonstrating their potential as

hotspots in ride-hailing services. However, the limitation lies in the trade-off between cluster quantity and quality. A more comprehensive understanding of ride-hailing hotspots is achieved, emphasizing the need for a balance between cluster granularity and silhouette coefficients. Other researchers can use our method of analyzing hotspots through PUDO patterns to find hotspots in other transportation networks. This approach can help improve efficiency and help drivers optimize the routing of transportation services.

Future studies should consider utilizing datasets from multiple ride-hailing services to enhance the generalizability of the analysis results. And, additional factors such as weather or special events in certain areas should be taken into account in the PUDO analysis to provide a more comprehensive understanding of the factors that influence PUDO patterns.

ACKNOWLEDGMENT

This work was supported by Directorate General of Higher Education, Research, and Technology, the Ministry of Education, Culture, Research, and Technology Indonesia under Grant No. 1911/UN1/DITLIT/Dit-Lit/PT.01.03/2022.

REFERENCES

[1] X. Shen, L. Wang, Q. Pei, Y. Liu, and M. Li, "Location Privacy-Preserving in Online Taxi-Hailing Services," Peer-to-Peer Netw. Appl., vol. 14, no. 2021, pp. 69–81, 2021.

[2] Z. Xiong, Jian Li, and H. Wu, "Understanding Operation Patterns of Urban Online Ride-Hailing Services: A Case Study of Xiamen," Transp. Policy, vol. 101, no. August 2020, pp. 100–118, 2021, doi: 10.1016/j.tranpol.2020.12.008.

[3] B. Zhang, S. Chen, Y. Ma, T. Li, and K. Tang, "Analysis on Spatiotemporal Urban Mobility Based on Online Car-Hailing Data," J. Transp. Geogr., vol. 82, no. February 2019, p. 102568, 2019, doi: 10.1016/j.jtrangeo.2019.102568.

[4] M. Veloso, S. Phithakkitnukoon, and C. Bento, "Sensing Urban Mobility with Taxi Flow," in Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, 2011, pp. 41–44, doi: 10.1145/2063212.2063215.

[5] Y. Liu, C. Kang, S. Gao, Y. Xiao, and Y. Tian, "Understanding Intra-Urban Trip Patterns From Taxi Trajectory Data," J. Geogr. Syst., vol. 14, no. 4, pp. 463–483, 2012, doi: 10.1007/s10109-012-0166-z.

[6] F. A. Silva, C. Celes, and A. A. F. Loureiro, "Filling the Gaps of Vehicular Mobility Traces Categories and Subject Descriptors," Proc. 18th ACM Int. Conf. Model. Anal. Simul. Wirel. Mob. Syst. (MSWiM '15), pp. 47–54, 2015.

[7] A. Keler, J. M. Krisp, and L. Ding, "Detecting Vehicle Traffic Patterns in Urban Environments Using Taxi Trajectory Intersection Points," Geo-Spatial Inf. Sci., vol. 20, no. 4, pp. 333–344, 2017, doi: 10.1080/10095020.2017.1399672.

[8] M. Fang, L. Tang, Z. Kan, X. Yang, T. Pei, and Q. Li, "An Adaptive Origin-Destination Flows Cluster-Detecting Method to Identify Urban Mobility Trends," pp. 1–17, 2021, [Online]. Available: http://arxiv.org/abs/2106.05436.

[9] E. R. Magsino, A. J. Abello, and J. M. Lalusin, "Taxi Hotspots Identification through Origin and Destination Analysis of Taxi Trips using K-means Clustering and H-indexing," J. Phys. Conf. Ser., vol. 1997, no. 1, 2021, doi: 10.1088/1742-6596/1997/1/012006.

[10] S. Mohammed, A. H. Alkhereibi, A. Abulibdeh, R. N. Jawarneh, and P. Balakrishnan, "GIS-based spatiotemporal analysis for road traffic crashes; in support of sustainable transportation Planning," Transp. Res. Interdiscip. Perspect., vol. 20, no. January, p. 100836, 2023, doi: 10.1016/j.trip.2023.100836.

[11] S. Dutta, A. Das, and B. K. Patra, "CLUSTMOSA: Clustering For GPS Trajectory Data Based on Multi-Objective Simulated Annealing to Develop Mobility Application," Appl. Soft Comput., vol. 130, p. 109655, 2022, doi: 10.1016/j.asoc.2022.109655.

[12] S. Wang, Z. Bao, J. S. Culpepper, and G. Cong, "A Survey on Trajectory Data Management, Analytics, and Learning," ACM Comput. Surv., vol. 54, no. 2, 2021, doi: 10.1145/3440207.

[13] D. Zhou, R. Hong, and J. Xia, "Identification of Taxi Pick-Up and Drop-Off Hotspots Using The Density-Based Spatial Clustering Method," in CICTP 2017: Transportation Reform and Change - Equity, Inclusiveness, Sharing, and Innovation - Proceedings of the 17th COTA International Conference of Transportation Professionals, 2018, no. January, pp. 196–204, doi: 10.1061/9780784480915.020.

[14] X. Qian, W. Zhang, S. V. Ukkusuri, and C. Yang, "Optimal Assignment and Incentive Design in The Taxi Group Ride Problem," Transp. Res. Part B Methodol., vol. 103, pp. 208–226, 2017, doi: 10.1016/j.trb.2017.03.001.

[15] A. Fielbaum, "Optimizing a Vehicle's Route in an On-Demand Ridesharing System in Which Users Might Walk," J. Intell. Transp. Syst. Technol. Planning, Oper., vol. 26, no. 4, pp. 432–447, 2022, doi: 10.1080/15472450.2021.1901225.

[16] J. Tang, Urban Travel Mobility Exploring with Large-Scale Trajectory Data. Elsevier Inc., 2018.

[17] R. K. Gunawan and Susilawati, "A Study of Spatiotemporal Distribution of Mobility-On-Demand in Generating Pick-Up/Drop-Offs Location Placement," Smart Cities, vol. 4, no. 2, pp. 746–766, 2021, doi: 10.3390/smartcities4020038.

[18] H. Yu, Z. Li, G. Zhang, P. Liu, and J. Wang, "Extracting and Predicting Taxi Hotspots in Spatiotemporal Dimensions Using Conditional Generative Adversarial Neural Networks," IEEE Trans. Veh. Technol., vol. 69, no. 4, pp. 3680–3692, 2020, doi: 10.1109/TVT.2020.2978450.

[19] L. Zhang, Z. He, X. Wang, Y. Zhang, J. Liang, and G. Wu, "Pick-Up Point Recommendation Using Users' Historical Ride-Hailing Orders," in Wireless Algorithms, Systems, and Applications, 2022, pp. 393–405.

[20] Y. Wang and J. Ren, "Taxi Passenger Hot Spot Mining Based on a Refined K-Means++ Algorithm," IEEE Access, vol. 9, pp. 66587–66598, 2021, doi: 10.1109/ACCESS.2021.3075682.

[21] R. Rafiq and M. G. McNally, "An Exploratory Analysis of Alternative Travel Behaviors of Ride-Hailing Users," Transportation (Amst)., vol. 50, no. 2, pp. 571–605, 2023, doi: 10.1007/s11116-021-10254-9.

[22] X. Huang, Y. Yin, S. Lim, G. Wang, B. Hu, and J. Varadarajan, "Grab-Posisi: An Extensive Real-Life GPS Trajectory Dataset in Southeast Asia," in Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Prediction of Human Mobility, PredictGIS 2019, 2019, pp. 1–10, doi: 10.1145/3356995.3364536.

[23] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226–231.

[24] D. Deng, "DBSCAN Clustering Algorithm Based on Density," Proc. - 2020 7th Int. Forum Electr. Eng. Autom. IFEEA 2020, pp. 949–953, 2020, doi: 10.1109/IFEEA51475.2020.00199.

[25] J. Ha, M. Kambe, and J. Pe, Data Mining: Concepts and Techniques, 3rd ed. Elsevier Inc., 2011.

[26] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to The Interpretation and Validation of Cluster Analysis," J. Comput. Appl. Math., vol. 20, no. C, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

Learning Engagement of Children with Dyslexia Through Tangible User Interface: An Experiment

Siti Nurliana Jamali[1], Novia Admodisastro[2], Azrina Kamaruddin[3], Sa'adah Hassan[4]
Faculty of Computer Science and Information Technology
Department of Software Engineering and Information System
Universiti Putra Malaysia, Serdang, Malaysia

***Abstract*—This paper presents the evaluation of a mobile application employing Tangible User Interface (TUI) technology to enhance the educational involvement of children experiencing dyslexia. The primary objective of this application is to assist these children in overcoming challenges related to reading, spelling, pronunciation, and writing, issues often associated with lower self-esteem and dissatisfaction in an academic setting. The study adopts a User-Centered Design (UCD) approach, focusing on the specific needs and preferences of children with dyslexia during development. The evaluation involved 30 children with dyslexia, divided into two groups: a control group utilizing the non-tangible DisleksiaBelajar mobile app (DB) and a treatment group utilizing the DisleksiaBelajar 3D Tangible (DB3dT) app, which incorporates tangible elements. Results indicated that the DB3dT app achieved significantly higher usability scores (79.5%) compared to the DisleksiaBelajar app (51%). Furthermore, the treatment group utilizing the DB3dT app surpassed the control group in learning performance. In summary, the evaluation demonstrated that integrating tangible elements into the DB3dT app notably enhanced the learning experience for children with dyslexia when compared to the non-tangible DisleksiaBelajar app. The children exhibited increased engagement and a willingness to repeat activities, suggesting potential advancements in learning outcomes and performance.**

Keywords*—*Dyslexia; Tangible User Interface; mobile application; user centered design; engagement

I. INTRODUCTION

Dyslexia is a learning disorder which is commonly found to have difficulty explicitly in language. Children with dyslexia tend to have language difficulty that leads them to be incompetent in reading skills, word recognition, differentiate the sound of letters, recognize mirror letters, and create the syllables in the sentences. Besides, dyslexic students also may face difficulty with spelling, writing, and speaking. In current teaching method, dyslexic students have limitation as the study material unable to offer feedback in learning and multisensory technique is not being embraced appropriately [1].

Multisensory techniques are required to allow dyslexic students to use all senses such as touch, see, hear, and kinesthetic movements in learning to read, make sound of the letters correctly and recognize, and distinguish the mirror letters. Besides, the current teaching module has constraints in terms of delivery to the students while learning. For instance, teachers are required to teach single sound values or letter sounds, word recognition and phonological awareness module such as rhyming and blending the words to the students which relying heavily on the teachers as well as required to teach one- to -one to students which lead to extremely labor intensive, and prolonged in performing the teaching procedure. This teaching module can be a real challenge to deliver and difficult to produce the sound of the letters correctly to students.

An appropriate tangible interaction learning model is developed to promote interactive and engaging learning experiences for dyslexic students to improve the current teaching approach [32]. One technique that has extremely beneficial in learning for dyslexic students is incorporating Tangible User Interface (TUI). TUI provides numerous advantages such as fun learning, collaboration, and support for children with dyslexia in learning mostly in language [2, 3]. TUI allows users to interact with physical objects in the real world, connecting to the digital information in the virtual world and including spatial and embodied facilitation. The advancement of interactive systems has been propelled by the increasing availability of novel devices and methods of interaction. Users now have access to a range of innovative interactive technologies across various application domains, including natural user interfaces, multi-touch displays, cameras, and sensor-based interactions [28].

TUI has been applied with success to children with special education needs. The works in [12, 13, 14, 15, 16, 17, 31] collectively demonstrate the benefits of tangible interaction for children with Dyslexia and Attention Deficit Hyperactivity Disorder (ADHD). TUI provides a suitable method for engaging children in playful learning and has been shown to facilitate improvements in reading, spelling, writing, and letter sound correspondence skills, as well as higher retention abilities in learning environments. The incorporation of tactile and multisensory elements enhances the learning experience, promoting exploratory activities and offering a playful learning environment that supports the unique needs of children with dyslexia. Overall, TUI has proven to be an effective approach in supporting the learning and development of children with dyslexia.

In research [29] discussed systematic mapping of toy user interfaces which focus on physical tangible toys and revealed the opportunities of TUI in education domain. As an example, in [30], the authors furnish a comprehensive examination of the TUI. They delve into its functional characteristics, present various application cases, and explore the design and application considerations related to TUI in the context of

education. This paper presents the evaluation of TUI system using experiment for children with dyslexia in improving learning engagement.

The structure of the study is based on the following: Section II will present the literature review, Section III will present the methodology used in the research, Section IV and Section V will present the results and discussion, Section VI will present the conclusion and future work.

II. Literature Review

A. Tangible User Interface (TUI)

Over the years the benefits of TUI have been rapidly expanded to build embedded and cooperating user experiences. TUI allows the interaction of physical items known as tangibles with computer applications. TUI has several inherent benefits compared to traditional user interfaces (e.g., keyboard or mouse interfaces), with a natural environment and rapid haptic feedback when users get both digital and physical feedback. Research demonstrates TUI enables students to participate in learning activities cognitively, emotionally, physically, and socially. TUI is also recognizable to students who have difficulty with learning such as dyslexics. Dyslexia is an impairment in the language that affects reading, writing, speaking, and listening. The conventional dyslexia learning using multi-sensory procedure has proven to be effective in assisting individuals to learn. However, owing to its rigorous, extended, and one-on-one teaching procedure, this method is highly demanding.

For children with dyslexia, who benefit from multisensory learning, TUI offers advantages through tactile and kinesthetic modalities that are absent in GUI. These modalities enhance the learning experience for these children, providing them with a more engaging and effective learning environment. For children with dyslexia, it is crucial to incorporate tactile and kinesthetic senses when designing TUI systems for teaching them to read and write. By considering their specific needs, the design space must include necessary guidelines that highlight not only the interface but also the learning styles, learning activities, dyslexia learning methods, student's level, and feedback.

TUI in children's learning has garnered significant interest due to its benefits. One such benefit is that tangible interaction, which utilizes embodiment and various forms of feedback (e.g., audio, visual, and haptic), supports different learning styles of children [4]. TUI in learning is commonly incorporated into play activities, which have the potential to promote the cognitive, motor, and physical development of children. When children play, they engage their cognition to think, explore, and enjoy the learning process. However, ensuring children's engagement in learning poses significant challenges [5, 6].

B. Engagement

One of the most important aspects in learning performance is student engagement. This is because level of engagement leads to improved retention, which improves learning performance [7]. In the teaching and learning of children with dyslexia, engagement functions as a nonverbal behavior that necessitates attention. In the context of children with dyslexia, learning engagement is often a big struggle due to the difficulties they face such as They may feel frustration and anxiety due to their difficulties in reading and writing, leading to negative attitudes toward learning, emotional outbursts, or withdrawal from academic tasks [8]. These struggles can impact self-esteem, giving rise to feelings of inadequacy and reluctance to participate in school activities [9]. As a coping mechanism, children with dyslexia may exhibit avoidance behaviors, such as avoiding reading or writing tasks, resisting class participation, or hiding their difficulties from others [10]. Consequently, they may struggle to maintain focus during reading or writing tasks, leading to distractibility and difficulty staying on task. This can significantly affect their overall academic performance and hinder their ability to engage in learning activities and follow instructions [11].

C. Related Works

TUI has been utilized as an assistive technology in educational settings for children with learning difficulties. In this context, six TUI learning systems related to language learning for children with special needs are discussed. The first system, Character Alive [12] a TUI system to support children with dyslexia aged between 5-7 years old in reading and writing skill was developed. The system emphasized on the learning Chinese literacy acquisitions for Mandarin language and provides multisensory approach to allow children in reading and writing of the Chinese characters and words. The system incorporates tangible objects which use dynamic color cues, 2D tangible cards and provide intuitive feedback such as audio, tactile and visual using animations representations. The function of using color cues in the system to inform children on the common patterns of similar characters. Besides, children can arrange and organize the 2D tangible cards which allow them on the tactile feedback and kinesthetic learning environment. Moreover, the use of animation in the system can capture children attention and enable them to memorize the Chinese characters effectively. Though, the work is still in progress which the researcher has not conducted any usability testing to evaluate in terms of the effectiveness of the system.

The second system PhonoBlocks [13] a TUI system developed for reading that uses dynamic colour cues embedded in 3D tangible letters to provide additional decoding information and modalities for children aged 5-8 years old, who are having difficulty learning to decode English letter-sound pairs. The work has addressed several TUI design guidelines in assisting dyslexic to read such as spatiality, various types of interaction modalities, and multiple ways of letter representation as well as structured procedures. PhonoBlocks allows simultaneous use of visual, auditory as well as kinesthetic or tactile approach in both physical and digital representations. It also focuses on the 3D design of tangible letters which facilitates dyslexic children to learn letter as a basis rather than words because they often struggle with letter-sound correspondences and mirror-letter such as b and d.

Third, Block Talks [14] is a system that combines TUI and augmented reality (AR) to support children aged 8-10 in learning English sentence construction. The system utilizes

tangible blocks that represent different elements of a sentence, such as subjects, verbs, and objects. Children can physically manipulate and arrange these blocks to create meaningful sentences. This hands-on approach provides a structured and scaffolded learning experience that fosters language development and helps children grasp grammar rules and sentence formation. The engagement and fine motor skills involved in the tangible interaction promote a deeper understanding of sentence structure. Additionally, AR technology enhances the learning experience by providing visual and auditory feedback, contextual information, and interactive storytelling elements. For example, through the AR display on the screen, children can better comprehend sentence construction concepts and see their constructed sentences come to life. Evaluation results show that Block Talks effectively engages children and promotes their understanding of sentence structure. The use of TUI with the blocks stimulates hands-on learning and facilitates the development of language skills, including grammar and syntax.

Fourth, Interactive Fruit Panel (IFP) [15] is a TUI serious game developed to support children with special needs in learning an alternative communication system. IFP offers three key advantages in facilitating communication skills for children with special needs. Firstly, it utilizes a tangible interface in the form of a fruit panel, allowing children to physically interact with the panel. This hands-on approach promotes engagement, fine motor skills, and a multisensory learning experience. Secondly, IFP incorporates a serious game element, making the learning process enjoyable and motivating for children. Finally, IFP is specifically designed to address the unique needs of children with special needs in learning alternative communication. It provides a structured and intuitive interface that supports language and communication development, enabling children to express themselves effectively. Evaluation findings indicate that when children played with IFP, they showed increased concentration compared to using traditional games. Moreover, the use of IFP resulted in less distraction for the children.

Fifth, Tactile Letters is a multimodal TUI tabletop system developed to teach English alphabet sounds to children with dyslexia aged five to six years old [16]. The system utilizes texture cues in the form of two sets of tangible letters, each consisting of 24 letter cards. Children can choose to interact with the 3D tangible letters or use the letter cards on the interactive tabletop. When children connect the tangible letters correctly, audio feedback is provided to reinforce their learning. The design guidelines of this work emphasize spatiality, allowing children with dyslexia to decode and arrange the letters in their environment. Additionally, multiple senses, including audio, visual, and tactile, are incorporated into the prototype. By providing a hands-on and multisensory learning experience, Tactile Letters aims to improve letter recognition and enhance visual processing for children with dyslexia. The evaluation results indicate that the incorporation of texture cues successfully captures the participants' attention and increases their engagement with the program. The inclusion of frequent kinesthetic movement promotes brain activity and aids in word retention, surpassing the effectiveness of traditional flashcards. Participants expressed interest in the program and gained knowledge from the phonics-based reading program. Tactile Letters serves as a potential reading tool for children with dyslexia, assisting them in independent reading practice and learning, complementing their regular classroom instruction.

Six, TraceIt is another tool designed to support children with dyslexia in reading through a hands-on and interactive learning activity [17]. TraceIt utilizes air tracing interaction to improve letter formation and enhance visual processing. The program incorporates color-based physical objects and offers a multisensory learning experience. The evaluation results reveal that the air tracing interaction technique successfully captures the participants' attention and increases their engagement with the program. The inclusion of frequent kinesthetic movement promotes brain activity and aids in word retention, surpassing the effectiveness of traditional flashcards. Participants not only expressed interest in the program but also gained knowledge from the phonics-based reading program. TraceIt demonstrates good potential as a reading tool for children with dyslexia, providing a playful and interactive learning environment that complements their regular classroom instruction.

The works by [12, 14, 15, 16, 17] collectively demonstrate the benefits of TUI for children with dyslexia. TUI provides a suitable method for engaging children in playful learning and has been shown to facilitate improvements in reading, spelling, writing, and letter sound correspondence skills, as well as higher retention abilities in learning environments. Overall, TUI has proven to be an effective approach in supporting the learning and development of children with dyslexia.

III. METHODOLOGY

User-Centered Design (UCD) was adopted in this study, which involves understanding the user and their needs and context throughout the entire process from user requirements to evaluation stages. UCD consists of four phases: understanding the context, specifying user requirements, designing solutions, and evaluating against the requirements [18]. In this research, UCD was incorporated to develop a prototype and gain a deep understanding of users by involving them in the design process and product development. The focus was on meeting users' needs and improving their experience with the proposed solutions.

A. The Experiment

A quasi experiment was conducted in DAM centers in Ampang and Bandar Baru Bangi that aimed to evaluate the effectiveness of the DB3dT app in engaging students with dyslexia. Due to the limited number of participants available for randomization, a quasi-experiment was chosen. The participants in this study were children with dyslexia who were already attending predetermined classes in the DAM centers and were assigned to an eight-week intervention conducted within their school setting. These children were classified based on their level of study, which included beginner, intermediate, and advanced levels. The experiment consisted of two groups: the control group and the treatment group. The control group comprised 15 children with dyslexia who did not receive any stimulus. In their case, a non-tangible

approach using the DisleksiaBelajar mobile app was provided. On the other hand, the treatment group consisted of 15 children with dyslexia who received stimulus in the form of the DB3dT app. For this group, tangible objects such as tangible cards and toys were prepared to enhance their interaction with the app (see Fig. 1). The children in the treatment group engaged with the DB3dT app through task activities specifically prepared for them.

Fig. 1. The tangible objects- flashcards, 3D alphabet and toy.

B. The Procedure

The experiment was conducted on school property, utilizing a private space provided by the school principal. The location for the evaluation activity was assigned by the teacher and chosen based on its convenience, accessibility, and suitability for video recording purposes [19]. The decision to conduct the experiment in a school setting was aimed at providing the participants with a familiar and comfortable environment throughout the procedure [20]. Creating a comfortable atmosphere was important to promote a sense of ease for the participants [21] and ensure their comfort during the session [22].

Fig. 2. Experiment with children.

The session began with an introduction between the student, facilitator, and observer to help the student become acquainted with the setting and the session. The facilitator then explained the instructions regarding the instruments and the tasks required. In the control group, fifteen children were instructed to use the non-tangible approach using the DisleksiaBelajar mobile app. On the other hand, the treatment group, comprised of another fifteen children, were asked to use the tangible approach with the DB3dT app. Throughout the experiment, a facilitator was present near the students to assist them in using the application. An observer was also seated in front of the students to observe their engagement through facial expressions and behavior. Considering the vulnerability factors associated with the participants, such as their age and disabilities, maintaining a high-quality relationship between the facilitator and students was crucial [23] (see Fig. 2). Additionally, the facilitator communicated instructions in the Malay language to ensure clarity and understanding for the children.

During the session, whenever a student required assistance, such as when they needed help with spelling specific terms or when they were unfamiliar with the vocabulary of an object in Malay, support was provided and recorded. If a student struggled significantly with the exercises and remained inactive for a prolonged period, the question was skipped to save time and minimize demotivation. Two mobile tablets were utilized during the experiment, specifically for the DisleksiaBelajar mobile app and the DB3dT app. The DisleksiaBelajar mobile app was developed specifically for children aged between 6 to 12 years old with dyslexia, aiming to enhance their Malay language skills. Both instruments were introduced in this quasi-experiment as entirely new approaches. This was done to ensure that neither group of dyslexic children had any prior knowledge or bias when using these apps. The DB3dT app employed a tangible approach, while the DisleksiaBelajar mobile app utilized a non-tangible approach. These two apps had similar activity modules focusing on phonology, spelling, and reading skills. Prior to the start of the experiment, the children with dyslexia were introduced to and familiarized with the mobile tablets, including learning how to use the camera, and understanding the functions and settings of the devices. For the experiment setup, materials such as a mobile tablet, a video recorder, and a mobile phone were used. The session lasted for 45 minutes and was recorded using a video camera for further analysis.

C. The Instruments

Several data collection instruments were used in this experiment: Tangible approach using DB3dT application, non-tangible approach using DisleksiaBelajar application, System Usability Scale (SUS), Again-Again Table and Observation Form. The SUS serves the purpose of evaluating the system's usability from the perspective of dyslexic children. In addition, the use of a five-point Smileyometer scale, as shown in Fig. 3, allows students to rate their experience with the SUS ranging from 1 to 5, with a rating of 5 representing the most positive response. Fig. 4 displays the ten items that are evaluated in this scale. To ensure comprehension among children, the items were translated into Malay language without altering their original meaning. This

instrument has been validated and previously used in the study conducted by [24]. As some children may not be able to read fluently, facilitators assist in reading the items. The Again-Again table was utilized to evaluate enjoyment, engagement and confirm acceptance of the DB3dT app by children with dyslexia that using 3-point Likert scale of 'Yes, Maybe, No' (*Ya, Mungkin, Tidak*) as shown in Fig. 5. As described by [25], human emotion should be considered when validating user acceptance for assessing user experience. The Again-Again table from the Fun Toolkit [26] was utilized as a self-reporting approach for children. According to [24, 27], the Again-Again table was utilized because students are highly likely to experience enjoyment when an activity is engaging, making them willing to do it again.

D. The Application

The DisleksiaBelajar 3d Tangible (DB3dT) application was developed based on the proposed learning model [32]. The DB3dT app enables children to interact with digital information using tangible objects such as tangible letter cards, alphabet blocks, and toys in the physical environment. In study [12], the author developed a reading augmented reality using 2D and 3D cards to compose a character within a word and sentences in supporting children Chinese reading and writing skills.

This DB3dT facilitates phonology, spelling, and reading skill development specifically tailored to dyslexic learning patterns. Through intuitive interaction with tangible objects, children can construct words from syllables and view augmented reality 3D overlay content on a screen during the learning activity. The application incorporates various sensory experiences, including tactile, auditory, visual, and kinesthetic elements, to strengthen literacy skills. There are five learning activity modules in DB3dT app as depicted in Fig. 6.

Fig. 3. Smileyometer.

	Soalan	Skala
1	Saya akan sentiasa menggunakan aplikasi DB3dT ini.	
2	Aplikasi DB3dT ini mudah digunakan.	
3	Saya berjaya menggunakan aplikasi DB3dT ini tanpa bantuan orang lain.	
4	Aplikasi DB3dT ini berfungsi dengan baik.	
5	Saya suka menggunakan aplikasi DB3dT ini.	
6	Saya faham arahan yang digunakan dalam aplikasi DB3dT ini.	
7	Saya seronok apabila menggunakan aplikasi DB3dT ini.	
8	Aplikasi DB3dT ini membantu saya dalam mempelajari Bahasa Melayu dengan baik.	
9	Saya berjaya menggunakan aplikasi DB3dT ini tanpa sebarang masalah.	
10	Saya yakin menggunakan aplikasi DB3dT ini.	

Fig. 4. The 10-items for system usability scale.

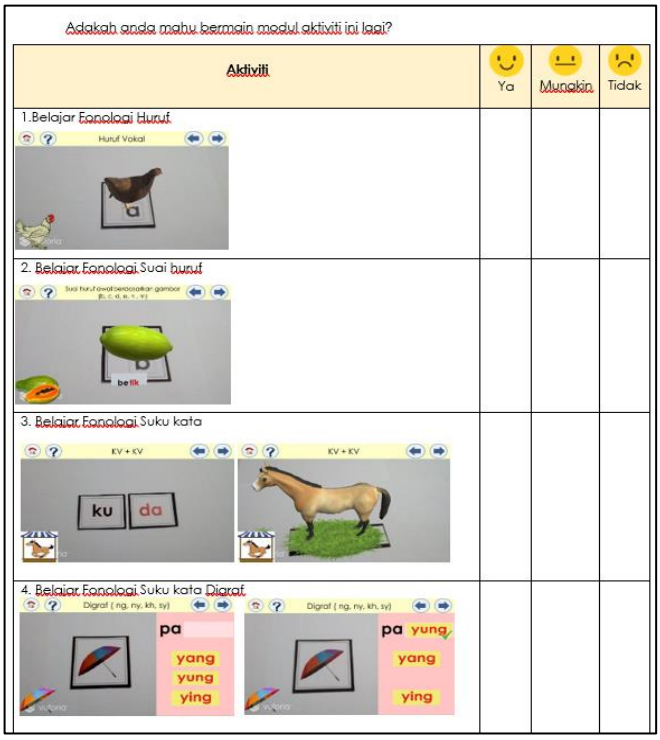


Fig. 5. Again-Again table.

Fig. 6. Five learning activity modules in DB3dT app.

1) Huruf, Suai Huruf and Suku Kata modules: This module focuses on phonology skills, specifically vowels, consonants, and syllable exercises. The vowel activity evaluates the student's knowledge of vowels, requiring them to differentiate vowel sounds and match the world's first letter in the Huruf module. Pictures are provided alongside the words to assist students in selecting the correct vowel. On the other hand, the consonant practice addresses five common consonant mistakes in the Malay language, as shown in Fig. 7. This task assesses a learner's ability to identify consonant sounds and match them correctly. The syllables exercise consists of two types: Vowel + Consonant Vowel (V+CV) and three-syllable combinations (CV+CV+CV), among others. This activity evaluates the learner's ability to recognize pictures, choose the correct syllable, and drag it into the appropriate word. To provide a challenge, similar syllables are also presented as distractors. In the Suai Huruf module, students match initial letters with corresponding pictures. This module introduces six letters: b, d, c, e, n, and m, which are

known to cause confusion among dyslexic children. The objective is to help dyslexic students understand and recognize these letters, as shown in Fig. 8. The Suku Kata module focuses on sorting and recognizing CV and CVCV syllables using picture hints, as depicted in Fig. 9.

Fig. 7. Vowel and consonant activity in the *Huruf* module.

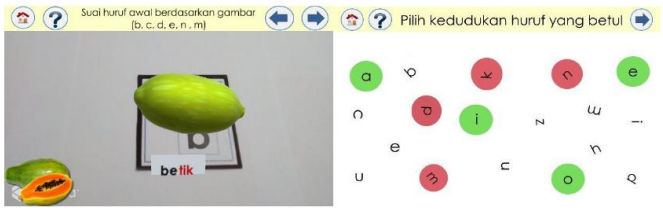


Fig. 8. Matching initial letters with picture and identify correct letter position in the *Suai Huruf* module.

Fig. 9. Sorting syllables using CVC and CVCV in the *Suku Kata* module.

2) Ejaan (Spelling) module: The Ejaan module is designed to assist dyslexic children in identifying spelling based on picture hints provided at the bottom left of the screen. The students are asked to manipulate tangible letters to construct the correct spelling. Each letter is accompanied by its corresponding sound. An example of the activity involves spelling the word "kuda" using CVCV words, as shown in Fig. 10. Additionally, dyslexic students learn to spell digraphs (ny, ng, kh, sy) and diphthongs (ai, au, oi), as depicted in Fig. 11. Finally, the module includes an activity where dyslexic students need to identify the correct spelling of colors, helping them practice spelling words correctly, as depicted in Fig. 12.

Fig. 10. Spell and divide the CVCV in the *Ejaan* module.

Fig. 11. Spell and divide the CVCV in the *Ejaan* module.

Fig. 12. Choosing correct colors spelling in the *Ejaan* module.

3) Bacaan (Reading) module: This module provides children with short passages themed around animals in the zoo. Students have the option to read the passages on their own or listen to the system narrator by pressing the audio icon. They can then select and interact with tangible animal figurines displayed on the screen. Four animal figurine toys are available for selection: Tiger (Harimau), Rhinoceros (Badak sumbu), Turtle (Penyu), and Orang Utan. Students can scan the tangible animals, triggering an augmented video of the animal along with accompanying sounds and a narration of its story. Following the Bacaan module, dyslexic children can engage in comprehension exercises by clicking on the Latihan icon at the top of the screen. These exercises require them to answer questions based on the animal stories they previously learned, with three exercises provided for each animal, as shown in Fig. 13.

Fig. 13. Reading a short passage and learning comprehension modules.

IV. Results

In this study the primary objective is to examine the effectiveness of the DB3dT app to enhance student engagement in learning for children with dyslexia through a

quasi-experiment. We presented descriptive statistics to determine the average on-task behavior of the students as a measure of the effectiveness of the DB3dT app. The experiment results were analyzed for usability testing using the System Usability Scale (SUS), usability observations, user experience using the Again-Again Table, and a checklist of children's performance in learning activities.

A. Descriptive Statistic

According to the results presented in Table I, the average on-task time for beginner students in the control group, who used the non-tangible approach (DisleksiaBelajar mobile app), was recorded as 14 minutes and 93 seconds (SD= 2.72). In comparison, for the treatment group that used the tangible approach (DB3dT app), the average on-task time was recorded as 33 minutes and 58 seconds (SD= 6.18). This indicates that students in the treatment group were more engaged with the activity modules, even at the beginner level. Furthermore, among students at the intermediate level, those in the control group exhibited an average on-task time of 12 minutes and 49 seconds (SD= 1.37), while the treatment group recorded an average of 31 minutes and 30 seconds (SD= 7.74). This translates to a twofold increase in on-task time for the treatment group indicating that students using the DB3dT app retained their engagement for a longer period compared to those using the DisleksiaBelajar mobile app. Some students mentioned that the DisleksiaBelajar mobile app was too easy for them, leading them to complete tasks quickly. Finally, the analysis of average on-task time between the control group and treatment group for advanced students was conducted.

The results showed that students in the control group were engaged with the activity modules for approximately 11 minutes and 39 seconds (SD= 2.72) using the DisleksiaBelajar mobile app, whereas the treatment group recorded an average of 22 minutes and 54 seconds (SD= 3.73) using the DB3dT app. These values indicate that there was a twofold difference in on-task time between the two activity modules due to the higher mastery level of the advanced students. Since advanced students were more proficient in performing the activity modules, their engagement time was shorter compared to beginner and intermediate levels. Nonetheless, the treatment group still retained twice the engagement time compared to the control group. This suggests that using the tangible approach (DB3dT app) in the treatment group resulted in higher engagement levels with the activity modules compared to the non-tangible approach used in the DisleksiaBelajar mobile app. In general, the adoption of a tangible approach in learning activity modules plays a crucial role in measuring student engagement based on on-task time. The results clearly indicate that students using the DB3dT app spend more time engaged compared to students using the DisleksiaBelajar mobile app.

B. System Usability Scale

Based on these scoring rules, the average score for the DB3dT app (tangible approach application) is 79.5%, whereas the average score for the DisleksiaBelajar app (non-tangible application) is 51%. In conclusion, the DB3dT app with its integration of tangible elements like augmented models, animations, videos, and text, was found to be more usable for students with dyslexia compared to the DisleksiaBelajar mobile app. This can be seen in the analysis of each questionnaire item, which showed more positive responses (odd numbered items) for the DB3dT app versus the DisleksiaBelajar app (see Fig. 14). The DB3dT app also had fewer negative responses (even numbered items) than the DisleksiaBelajar app for most items. For item 5, which asked about system function integration, the treatment group scored 4.3 and the control group scored 4.1, a minor 0.2-point difference. This small difference is likely because students felt the functions were good in both apps. Overall, the tangible elements incorporated in the DB3dT app made it more engaging and usable than the non-tangible DisleksiaBelajar mobile app for children with dyslexia.

TABLE I. RESULT FOR ON-TASK TIME FOR BOTH GROUPS

	Control Group					Treatment Group				
Beginner	18.00	17.17	15.17	12.24	12.1	37.37	24.83	30.19	40.54	34.99
Average	14.93					33.58				
SD	2.72					6.18				
Intermediate	11.87	14.34	13.51	11.74	11.03	27.31	26.3	43.99	33.4	25.51
Average	12.49					31.30				
SD	1.37					7.74				
Advanced	10.37	14.89	13.52	9.89	8.3	27.74	18.63	21.12	20.21	25.02
Average	11.39					22.54				
SD	2.72					3.73				

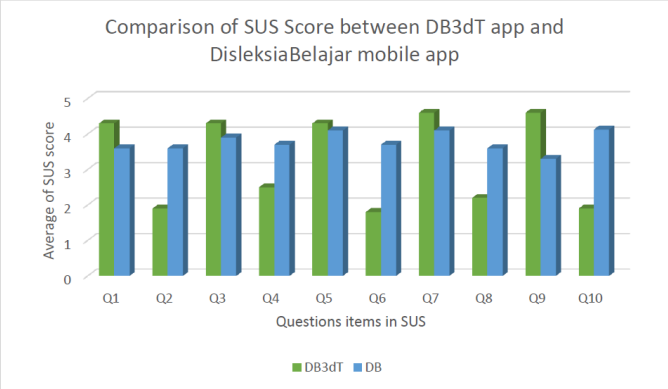


Fig. 14. Comparison of SUS score for DB3dT app and DisleksiaBelajar mobile app.

C. Usability from the Observation

In this study, observational notes were utilized to capture the students' signs, expressions, and behavior while interacting with both tangible and non-tangible applications. These notes were used to make a comparison between the DB3dT app as a tangible application and the DisleksiaBelajar mobile app as a non-tangible application as shown in Table II. During the experiment, it was observed that each student had different abilities in responding to the applications. Students at the beginner level encountered challenges in answering comprehension questions and struggled with reading lengthy passages. On the other hand, students at the intermediate and advanced levels demonstrated confidence in engaging with the activities and were able to answer comprehension questions effectively. Based on the observations, two distinct types of students emerged: those who were on-task (engaged) and those who were off-task (disengaged). Engaged students exhibited dynamic concentration and active interaction with the application. They required less supervision and demonstrated independent participation. In contrast, disengaged students displayed various behavioral issues such as excessive talking, difficulty sitting still, hyperactivity, fear of making mistakes, easy distractibility, and a tendency to give up quickly during the activities. The results indicated that the tangible approach offered additional advantages for both engaged and disengaged students. Using the DB3dT app, both types of students remained engaged for longer periods. The DB3dT app's activities allowed students to interact with the learning materials based on their level of engagement, encompassing cognitive, behavioral, and emotional aspects.

D. User Experience using Again-Again Table

The Again-Again table was also used to evaluate enjoyment and engagement for children with dyslexia. The children scored each DB3dT app activity interface that they wanted to play again using a 3-point Likert scale of 'Yes, Maybe, No'. Thirty children completed the scoring. Frequencies of the scales were calculated to analyze the activities. This study relates to the work of [27], which used enjoyment as a metric to quantify students' emotions while interacting with a mobile app for those with speech delays. The result in Table III indicates that most children highly enjoyed certain DB3dT activities, such as *Fonologi Suku Kata* and *Ejaan*, with all students willing to play them again. They found these activities easy and engaging, attributing their enjoyment to the fun tangible letter cards they could arrange on the board, interesting 3D visuals on screen, and letter sound feedback. Additionally, the children needed minimal assistance during these tasks, displaying genuine enjoyment and enthusiasm while using the app. In contrast, activities like *Fonologi Huruf*, *Fonologi Suku Kata Digraf*, and *Latihan Pemahaman* had only 73% agreement among students to play again. Some students found these activities more challenging due to reading difficulties, needing more facilitator support to identify sentences.

TABLE II. COMPARISON OF STUDENT'S RESPONSE FROM OBSERVATIONAL NOTES

Study Level	DisleksiaBelajar mobile app (DB) (non-tangible)	DB3dT app (tangible)
Beginner	• Displayed no reaction when they were asked about the letters. • Kept on trying and making mistakes of each letter.	• Amazed with the tangible objects. • Kept playing with the tangible cards and figurine toys.
	• Unable to sit still and were restless.	• Got distracted with the tangible objects/ cards.
	• Confused by some letters, especially 'b' and 'd', and were unable to read long passages	• Unable to sit still and restless.
	• Asked for help from the facilitator when doing the activity.	• Followed the words and letters when learning letters.
	• Bored and wanted to quit the activity.	• Abled to engage with the learning activity even when facing technical issue.
Intermediate	• Showed excitement when answering the questions.	• Displayed astonishment as soon they saw 3D objects appearing on the screen.
	• Confidently saying 'yes' and finding it easy after completing the activity.	• Clapped their hands when scanning tangible objects.
	• Confused by some letters, especially 'b' and 'd', and were unable to read long passages	• Smiled while doing the activity.
Advanced	• Showed no expression and were bored because the activity was deemed easy.	• Abled to read the passages of the animal story.
	• Felt shy but still wanted to do the activity.	• Excited to use the black board to arrange letters.
	• Displayed no expression when they answered the questions correctly.	• Smiled when they were able to answer the learning comprehension questions.

Additionally, 80% of children agreed to play the *Fonologi Suai Huruf* activity again, while only 53% agreed for the *Bacaan* activity. Feedback showed these activities were more difficult, with some struggling to read fluently, blend syllables, and comprehend long sentences and passages. Overall, the findings indicate children prefer activities with tangible objects like cards and letters, as these elements greatly influence their interest in learning. To enhance the learning experience, incorporating visual, audio, and kinesthetic

components is crucial. In conclusion, the Again-Again table demonstrates the DB3dT app is enjoyable to use. This confirms that the children's experiences and feelings while engaging with the DB3dT app, as revealed in Table III, align with the Again-Again table results. The tangibility elements made activities more engaging and fun for children with dyslexia.

E. Performance Checklist

In this experiment, student engagements were evaluated using a performance checklist that consisted of various questions related to their learning activities. The purpose of utilizing this checklist was to assess their progress in the learning tasks and obtain a comprehensive understanding of their learning outcomes. The questions covered different aspects of phonology learning, including identifying vowel and consonant letters, matching letters, learning syllable patterns (CVCV, CVC), spelling, and reading short paragraphs with 4, 5, and 6 sentences. The total possible score for these questions was 40. Both groups were given this performance checklist as an exercise based on the learning modules they had previously completed.

TABLE III. FREQUENCY RESPONSE TO AGAIN-AGAIN TABLE FOR DB3DT APP

Activity	Would you like to play again?	Frequency
Fonologi Huruf (Letter Phonology)	Yes	**11**
	Maybe	**4**
	No	-
Fonologi Suai Huruf (Letters Matching Phonology)	Yes	**11**
	Maybe	**3**
	No	**1**
Fonologi Suku Kata (Syllables Phonology)	Yes	**15**
	Maybe	-
	No	-
Fonologi Suku Kata Digraf (Digraph Syllables Phonology)	Yes	**11**
	Maybe	**4**
	No	-
Fonologi Suku Kata Diftong (Diphthong Syllables Phonology)	Yes	**12**
	Maybe	-
	No	**3**
Ejaan (Spelling)	Yes	**15**
	Maybe	-
	No	-
Bacaan (Reading)	Yes	**8**
	Maybe	**6**
	No	**1**
Latihan Pemahaman (Comprehension Exercise)	Yes	**11**
	Maybe	**2**
	No	**2**

The average total score of all 15 participants was 67.5%. It is worth noting that the control group scored slightly lower, at 60%, while the treatment group achieved a score of 65% at the beginner level, as depicted in Fig. 15. At the intermediate level, the control group attained an average score of 72%, whereas the treatment group scored 77% (see Fig. 16). Lastly, at the advanced level, the control group obtained an average score of 69.5%, whereas the treatment group excelled with an average score of 89.5% (see Fig. 17). These results suggest that the children with dyslexia in the treatment group displayed higher levels of engagement in the learning activities across all three proficiency levels compared to the control group.

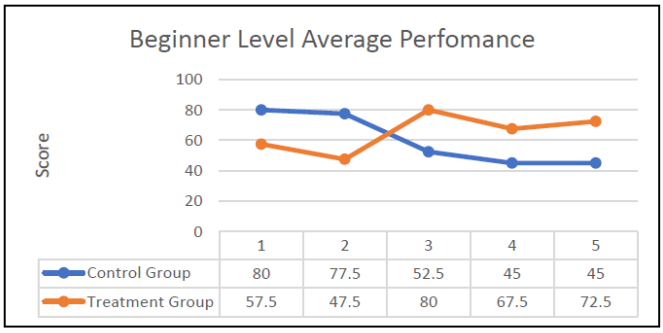


Fig. 15. Beginner level average performance.

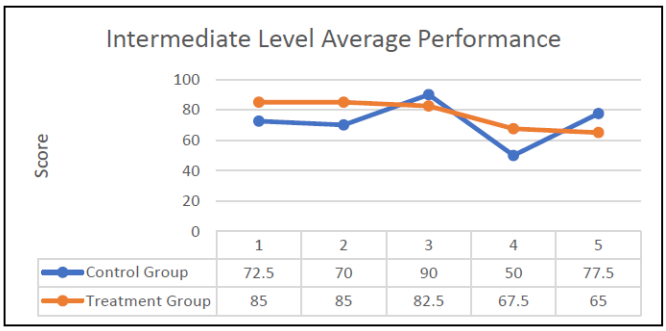


Fig. 16. Intermediate level average performance.

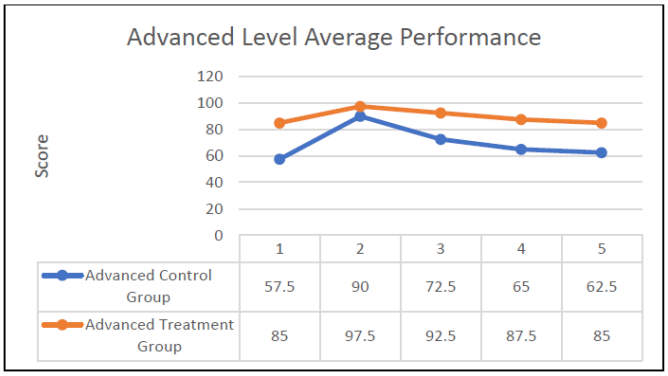


Fig. 17. Advanced level average performance.

V. DISCUSSION

This section discusses on evaluating the usability aspects and engagements of the DB3dT app. The evaluation encompasses the main aspect which is the effectiveness of the DB3dT app in promoting children's engagement. To evaluate the effectiveness of the DB3dT app in promoting student engagement, a quasi-experiment was conducted involving dyslexic students. The students were divided into control and treatment groups to compare the results based on their on-task time when using the non-tangible approach using the

DisleksiaBelajar app (control group) and the tangible approach using the DB3dT app (treatment group). The on-task time was utilized as a measure of student engagement. The results indicated that the treatment group, which used the DB3dT app, exhibited better and longer engagement compared to the control group using the DisleksiaBelajar mobile app, additionally, the hypothesis testing revealed a significant effect, indicating that the DB3dT app had a significant impact on improving engagement for children with dyslexia in the learning process.

Additionally, the SUS and Again-Again table were used to measure DB3dT app usability. Results showed the DB3dT app (tangible approach) had higher average usability at 79.5% versus 51% for the DisleksiaBelajar app (non-tangible). The Again-Again table also measured enjoyment and engagement. Most DB3dT activities were found to be fun and enjoyable, with children willing to play them again, indicating higher engagement. This aligns with research showing increased performance from sustained learning engagement [33].

The learning performance checklist and exercises also showed the treatment group had higher engagement, with average total marks of 77.5% compared to 67.5% for the control group. The higher scores for the DB3dT app group indicate its effectiveness at retaining engagement. In summary, usability and engagement metrics showed tangibility elements in the DB3dT app enhanced the learning experience for children with dyslexia versus the non-tangible DisleksiaBelajar app. The children were more engaged and willing to replay activities, suggesting potential learning and performance improvements.

VI. Conclusion and Future Work

Overall, the effectiveness of the DB3dT app in promoting student engagement was evaluated by conducting a quasi-experiment that involved children with dyslexia. In conclusion, the DB3dT app demonstrated positive outcomes in supporting children with dyslexia learning the Malay language. The tangible elements incorporated in the DB3dT app increased engagement and understanding compared to the non-tangible mobile app. Based on the findings of this research, it is evident that the DB3dT app enhances learning activity and improves engagement among children with dyslexia. It also provides a usable approach for children in the learning process. Considering the constraints of limited time and funds, there are several recommendations for future research that can further enhance the understanding and application of the DB3dT app. For future work, we will replicate this research using students with diverse learning disabilities, such as dysgraphia or dyscalculia, would allow for an examination of how different types of learning disabilities impact the behavioral intention to use the DB3dT app accurately. This broader applicability of the results would provide valuable insights into its effectiveness across various learning disabilities.

Acknowledgment

We extend our sincere appreciation to the Dyslexia Association of Malaysia (DAM), Dyslexia Genius Malaysia (DGM), and the expert panels, including educational experts and dyslexia teachers, for their invaluable support in this research. A special acknowledgment goes to all the children with dyslexia who actively participated in this study. We are grateful to the university's ethics committee (JKEUPM) for approving our research application and express our thanks to the Ministry of Higher Education (MOHE) under FRGS/1/2019/ICT01/UPM/02/4 (vot. 5540289) for providing the necessary research funding.

References

[1] Hamid, S. S. A., Admodisastro, N., Kamaruddin, A., Manshor, N., & Ghani, A. A. A. (2017). Informing design of an adaptive learning model for student with dyslexia: a preliminary study. In Proceedings of the 3rd International Conference on Human-Computer Interaction and User Experience in Indonesia (pp. 67-75). https://doi.org/10.1145/3077343.3107577.

[2] Fan, M., Antle, A. N., & Cramer, E. S. (2016). Design rationale: opportunities and recommendations for tangible reading systems for children. In Proceedings of the the 15th international conference on interaction design and children (pp. 101-112). https://doi.org/10.1145/2930674.2930690.

[3] Falcao, T. P., & Price, S. (2010). Informing design for tangible interaction: a case for children with learning difficulties. In Proceedings of the 9th International Conference on Interaction Design and Children (pp. 190-193). ACM. https://doi.org/10.1145/1810543.1810568.

[4] Ishii, H., & Ullmer, B. (1997). Tangible bits: towards seamless interfaces between people, bits and atoms. In Proceedings of the ACM SIGCHI Conference on Human factors in computing systems (pp. 234-241). ACM.

[5] Gray, P. (2013). Free to learn: Why unleashing the instinct to play will make our children happier, more self-reliant, and better students for life, Basic Books.

[6] Ostroff, W. L. (2012). Understanding how young children learn: bringing the science of child development to the classroom. Ascd.

[7] Yu, Z., Yu, L., Xu, Q., Xu, W., & Wu, P. (2022). Effects of mobile learning technologies and social media tools on student engagement and learning outcomes of English learning. Technology, Pedagogy and Education, 31(3), 381-398.

[8] Snowling, M. J., Hulme, C., & Nation, K. (2020). Defining and understanding dyslexia: past, present and future. Oxford Review of Education, 46(4), 501–513. https://doi.org/10.1080/03054985.2020.1765756.

[9] Irma Rachmawati, K. S. P. S., & Pendidkan, J. B. (2019). Demographic characteristics, behavioral problems, andiq profileof children with dyslexia at dyslexia association of indonesia from january-june 2019: a quantitative study (Vol. 12). https://ojs.upsi.edu.my/index.php/JPB/article/view/3057.

[10] Francis, D. A., Caruana, N., Hudson, J. L., & McArthur, G. M. (2019). The association between poor reading and internalising problems: A systematic review and meta-analysis. Clinical Psychology Review, 67, 45-60. https://doi.org/10.1016/j.cpr.2018.09.002.

[11] Zuppardo, L., Serrano, F., Pirrone, C., & Rodriguez-Fuentes, A. (2023). More Than Words: Anxiety, Self-Esteem, and Behavioral Problems in Children and Adolescents with Dyslexia. Learning Disability Quarterly, 46(2), 77-91. https://doi.org/10.1177/07319487211041103.

[12] Fan, M., Fan, J., Antle, A. N., Jin, S., Yin, D., & Pasquier, P. (2019). Character Alive: A Tangible Reading and Writing System for Chinese Children At-risk for Dyslexia. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-6). https://doi.org/10.1145/3290607.3312756.

[13] Antle, A. N. (2015). PhonoBlocks : A Tangible System for Supporting Dyslexic Children Learning to Read. October. https://doi.org/10.1145/2677199.2687897.

[14] Fan, M., Baishya, U., McLaren, E.S., Antle, A.N., Sarker, S., Vincent, A.(2018): Block talks: a tangible and augmented reality toolkit for children to learn sentence construction. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems CHI

2018, Paper No. LBW056. Montreal QC, Canada (2018). https://doi.org/10.1145/3170427.3188576.

[15] Durango, I., Carrascosa, A., Gallud, J. A., & Penichet, V. M. (2018). Interactive fruit panel (IFP): a tangible serious game for children with special needs to learn an alternative communication system. Universal Access in the Information Society, 17, 51-65. https://doi.org/10.1007/s10209-016-0517-5.

[16] Fan, M., & Antle, A. N. (2015). Tactile letters: a tangible tabletop with texture cues supporting alphabetic learning for dyslexic children. In Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction (pp. 673-678). https://doi.org/10.1145/2677199.2688806.

[17] Teh, T. T. L., Ng, K. H., & Parhizkar, B. (2015). Traceit: An air tracing reading tool for children with Dyslexia. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9429, 356–366. https://doi.org/10.1007/978-3-319-25939-0_32.

[18] Good, A., & Omisade, O. (2019). Linking Activity Theory with User Centred Design: A Human Computer Interaction Framework for the Design and Evaluation of mHealth Interventions. Studies in Health Technology and Informatics, 263, 49–63. https://doi.org/10.3233/SHTI190110.

[19] Edwards, R. and Holland, J. (2013). What is qualitative interviewing?. London: Bloomsbury Academic.

[20] Punch, S. (2002). Research with Children. Childhood, 9(3), pp.321-341.

[21] Georgeson, J., Porter, J., Daniels, H. and Feiler, A., (2014) Consulting young children about barriers and supports to learning. European Early Childhood Education Research Journal, 22 (2), pp.198-212. https://doi.org/10.1080/1350293X.2014.883720.

[22] King, N. and Horrocks, C. (2010). Interviewing in qualitative research. Los Angeles: SAGE.

[23] Nind, Melanie (2008) Conducting qualitative research with people with learning, communication and other disabilities: methodological challenges (ESRC National Centre for Research Methods Review Paper, NCRM/012) National Centre for Research Methods 24pp.

[24] Admodisastro, N. (2021). Evaluation of Disleksia Belajar mobile app for assisting dyslexic junior school students to learn the Malay language. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(3), 2230-2235. https://turcomat.org/index.php/turkbilmat/article/view/1172.

[25] Méndez, A. Y. A., Ordóñez, C. A. C., Saltiveri, A. G., & Huitr, J. A. S. (2014). Evaluating interactive systems from an emotional perspective. Revista Científica Guillermo de Ockham, 12(1), 43–49. http://www.redalyc.org/articulo.oa?id=105332478005.

[26] Read, J. C., & MacFarlane, S. (2006). Using the fun toolkit and other survey methods to gather opinions in child computer interaction. Proceeding of the 2006 Conference on Interaction Design and Children IDC 06, 81. https://doi.org/10.1145/1139073.1139096.

[27] Tommy, C. A., Minoi, J. L., & Sian, C. S. (2019). Assessing Fun and Engagement in Mobile Applications for Children with Speech Delay. Applied Mechanics and Materials, 892, 79–87. https://doi.org/10.4028/www.scientific.net/amm.892.79.

[28] Rundo, L., Pirrone, R., Vitabile, S., Sala, E., & Gambino, O. (2020). Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. Journal of biomedical informatics, 108, 103479. https://doi.org/10.1016/j.jbi.2020.103479.

[29] de Albuquerque, A. P., & Kelner, J. (2019). Toy user interfaces: Systematic and industrial mapping. Journal of Systems Architecture, 97, 77-106. https://doi.org/10.1016/j.sysarc.2018.12.001.

[30] Zhou, Y., & Wang, M. (2015). Tangible user interfaces in learning and education. International Encyclopedia of the Social & Behavioral Sciences, 2, 20-25. https://doi.org/10.1016/B978-0-08-097086-8.92034-8.

[31] De la Guía, E., Lozano, M. D., & Penichet, V. M. (2015). Educational games based on distributed and tangible user interfaces to stimulate cognitive abilities in children with ADHD. British Journal of Educational Technology, 46(3), 664-678. https://doi.org/10.1111/bjet.12165.

[32] Jamali, S. N., Admodisastro, N., Kamaruddin, A., Ghani, A. A. A., & Hassan, S. (2019). Design guidelines of tangible interaction learning model for children with dyslexia. International Journal of Advanced Science and Technology, 28(2), 355-362.

[33] Liu, C. C., Chen, W. C., Lin, H. M., & Huang, Y. Y. (2017). A remix-oriented approach to promoting student engagement in a long-term participatory leasrning program. Computers & Education, 110, 1-15. https://doi.org/10.1016/j.compedu.2017.03.002.

FOREX Prices Prediction Using Deep Neural Network and FNF

Asmaa M. Moustafa[1]*, Mohamed Waleed Fakhr[2], Fahima A. Maghraby[3]
Arab Academy for Science Technology & Maritime Transport-College of Computing and Information Technology,
Heliopolis, Cairo Governorate, Egypt

***Abstract*—One of the largest financial markets on the planet is the foreign exchange (FOREX) market. Banks, retail traders, businesses, and individuals trade more than $5.1 trillion in FOREX daily. It is very challenging to predict prices in advance due to the market's complex, volatile, and highly fluctuating nature. In this study, the new FOREX Normalization Function (FNF) is proposed and used with different models to predict the prices of the AUD/USD, EUR/USD, USD/JPY, CHF/INR, USD/CHF, AUD/JPY, USD/CAD, and GBP/USD. Two models are proposed in this study. The first model contains FNF as a normalization and feature extractor, followed by a Convolutional Neural Network (CNN). The second model utilizes FNF and a Support Vector Regressor(SVR). The forecasts are set for a one-day timeframe, with predictions made for 1, 3, 7, and 15 days ahead. The efficient ability of the proposed method to solve the FOREX prediction problem is proven by performing experiments on nine real-world datasets from different currencies. Additionally, the models are evaluated using Mean Absolute Error (MAE) and Mean Squared Error (MSE). Applying the presented models to 9 different datasets improved the results by an average between 0.5% and 58% of MAE.**

***Keywords*—FOREX prediction; CNN; normalization function; SVR**

I. Introduction

FOREX, also known as "foreign exchange" involves changing one currency into another. Every day, traders trade trillions of dollars [1]. Due to significant currency rate fluctuations, this market is unpredictable, complicated, and subject to frequent changes [2]. The market is always open, although trading takes place in the four main time zones: European, Asian, Australian, and North American [3]. The opening and closing hours of each of these zones differ. The market is protected from scammers since it takes significant money to impact exchange rates. Over the past few decades, scholars have become increasingly interested in forecasting foreign exchange. Unlike the stock market, the foreign exchange market doesn't require large amounts of cash. Leverage is one of the most critical tools related to the market. Leverage is the process of increasing the future return on investment by using borrowed funds[4]. Traders, individuals, professionals making expensive purchases, entrepreneurs, and investors employ leverage. This approach is beneficial for those with little finances and is also a vital element of the FOREX market that attracts private and small investors.

Both technical analysis and fundamental analysis can be used to forecast FOREX prices. While technical analysis only uses historical time series data to make FOREX market predictions, the fundamental analysis considers various variables, including the company's and the nation's economic and industrial conditions [5]. Algorithmic trading refers to trading in which automated programmed algorithms implement orders instead of human traders. Algorithmic trading is utilized by hedge funds, pension funds, and other financial institutions [6], [7]. A lot of work has been put in by both academics and trading companies to find possible factors that could lead to much higher profits[8]. Numerous studies have attempted to forecast the movement of the FOREX market. The most crucial decision in FOREX is predicting the direction of currency price movement. Accurately forecasting currency prices can yield several advantages for traders and vice versa. In recent years, the academic community has made a lot of effort to develop machine learning models for FOREX market prediction.

On the other hand, numerous verifiable study types have been undertaken to understand and anticipate currency patterns in the FOREX market using machine learning algorithms. Generally, many methods are categorized into three categories: machine learning models, deep learning models, and hybrid forms.

Machine learning algorithms include Random Forest, Support Vector Machine, XGBoost, etc. Deep learning-based methods demonstrate how advanced neural models can significantly enhance prediction results. Like statistical methods, these methods require knowledge of effective signals to be utilized as input. The researcher utilized deep learning methods such as RNN, LSTM, CNN, GRU, and Transformers [2]. Long short-term memory (LSTM), a recurrent neural network (RNN), excels at modeling temporal patterns and is commonly employed in various tasks involving time series problems. CNNs are used to analyze price patterns by utilizing images of financial data as input [9]. This study attempts to answer the following questions: What normalization method improves FOREX prediction accuracy? What is the percentage of improvement on different dataset results? Which model, when used with the FOREX normalization function, gives the best results? Therefore, the objectives of this study are to utilize the FNF method with various machine and deep learning models and to compare the proposed model with two baseline models. Then, show the percentage of error reduction made by FNF.

This paper’s primary contributions are summed up as follows:

- FOREX Normalization Function FNF is proposed to predict FOREX prices.FNF generates 84 different normalized features.
- FNF and Convolutional Neural Networks FNF-CNN are used in the first model 1 dimension convolution layer applied on FNF features to predict the close price of the next 1,3,7,15 days.
- FNF and Support Vector Machine FNF-SVR are used in the second model. Four kernels are applied on FNF features to predict the close price of the next 1,3,7,15 days. The forecasts are set for a one-day timeframe. Applying the previous models to 9 different datasets improved the results by an average between 0.5% and 58% of MAE.

The rest of this research is organized in the following manner: Section II reviews the related work, Section III provides an overview of the key scientific concepts, and Section IV describes the proposed models and their architectures, used datasets, evaluation metrics, and training configuration. Section V contains the results of the proposed models compared to baselines, discussion and ablation study. Finally, the paper concludes in Section VI.

II. Related Work

Various methods have been used in previous years to forecast the FOREX market. Numerous approaches have been attempted, mostly based on Artificial Intelligence principles. Some methods contain just one processing technique, while others combine two or more techniques. Researchers have used a variety of linear and nonlinear models for FOREX forecasting. Naive models such as Exponential smoothing, Autoregressive Moving Average model (ARMA), and Autoregressive Conditional Heteroskedasticity models (ARCH) and their variants (GARCH, EGARCH, etc.) are some of the most often used strategies for modeling volatility in time series [10]. Using machine learning methods like artificial neural networks (ANN) has been the subject of extensive study in recent years. The outstanding quality of ANNs used for time series forecasting problems is their innate capacity for nonlinear modeling without any assumption regarding the statistical distribution being invalid based on the observations. The most popular is the multi-layer perceptron (MLP), which has one hidden layer. Support vector machines (SVM), initially designed to address classification issues, are currently used for time series forecasting. Least-square SVM (LS-SVM) and Dynamic Least-square SVM are two common SVM models for forecasting time series [11]. Some researchers have applied deep learning models to predict FOREX prices. Hee Kueh and Leonard have proposed a comprehensive intelligent system for automated FOREX trading. The algorithm utilizes an ensemble methodology to make decisions; each strategy preprocesses technical data in order to produce a distinct buy or sell signal. The ensemble model takes all the signals from the different strategies and uses majority voting logic to decide what to do next [12] .The problem with this system is its low accuracy, and it was tested only on the EUR/USD dataset. Pornwattanavichai and Maneeroj [13] proposed a cascading model for the FOREX market. They made forecasts using Fundamental Data and Technical indications based on BERT. They used the EUR/USD dataset from February 3, 2003, through February 28, 2020, which included 4,455 days. However, this system has limitations; it has not been tested on many datasets. Junior and Appiahene [14] developed a conceptual framework centered on a FOREX forecasting module that uses the Hurst test to determine whether a time series is predictable. They then applied a two-layer stacked LSTM architecture and correlation analysis to multiple currency datasets, including EUR/AUD, AUD/JPY, and AUD/USD. However, the problem with this framework is that it is not generalized to many currency pairs or window times. Dash S. and Sahu [15] employed a Deep Predictive Coding Network Optimized with a Reptile Search Algorithm for short-term forecasting over three days to forecast exchange rates of the CHF/INR, USD/EUR, and AUD/JPY currency pairs. However, this system produces a high mean absolute error value and does not support long-horizon forecasting.

Salman and Saeed U proposed the FLF-LSTM model to predict EUR/USD prices. They enhanced prediction using a custom loss function named FLF with a single LSTM and different activation functions [16]. Areej and Mohamed [17] used RBF, MLP, and SVM algorithms as classifiers to predict the direction of the price and compared them based on percentage classification performance. Ikhagvadorj and Tsendsuren [18] proposed a framework consisting of seven neural networks with different activation functions. The outputs of these neural networks are concatenated and then fed into the softmax layer to produce probabilities or importance weights for each neural network. This model uses extended Min-Max normalization for financial time series data. Haixu and Jiehui [19] used the Autoformer for long-term FOREX price prediction at different time steps (96, 192, 336, 720). Autoformer is a variation of the transformer that uses Auto-Correlation instead of self-attention. Auto-correlation focuses on the connections of sub-series among underlying periods, while self-attention focuses on the connection between time points.

In conclusion, previous research indicates low results, and the impact of preprocessing has not been researched in detail. Also, the datasets used were not diverse and did not contain a large number of values for the time horizon. This paper addresses the impact of data preprocessing and scaling to enhance price prediction by using FNF. Nine datasets are used in this research, and the horizon values are 1, 3, 7, and 15. FNF is used in this paper with different models. The proposed models outperform the baselines.

III. Background

In this section, the necessary context for introducing our method is presented. First, CNN, SVM, LSTM, and XGBoost are examined.

A. CNN

The Convolutional Neural Network (CNN) has demonstrated remarkable advances in several domains associated with pattern recognition and image processing throughout the previous decade. One of the primary advantages

of CNNs is their ability to effectively decrease the parameter count within Artificial Neural Networks (ANNs) [20].

1) Convolutional Neural Network Element: To develop a comprehensive understanding of (CNNs), it is necessary to examine their fundamental components. The input layer receives input and transfers it to the convolution layer[21]. The parameters of a convolutional neural network are arranged into an array of three-dimensional structural units called kernels or filters. Let us assume that the filter's dimensions in the qth layer are Fq × Fq × dq and L_q height of layer q, B_q width of layer q. The following equation defines the convolutional process from the q-th layer to the (q + 1)th layer [22].

$$h_{ijp}^{(q+1)}=\Sigma_{r=1}^{F_q}\Sigma_{s=1}^{F_q}\Sigma_{k=1}^{d_q} w_{rsk}^{(p,q)} h_{i+r-1,j+s-1,k}^{(q)} \quad (1)$$

$$\forall i \in \{1 \dots, L_q - F_q + 1\}$$

$$\forall j \in \{1 \dots, B_q - F_q + 1\}, \forall p \in \{1 \dots, d_{q+1}\}$$

The 3-dimensional tensor $W(p,q) = w_{ijk}^{(p,q)}$ represents the parameters of the pth filter in the qth layer. The indices i, j, and k represent the positions along the filter's height, width, and depth. The qth layer's feature maps are represented by the three-dimensional tensor $H(q) = h_{i,j,k}^{(q)}$ [22]

The pooling layer will then perform downsampling along the provided input's spatial dimension. Researchers use either max pooling or average pooling [23]. The fully-connected layer will perform the same functions observed in conventional artificial neural networks. It is also recommended that the Rectified Linear Unit (ReLU) activation function be employed between these layers to enhance performance [23]. The output layer generates the final prediction value [21]. A common CNN is shown in Fig. 1 [24].

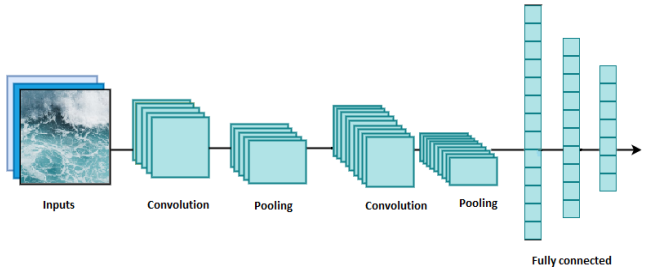


Fig. 1. Convolutional neural network architecture [24].

B. SVM

Support Vector Machines (SVM)can be used for prediction when the outcome is binary, multinomial, or continuous. Classification is a common term for regression with Bernoulli outcomes in statistical learning. Multinomial regression is also known as multiclass classification. The procedure is called regression when the results are continuous [25].

1) Support Vector Regression (SVR): In SVR, the ∈-insensitive loss function is minimized. If the loss is smaller than ∈, then the loss equals zero. The following equation (2) is utilized, known as the simple linear loss function:

$$L_\epsilon = \begin{cases} 0 & \text{, if } |y_i - f(x_i)| < \epsilon \\ |y_i - f(x_i)| - \epsilon, & \text{otherwise} \end{cases} \quad (2)$$

In support vector machines, kernels can achieve nonlinear regressions, such as radial basis function (RBF) and polynomial kernel [26]. The SVR can also be used with a linear kernel. The linear kernel equation is presented in (3), but nonlinear kernels are more flexible.

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij}\, x_{i'j} \quad (3)$$

$i \in \{1, \dots, n\}$ and $i' \in \{1, \dots, n\}$. We will call this inner product K . P is the number of x variables, and j=1,...,p. The variables p and x are mapped into a higher-dimensional space by nonlinear kernels [25], [26].

RBF, a radial basis function, is the most common option for a nonlinear kernel. Equation 4 presents it[25].

$$K(x_i, x_{i'}) = \exp\{-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\} \quad (4)$$

where $\gamma > 0$ is an additional parameter for adaptability. When a test observation is quite far from a training observation, the exponent becomes strongly negative, and $K(xi, xi')$ reaches zero. Sometimes, more variables, such as polynomials, must be added as a function of the original variables. Polynomial variables expand the number of regression variables. The following equation shows the polynomial equation [25] [26].

$$K(x_i, x_{i'}) = (\beta_0 + \gamma \sum_{j=1}^{p} x_{ij}\, x_{i'j})^d \quad (5)$$

Where $\gamma > 0$ and $\beta 0$ are additional parameters for adaptability, $\beta 0$ "biases" the similarity metric for all samples. Applying this kernel implies adding polynomial powers of the x variables [25].

C. LSTM

The Long Short-Term Memory (LSTM) is an architectural design of recurrent neural networks (RNNs) that aims to provide a more precise characterization of temporal sequences and their long-range dependencies in comparison to traditional RNNs [27]. In this section, LSTM architectures will be explored.

1) LSTM architectures: The Long Short-Term Memory (LSTM) model has specialized memory blocks within its recurrent hidden layer. The memory blocks in the network consist of memory cells that possess self-connections, enabling them to retain the temporal state of the network. These memory cells are accompanied by specialized multiplicative units known as gates, which regulate the information flow within the network. In the original architecture, each memory block comprised an input and output gate. The input gate regulates the influx of input activations into the memory cell, while the output gate regulates the transmission of cell activations from the current cell to the remaining components of the network [28] [29]. Subsequently, the forget gate was incorporated into the

memory block. Furthermore, the current Long Short-Term Memory architecture incorporates peephole connections that link the internal cells to the gates inside the same cell. This design allows the LSTM to acquire accurate output timing information [30].

An LSTM network calculates a mapping from an input sequence x = (x_1, ..., x_t) to an output sequence y = (y_1, ..., y_t) by calculating the network unit activations using the following equations iteratively from t = 1 to T:

$$i_t = \sigma(W_{ix}\, x_t + W_{im}\, m_{t-1} + W_{ic} ct\text{-}1 + b_i) \quad (6)$$

$$f_t = \sigma(W_{fx}\, x_t + W_{fm}\, m_{t-1} + W_{fc}\, ct\text{-}1 + b_f) \quad (7)$$

$$c_t = ft \odot ct\text{-}1 + it \odot g(W_{cx}\, x_t + W_{cm}\, m_{t-1} + b_c) \quad (8)$$

$$o_t = \sigma(W_{ox}\, x_t + W_{om}\, m_{t-1} + W_{oc}\, ct + b_o) \quad (9)$$

$$m_t = o_t \odot h(c_t) \quad (10)$$

$$y_t = \varphi(W_{ym}\ m_t + b_y) \quad (11)$$

The weight matrices W terms denote weight matrices, W_{ic} , W_{fc} , W_{oc} are represent diagonal weight matrices corresponding to peephole connections. σ represents the logistic sigmoid function, while b represents bias vectors (bi represents the input gate bias vector). The input gate, forget gate, output gate, and cell activation vectors i, f, o, and c are identical in magnitude to the cell output activation vector m. ⊙ represents the element-wise product of the vectors g and h, where g and h represent the cell input and cell output activation functions, respectively [31].

D. XGBoost

XGBoost, an abbreviation for extreme gradient boosting, is well recognized as a common, robust, and efficient implementation of gradient boosting [32]. XGBoost is an ensemble model that efficiently implements decision trees to create a composite model with superior prediction performance compared to individual techniques employed alone. The output of XGBoost is calculated using the following equation:

$$\hat{Y}_i^T = \sum_{k=1}^{T} f_k(x_i) = \hat{y}_i^{T-1} + f_T(x_i) \quad (12)$$

where $\hat{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{T-1}}$ i is the generated tree, $\boldsymbol{f}_{\mathbf{T}}(\boldsymbol{x}_{\mathbf{i}})$ is the newly created tree model, and T is the total number of tree models [33].

IV. Research Methodology

In this section, two models are proposed: FNF-CNN and FNF-SVR. The first model uses FNF to normalize data and extract new features; those features become input for CNN. The second model also uses FNF and then a Support Vector Machine. Finally, four different kernels are used. The architecture of the models is explained in this section.

A. Proposed Approach

1) Model I: FNF-CNN: Multiple models and preprocessing methods are used in this paper to enhance results and present the impact of FNF. For example, the Moving Average and normalization of the close price are calculated using the following equation:

$$FNFc = MAw(close) - close \quad (13)$$

Here, FNFc represents the normalized close price, and MAw is the Moving Average of window size w. The first proposed model calculates the FNF equation for open, high, low, and close. The time windows range from 1 to 21 days. Then, all moving average features are normalized (FNF) and fed to a 1-dimensional convolutional neural network (CNN). MSE is the loss function, and ReLU is the activation function, as shown in Fig. 2.

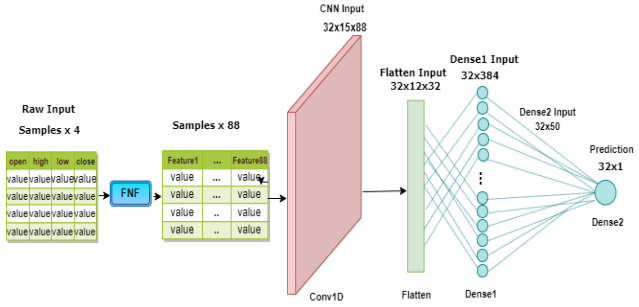


Fig. 2. The architecture of FNF-CNN to predict one day ahead.

A deep learning framework is used with extended Min-Max normalization [18] as the first baseline model. The raw data consists of open, high, low, and close prices from the previous time step, and the target is to predict close prices for the next day. The second baseline model is the Deep Predictive Coding Network Optimized with Reptile Search Algorithm (RSA-DPCN) [15]. RSA-DPCN is used to predict the future price of currency pairs for short-term time frames, such as three, seven, and 15 days ahead of the closing price of EURUSD, AUDJPY, and CHFINR [15]. Fig. 3 shows the proposed model to predict three days ahead of the closing price.

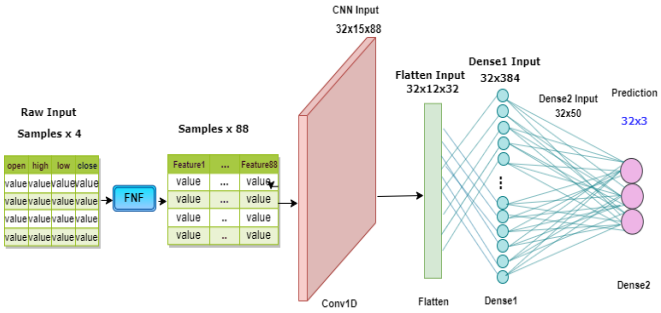


Fig. 3. This model is similar to the previous model in Fig. 2, but it is used to predict the price for the next three days.

2) Model II: FNF-SVR: The Moving Average from Windows 1 to 21 is calculated for all features in the second model. Then, all moving average features are scaled and fed to the Support Vector Regressor (SVR), as shown in Fig. 4. Different kernels are used to obtain the best results.

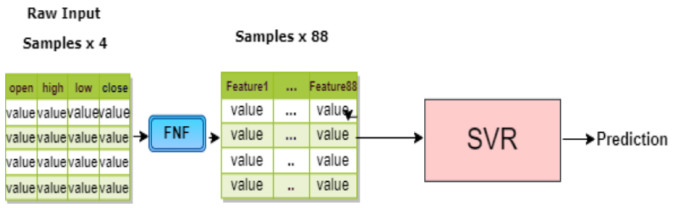


Fig. 4. Architecture of FNF-SVR.

XGBoost and LSTM with FNF are used to compare the results of different machine and deep learning models. For all models, MSE is used as a loss function. TensorFlow 2, Keras, Pandas, and XGBoost frameworks are used in the experiments.

B. Dataset

All models in this paper have been applied to the following datasets: The datasets in group 1 contain daily prices of (GBP/USD, EUR/USD, USD/CHF, USD/JPY, AUD/USD, USD/CAD) from 2000 to 2019. The datasets are partitioned into three parts: training (80%), validation (20%), and testing (the last 365 days)[18]. The dataset group 2 contains daily prices of AUD/JPY, CHF/INR, and EUR/USD from 2015 to 2020 [15].

C. Evaluation Metrics

Mean Absolute Error (MAE) is used to assess the model's performance and is expressed by Eq. (14) [34].

$$MAE = \frac{\sum_{t=1}^{T} |y_t - \hat{y_t}|}{T} \quad (14)$$

Here, y_t denotes the actual value of the price at period t, $\hat{y_t}$ Denotes the forecasted price value at period t, and T denotes the sample size. In other words, the MAE is the mean of the absolute difference between the predicted and actual prices throughout the test set. The actual price and the predicted price differ significantly when the MAE is high [34].

Mean Squared Error (MSE) is another method used to evaluate model performance. MSE is the average squared difference between the actual currency price and the values predicted by the model. Eq. (15) presents MSE [16].

$$MSE = \frac{\sum_{t=1}^{T} (y_t - \hat{y_t})^2}{T} \quad (15)$$

D. Training Configuration

In the first group of datasets, when the FNF-CNN model is used, the time windows range from 1 to 21 days. This range is used because increasing it above 21 leads to increased preprocessing without improving results and reducing the number below 22 reduces the accuracy of the results. The inputs to the first layer are (15x88), with nine previous time steps and 88 features.The number of previous steps was chosen as 9 because increasing it leads to increasing prediction error. The number of filters is 32, the kernel size is 4, and the next layer is flattened these hyperparaemeters chosen after many trails . These parameters were chosen after experimenting with the number of filters: 64 and 128. The experiments showed that using 32 filters leads to better results. Filter sizes of 8 and 16 were also tested, but they did not affect the improvement of the results. The last two layers are dense, and ReLU is used as the activation function..The FNF-LSTM model is a combination of FNF to extract features and LSTM to predict closing prices, and Adam is used as an optimizer. The learning rate and epochs of FNF-CNN and FNF-LSTM are 0.001 and 250, respectively. When we use a large number of epochs, overfitting occurs. Therefore, the appropriate value was equal to 250.The number of layers used is 5 and 3 in FNF-CNN and FNF-LSTM, respectively. The FNF- XGBoost model, which includes FNF and XGBoost Regressor, is used with 80 estimators and max depth=70, and the results look promising. All of these hyperparameters were selected based on many trials.

V. Results and Discussion

A. Results

This section shows the results compared to the results of baseline one and baseline two. FNF-SVR-RBF refers to the FNF-Support Vector Regressor model with a radial basis function. The FNF-SVR-p2 and p3 refer to the FNF-Support Vector Regression model with a polynomial degree 2 and 3 kernels, respectively. Table I shows the MAE results of baseline 1 compared to the proposed models. Table II is the same as Table I but for MSE.

Using FNF generates 84 features, and this variety of features affects results by reducing error.CNN supports efficient feature learning. This resulted in the proposed models performing better than the baseline 1. Additionally, it is worth noting that the SVR model with FNF yielded the best results among all the models in the tables. Fig. 5 shows the price compared to the FNF-SVR, FNF-CNN, and FNF-LSTM prediction for EUR/USD and USD/JPY from dataset group 1. These graphs show that the model has learned the market trend and predicts the prices based on the actual trend.

TABLE I. MAE Results of Baseline1 compared with Proposed Models

Models	EUR/USD	USD/JPY	USD/CHF	GBP/USD	USD/CAD	AUD/USD
Baseline1	0.0043	0.4479	0.0037	0.0062	0.0046	0.0034
FNF-SVR-RBF	0.003141	**0.302782**	**0.002922**	0.005400	0.003571	0.002756
FNF-SVR-p3	0.012265	1.144953	0.191367	0.091945	0.043742	0.005579
FNF-SVR-p2	0.010982	0.654655	0.191753	0.049823	0.020536	0.00724
FNF-SVR- Linear	**0.003112**	0.307247	0.003063	**0.005268**	**0.003515**	**0.002612**
FNF-CNN	0.004465	0.359933	0.013730	0.006086	0.003576	0.004962
FNF-LSTM	0.008481	0.86004	0.070747	0.014158	0.005835	0.004061
FNF-XGBoost	0.005732	0.480439	0.004759	0.105812	0.005034	0.004356

TABLE II. MSE RESULTS OF BASELINE1 COMPARED WITH PROPOSED MODELS

Models	EUR/USD	USD/JPY	USD/CHF	GBP/USD	USD/CAD	AUD/USD
Baseline1	0.000032	0.359160	0.000035	0.000069	0.000036	0.000020
FNF-SVR-RBF	**0.000017**	**0.161011**	**0.000015**	**0.000050**	0.000022	**0.000013**
FNF-SVR-p3	0.000171	1.513236	0.036760	0.008682	0.001948	0.000047
FNF-SVR-p2	0.000137	0.561193	0.036908	0.002587	0.000446	0.000067
FNF-SVR- Linear	**0.000017**	0.166712	0.000016	0.000048	**0.000021**	0.000012
FNF-CNN	0.000028	0.203918	0.000269	0.000061	0.000022	0.000033
FNF-LSTM	0.000083	0.868147	0.005479	0.000249	0.000044	0.000023
FNF-XGBoost	0.000050	0.382606	0.000036	0.012168	0.000041	0.000031

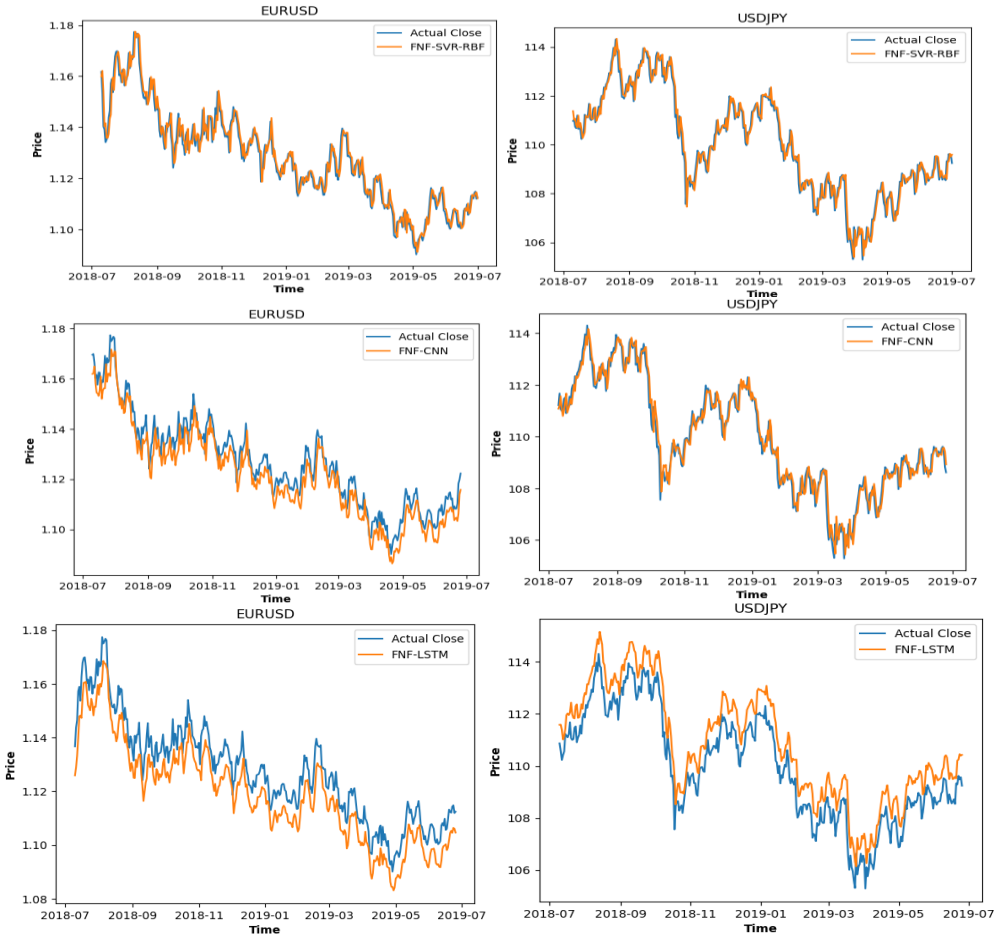


Fig. 5. On the left, the Actual closing price of EUR/USD compared to FNF-SVR-RBF, FNF-CNN, and FNF-LSTM predictions, and the right shows USD/JPY dataset results.

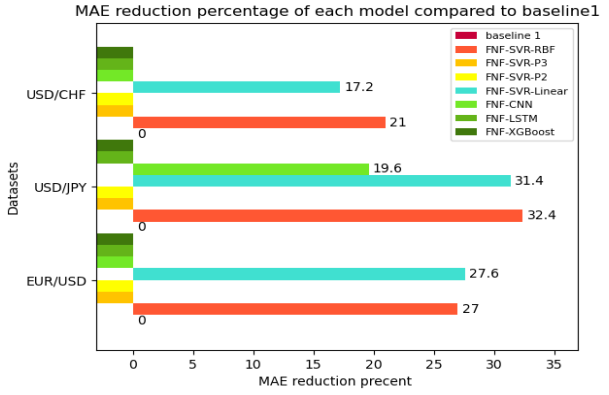


Fig. 6. The MAE reduction percentage of each model compared to baseline 1 applied on USD/CHF, USD/JPY, and EUR/USD datasets

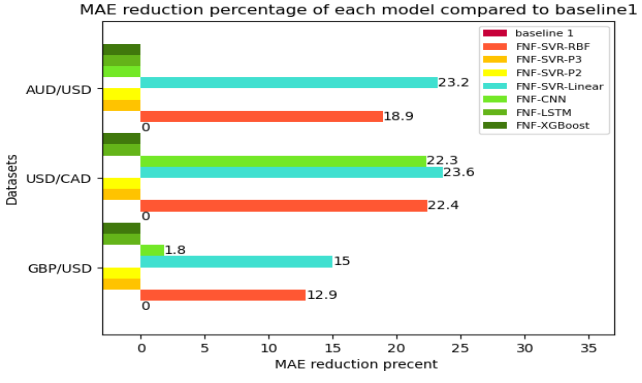


Fig. 7. The percentage reduction in MAE for each model, compared to baseline 1, applied to the AUD/USD, USD/CAD, and GBP/USD datasets.

In the previous results, using FNF with SVR and CNN outperformed baseline 1 in most datasets. Fig. 6 and 7 show the percentage reduction in MAE for each model compared to baseline 1. The preprocessing and feature extraction stages are the main reasons for error reduction.

Table III shows the results for the AUD/JPY dataset of baseline 2 compared to the proposed models. Table IV and Table V are the same as Table III but for the CHF/INR and EUR/USD datasets. The 84 features extracted using FNF also enhance the result of multistep prediction, like 3, 7, and 15 days. From the results, the FNF-SVR and FNF-CNN models outperform baseline two on the AUD/JPY and CHF/INR datasets. Based on CHF/INR and AUD/JPY datasets, we find that the results of the FNF-SVR model are close to the results of the FNF-CNN model.FNF-LSTM also outperforms baseline 2 in most datasets because LSTM learns temporal dependency.

TABLE III. AUD/JPY RESULTS FOR 3,7,15 DAYS AHEAD PREDICTION

Horizon	**3 days**		**7 days**		**15 days**	
Metrics	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
Baseline 2	1.148	1.317904	1.258	1.582564	1.298	1.684804
FNF-SVR-RBF	0.489560	0.425518	0.644829	0.728458	0.902696	1.256283
FNF-SVR-p3	0.498748	0.421315	**0.630987**	**0.680119**	**0.852732**	**1.157992**
FNF-SVR-p2	**0.482397**	0.412050	0.806809	1.105423	0.872990	1.196919
FNF-SVR-Linear	0.539413	0.488757	0.797897	1.045897	1.286681	2.590383
FNF-CNN	0.495194	**0.40988**	0.847844	1.208709	1.267425	2.508938
FNF-LSTM	0.497183	0.426937	0.709538	0.863648	1.038998	1.656788
FNF-XGBoost	0.633611	0.693637	0.806809	1.105423	1.292051	2.972607

TABLE IV. CHF/INR RESULTS FOR 3,7,15 DAYS AHEAD PREDICTION

Horizon	**3 days**		**7 days**		**15 days**	
Metrics	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
Baseline 2	0.5148	0.265019	1.1148	1.242779	1.5148	2.294619
FNF-SVR-RBF	0.491143	0.367525	0.687234	0.748931	1.087740	1.849327
FNF-SVR-p3	1.281250	2.016307	1.136352	1.708272	1.032554	1.522211
FNF-SVR-p2	0.555829	0.451755	0.590223	0.548790	**0.782684**	**1.036713**
FNF-SVR- Linear	0.380511	0.241400	0.740709	0.843808	1.626221	4.194043
FNF-CNN	**0.368026**	**0.227829**	**0.540637**	**0.463773**	1.022167	1.716291
FNF-LSTM	1.122078	1.596323	0.728758	0.810617	1.140256	1.820308
FNF-XGBoost	2.640845	8.427829	3.065497	10.727666	3.330058	12.443699

TABLE V. EUR/USD RESULTS FOR 3,7,15 DAYS AHEAD PREDICTION

Horizon	**3 days**		**7 days**		**15 days**	
Metrics	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
Baseline 2	**0.001**	**0.000001**	**0.0017**	**0.00000289**	**0.0067**	**0.0000449**
FNF-SVR-RBF	0.005288	0.000046	0.007573	0.000100	0.011719	0.000245
FNF-SVR-p3	0.005756	0.000053	0.008000	0.000109	0.012118	0.000257
FNF-SVR-p2	0.473575	0.351067	0.007794	0.000104	0.011822	0.000249
FNF-SVR- Linear	0.005490	0.000049	0.007681	0.000102	0.011551	0.000240
FNF-CNN	0.006116	0.00006	0.009524	0.000142	0.01604	0.000409
FNF-LSTM	0.009236	0.000117	0.02616	0.000764	0.01023	0.000168
FNF-XGBoost	0.006295	0.000068	0.009578	0.000158	0.012472	0.000297

The following section shows the percentage of error reduction by many models compared to baseline2. There is a significant reduction in errors in the AUD/JPY and CHF/INR datasets. The improvement is observed for the 3, 7, and 15 horizons, as shown in Fig. 8. However, using the proposed models on the EUR/USD dataset does not enhance the results because the strength of trend and seasonality in the EUR/USD time series is low.

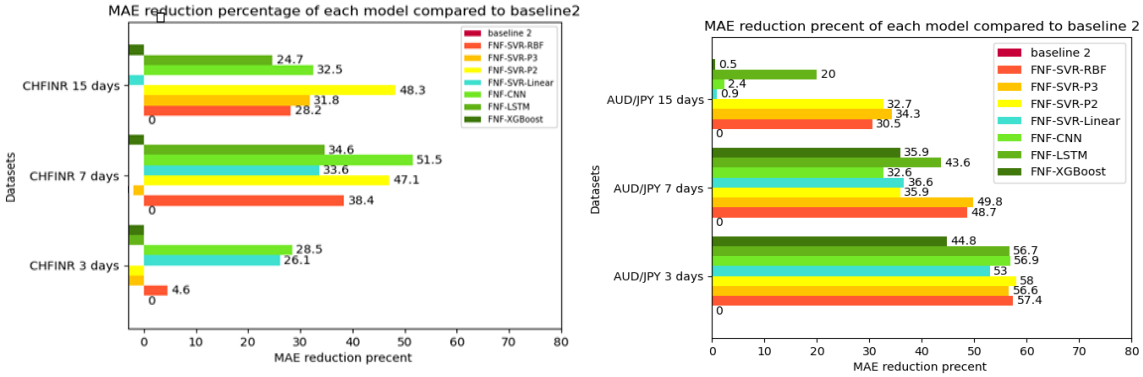


Fig. 8. The MAE reduction percentage of each model, compared to baseline 2, was applied to the AUD/JPY and CHF/INR datasets.

The results appear to vary between different datasets because each dataset has different statistical properties, and the strength of trend and seasonality differs from one time series to another.

B. Discussion and Ablation Study

This section presents an ablation study for the proposed models, FNF-SVR and FNF-CNN, and compares their results with and without FNF. Additionally, this section includes an ablation study applied to the FNF-LSTM model.

1) FNF-CNN: This study examines the impact of using FNF with CNN and the effect of each layer in CNN. The following tables demonstrate the impact of each layer in CNN. In the first baseline, the objective is to predict the next closing price of the next day. The USD/JPY Dataset from group 1, which was used in the first baseline, is shown in Table VI. Table VI illustrates the impact of removing specific layers and FNF on the USD/JPY dataset from group 1. Numbers in the header of the table refer to different model element combinations. The Conv1D refers to the one-dimensional convolution layer, and MaxPooling1D refers to the one-dimensional max pooling layer. The results of the proposed model outperform all other compared components of the models. Tables VII and VIII present the results of USD/CHF and AUD/USD datasets. In most datasets of the first baseline, the FNF-CNN model outperforms other candidate CNN components. If we look at the results of the second column (FNF-CNN) and fifth column (model element combinations 3), which shows the results of the model with FNF and without it, the impact of scaling and feature extraction enhances results in most datasets. Using FNF with the CNN model improved the results by 12.3% and 26.0% in terms of MAE for the datasets USD/JPY and AUD/USD, respectively. These tables also confirm that the layers used in FNF-CNN are the ones that generally produce the best results on different datasets.

TABLE VI. FNF-CNN ABLATION STUDY APPLIED ON USD/JPY FROM GROUP 1 DATASET

Component	FNF-CNN	**1**	**2**	**3**	**4**	**5**
Conv1D	✓	✓	✓	✓	✓	✓
MaxPooling1D		✓	✓		✓	
Flatten	✓	✓	✓	✓	✓	✓
dense1	✓	✓	✓	✓		
dense2	✓	✓	✓	✓	✓	✓
FNF	✓	✓				
MAE	**0.359933**	0.374199	0.611925	0.410345	0.747184	0.459767
MSE	**0.203918**	0.225625	0.512484	0.267927	0.856830	0.307425

TABLE VII. FNF- CNN ABLATION STUDY APPLIED ON USD/CHF FROM GROUP 1 DATASET

Component	FNF-CNN	**1**	**2**	**3**	**4**	**5**
Conv1D	✓	✓	✓	✓	✓	✓
MaxPooling1D		✓	✓		✓	
Flatten	✓	✓	✓	✓	✓	✓
dense1	✓	✓	✓	✓		
dense2	✓	✓	✓	✓	✓	✓
FNF	✓	✓				
MAE	**0.013730**	0.019878	0.023732	0.012414	0.024553	0.036576
MSE	**0.000269**	0.000669	0.000616	0.000206	0.000643	0.001400

TABLE VIII. FNF-CNN ABLATION STUDY APPLIED ON AUD/USD FROM GROUP 1 DATASET

Component	FNF-CNN	**1**	**2**	**3**	**4**	**5**
Conv1D	✓	✓	✓	✓	✓	✓
MaxPooling1D		✓	✓		✓	
Flatten	✓	✓	✓	✓	✓	✓
dense1	✓	✓	✓	✓		
dense2	✓	✓	✓	✓	✓	✓
FNF	✓	✓				
MAE	0.004962	**0.003877**	0.009933	0.006707	0.009564	0.008913
MSE	0.000033	**0.000023**	0.000116	0.000057	0.000106	0.000092

The second baseline target predicts the closing price of 3, 7, and 15 days ahead. Tables IX, X, and XI show the results of the CNN ablation study on the CHF/INR dataset from group 2. The target is to predict the closing price for the next 3, 7, and 15 days. The results show that using FNF improves performance compared to not using it because FNF generates many features that enhance prediction accuracy. Additionally, these tables validate that the FNF-CNN layers are those that typically yield the best results across a variety of horizons.Using FNF with the CNN model improved the results by 51.1%, 16.0%, and 17.0% MAE for the datasets CHF/INR when horizons equal 3, 7, and 15, respectively.

TABLE IX. FNF-CNN ABLATION STUDY APPLIED ON CHF/INR FROM GROUP 2 DATASETS TO PREDICT 3 DAYS

Component	FNF-CNN	**1**	**2**	**3**	**4**	**5**
Conv1D	✓	✓	✓	✓	✓	✓
MaxPooling1D		✓	✓		✓	
Flatten	✓	✓	✓	✓	✓	✓
dense1	✓	✓	✓	✓		
dense2	✓	✓	✓	✓	✓	✓
FNF	✓	✓				
MAE	0.368026	**0.364406**	0.748122	0.753149	0.884571	27.391575
MSE	0.227829	**0.213372**	0.792341	0.791035	1.11417	2195.586932

TABLE X. FNF-CNN ABLATION STUDY APPLIED ON CHF/INR FROM GROUP 2 DATASETS TO PREDICT 7 DAYS

Components	FNF-CNN	**1**	**2**	**3**	**4**	**5**
Conv1D	✓	✓	✓	✓	✓	✓
MaxPooling1D		✓	✓	✓	✓	
Flatten	✓	✓	✓	✓	✓	✓
dense1	✓	✓		✓		
dense2	✓	✓	✓	✓	✓	✓
scaling	✓	✓	✓			
MAE	**0.540637**	0.622345	0.707641	0.643276	0.624963	12.457177
MSE	**0.463773**	0.649072	0.786002	0.677533	0.659157	941.73922

TABLE XI. FNF-CNN ABLATION STUDY APPLIED ON CHF/INR FROM GROUP 2 DATASETS TO PREDICT 15 DAYS

Component	FNF-CNN	**1**	**2**	**3**	**4**	**5**
Conv1D	✓	✓	✓	✓	✓	✓
MaxPooling1D		✓	✓	✓	✓	
Flatten	✓	✓	✓	✓	✓	✓
dense1	✓	✓		✓		
dense2	✓	✓	✓	✓	✓	✓
scaling	✓	✓	✓			
MAE	**0.735953**	1.013976	0.777572	0.887162	0.868603	27.690644
MSE	**0.917616**	1.69364	0.956507	1.195118	1.157751	2204.428353

2) Impact of Kernels and FNF on the SVR model: This section presents the impact of FNF with different SVR kernels. The USD/CHF and AUD/USD datasets from group 1 are used to predict the next day's closing price. The FNF results outperform those without it, as shown in Fig. 9. In the second baseline, the target is to predict the closing price in the next 3, 7, and 15 days; the MultiOutputRegressor from the Keras library was used to do this.

To discuss the previous results in detail, we will explain the percentage of improvement in the results for each time series separately. Applying FNF-SVR on the USD/JPY dataset from group 1 enhances the results by 0.2%, 0.7%, 0.5%, and 2.6% of MAE when using RBF, Poly degree 3, Poly degree 2, and linear kernels, respectively. Using FNF and SVR on the AUD/USD dataset from group 1 enhances the results by 4.2%, 76.1%, and 32.2% of MAE when using RBF, Poly degree 3, and Poly degree 2 kernels, respectively. When we apply the SVR-FNF model on the CHF/INR dataset from group 2 to predict the next three days, the results enhance by 6.7%, 99.7%, and 29.5% of MAE when using RBF, Poly degree 3, and Poly degree 2 kernels, respectively. When we applied the previous experiment on the same dataset but with a horizon equal to 7 days, the results enhanced by 86.7%, 99.7%, and 52.0% of MAE. However, when the horizon is equal to 15 days, the results enhanced by 84.3%, 99.8%, and 34.6%, respectively.

3) FNF-LSTM: In this section, the FNF-LSTM ablation study is presented. The following Datasets, USD/JPY, USD/CHF, and AUD/USD from group 1, are used and displayed in Fig. 10. This figure shows MAE is reduced when FNF is used. Fig. 11 shows the results of the same experiment on the CHF/INR and AUD/JPY datasets from group 2. The results demonstrate that the use of FNF enhances prediction in most datasets.

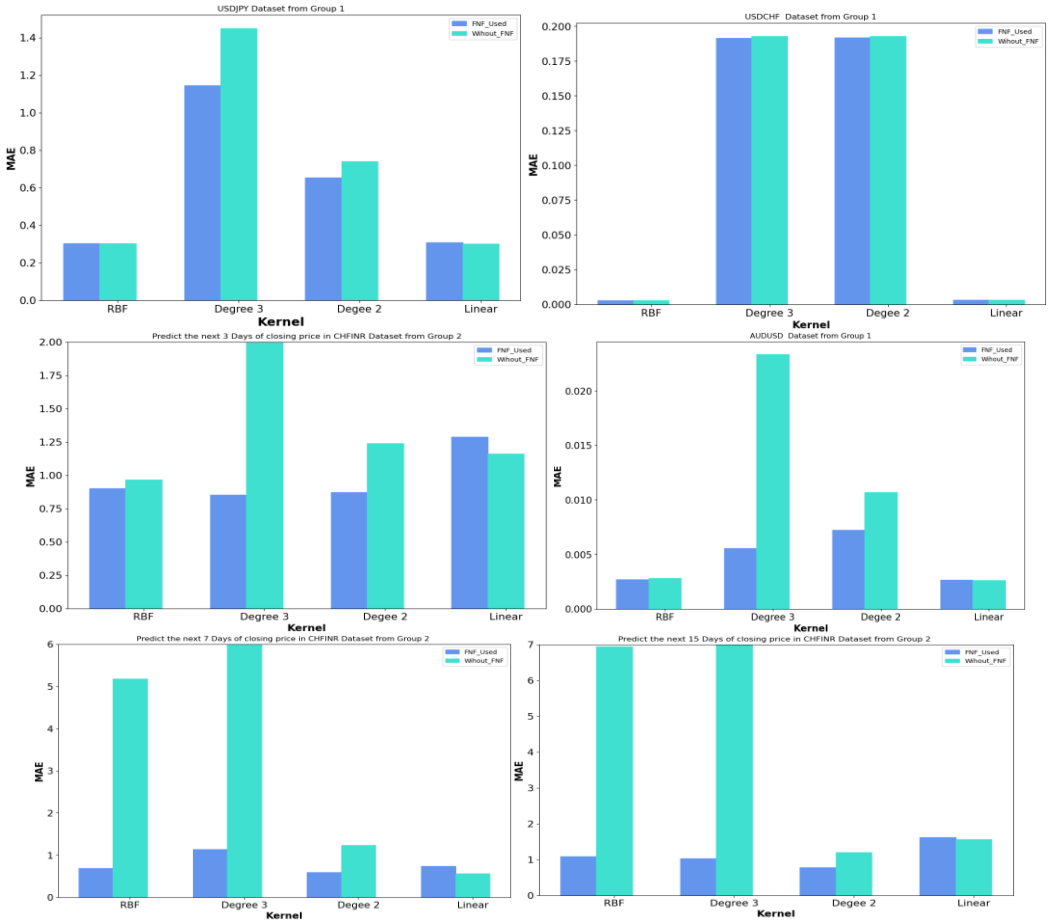


Fig. 9. (a) Impact of FNF and SVR kernels applied on USDJ/PY, USD/CHF, and AUD/USD datasets from group 1 at the top and middle left of this figure. (b) The last three charts show the effect of FNF and Kernels on the CHF/INR dataset from Group 2 to predict the next 3, 7, and 15 days.

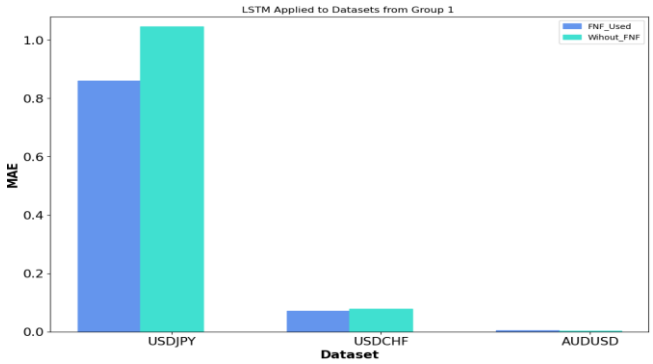


Fig. 10. Impact of FNN with LSTM applied to the USD/JPY, USD/CHF, and AUD/USD datasets from Group 1 to predict the next-day closing price.

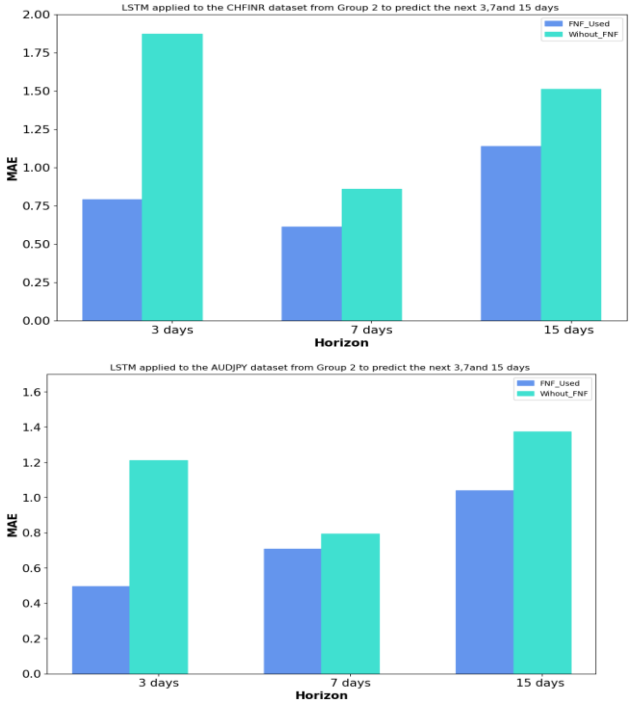


Fig. 11. FNF-LSTM Ablation Study applied to CHF/INR and AUD/JPY Datasets from Group 2 to predict the next 3, 7, and 15 days.

Because every dataset has unique statistical characteristics, and each time series has a variable strength of trend and seasonality, the results appear to differ between them. LSTM can learn temporal dependencies, but when we use FNF-LSTM on the USD/JPY and USD/CHF datasets from group 1, the results are enhanced by 17.8% and 9.5%, respectively, while the results of AUD/USD are not enhanced. Applying FNF-LSTM on the CHF/INR dataset from group 2 to predict the next 3, 7, and 15 days enhances results by 57.8%, 28.7%, and 24.6% of MAE, respectively. Applying FNF-LSTM on the AU/DJPY dataset from group 2 to predict the next 3, 7, and 15 days enhances results by 58.9%, 10.6%, and 24.4% of MAE, respectively.

VI. Conclusion

This paper proposed a FOREX normalization function used as a preprocessing method. This function is used with a machine learning model (SVR) and deep learning model(CNN) to enhance FOREX price prediction. Moving Averages and Scaling on raw data are essential steps to minimize error. Nine FOREX datasets are used in different horizons (1, 3, 7, and 15 days). Mean Absolute Error and Mean Squared Error are used to evaluate all models. The best performance results come from FNF-SVR and FNF-CNN. In this research, we compare different models with FNF and without it. The comparison between the proposed models and the two baseline models shows that our proposed models outperform the baseline models. The development of these proposed models is still in its early stages. Since the models present an exciting and potentially successful research topic, many enhancements must be investigated. The importance of this study is that it reduced the prediction error, which researchers can use this study to build decision support systems used in automated trading. Traders can also use its results to help make decisions to buy and sell currencies.

The limitation of this study is that we did not test the presented models on long-term prediction and did not test them on more datasets. In future work, the same proposed models will be used with different activation functions to find enhanced activation functions for FOREX and train the model using more datasets with varying time frames.

References

[1] M. Ozturk, I. H. Toroslu, and G. Fidan, “Heuristic based trading system on Forex data using technical indicator rules,” Appl Soft Comput, vol. 43, pp. 170–186, 2016.

[2] M. Ayitey Junior, P. Appiahene, O. Appiah, and C. N. Bombie, “Forex market forecasting using machine learning: Systematic Literature Review and meta-analysis,” J Big Data, vol. 10, no. 1, Dec. 2023, doi: 10.1186/s40537-022-00676-2.

[3] S. Masry, A. Dupuis, R. B. Olsen, and E. Tsang, “Time zone normalization of FX seasonality,” Quant Finance, vol. 13, no. 7, pp. 1115–1123, 2013.

[4] B. J. de Almeida, R. F. Neves, and N. Horta, “Combining Support Vector Machine with Genetic Algorithms to optimize investments in Forex markets with high leverage,” Appl Soft Comput, vol. 64, pp. 596–613, 2018.

[5] H. Jamali, Y. Chihab, I. García-Magariño, and O. Bencharef, “Hybrid Forex prediction model using multiple regression, simulated annealing, reinforcement learning and technical analysis,” IAES International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 892–911, Jun. 2023, doi: 10.11591/ijai.v12.i2.pp892-911.

[6] H. H. Zahrah and J. Tirtawangsa, “Algorithmic Forex Trading Using Q-learning,” in IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, 2023, pp. 24–35.

[7] A. Shavandi and M. Khedmati, “A multi-agent deep reinforcement learning framework for algorithmic trading in financial markets,” Expert Syst Appl, vol. 208, p. 118124, 2022.

[8] B. Huang, Y. Huan, L. Da Xu, L. Zheng, and Z. Zou, “Automated trading systems statistical and machine learning methods and hardware implementation: a survey,” Enterprise Information Systems, vol. 13, no. 1. Taylor and Francis Ltd., pp. 132–144, Jan. 02, 2019. doi: 10.1080/17517575.2018.1493145.

[9] A. H. Moghaddam and S. Momtazi, “Image processing meets time series analysis: Predicting Forex profitable technical pattern positions,” Appl Soft Comput, vol. 108, p. 107460, 2021.

[10] S. S. Chung and S. Zhang, “Volatility estimation using support vector machine: Applications to major foreign exchange rates,” Electronic Journal of Applied Statistical Analysis, vol. 10, no. 2, pp. 499–511, 2017.

[11] X.-L. Gong, X.-H. Liu, X. Xiong, and X.-T. Zhuang, “Forecasting stock volatility process using improved least square support vector machine approach,” Soft comput, vol. 23, pp. 11867–11881, 2019.

[12] L. K. Y. Loh, H. K. Kueh, N. J. Parikh, H. Chan, N. J. H. Ho, and M. C. H. Chua, “An Ensembling Architecture Incorporating Machine Learning Models and Genetic Algorithm Optimization for Forex Trading,” FinTech, vol. 1, no. 2, pp. 100–124, Mar. 2022, doi: 10.3390/fintech1020008.

[13] A. Pornwattanavichai, S. Maneeroj, and S. Boonsiri, “BERTFOREX: Cascading Model for Forex Market Forecasting Using Fundamental and Technical Indicator Data Based on BERT,” IEEE Access, vol. 10, pp. 23425–23437, 2022, doi: 10.1109/ACCESS.2022.3152152.

[14] M. Ayitey Junior, P. Appiahene, and O. Appiah, “Forex market forecasting with two-layer stacked Long Short-Term Memory neural network (LSTM) and correlation analysis,” Journal of Electrical Systems and Information Technology, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s43067-022-00054-1.

[15] S. Dash et al., “A Novel Algorithmic Forex Trade and Trend Analysis Framework Based on Deep Predictive Coding Network Optimized with Reptile Search Algorithm,” Axioms, vol. 11, no. 8, Aug. 2022, doi: 10.3390/axioms11080396.

[16] S. Ahmed, S. U. Hassan, N. R. Aljohani, and R. Nawaz, “FLF-LSTM: A novel prediction system using Forex Loss Function,” Applied Soft Computing Journal, vol. 97, Dec. 2020, doi: 10.1016/j.asoc.2020.106780.

[17] A. A. Baasher and M. W. Fakhr, “Forex trend classification using machine learning techniques,” in Proceedings of the 11th WSEAS international conference on Applied computer science, World Scientific and Engineering Academy and Society (WSEAS) Stevens Point …, 2011, pp. 41–47.

[18] L. Munkhdalai, T. Munkhdalai, K. H. Park, H. G. Lee, M. Li, and K. H. Ryu, “Mixture of Activation Functions with Extended Min-Max Normalization for Forex Market Prediction,” IEEE Access, vol. 7, pp. 183680–183691, 2019, doi: 10.1109/ACCESS.2019.2959789.

[19] H. Wu, J. Xu, J. Wang, and M. Long, “Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting,” Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.13008

[20] O. Bayat, S. Aljawarneh, H. F. Carlak, International Association of Researchers, Institute of Electrical and Electronics Engineers, and Akdeniz Üniversitesi, Proceedings of 2017 International Conference on Engineering & Technology (ICET’2017) : Akdeniz University, Antalya, Turkey, 21-23 August, 2017.

[21] R. Tekchandani and N. Kumar, Applied Deep Learning: Design and implement your own Neural Networks to solve real-world problems (English Edition). BPB Publications, 2023.

[22] AGGARWAL, Charu C., et al. Neural networks and deep learning. "Springer", 2018, 10.978: 3.

[23] K. O’Shea and R. Nash, “An Introduction to Convolutional Neural Networks,” Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.08458

[24] A. Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.," 2022.

[25] N. Guenther and M. Schonlau, "Support vector machines," 2016.

[26] wwwit-ebooksinfo, Introduction to Machine Learning with Python. 2016. [Online]. Available: http://safaribooksonline.com

[27] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," Feb. 2014, [Online]. Available: http://arxiv.org/abs/1402.1128

[28] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," Neural Comput, vol. 12, no. 10, pp. 2451–2471, 2000.

[29] D. M. Shprekher, G. I. Babokin, and E. B. Kolesnikov, "Application of neural networks for prediction of insulation condition in networks with isolated neutral," in 2019 International Russian Automation Conference (RusAutoCon), IEEE, 2019, pp. 1–6.

[30] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning Precise Timing with LSTM Recurrent Networks," 2002. [Online]. Available: www.idsia.ch

[31] H. H. Sak, A. Senior, and B. Google, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling."

[32] SERRANO, Luis. Grokking Machine Learning. "Simon and Schuster", 2021.

[33] S. Ben Jabeur, S. Mefteh-Wali, and J. L. Viviani, "Forecasting gold price with the XGBoost algorithm and SHAP interaction values," Ann Oper Res, 2021, doi: 10.1007/s10479-021-04187-w.

[34] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?," Geosci. Model Dev. Discuss, vol. 7, pp. 1525–1534, 2014, doi: 10.5194/gmdd-7-1525-2014.

A New Steganography Method for Hiding Text into RGB Image

AL-Hasan Amer Ibrahim[1], Ruaa Shallal Abbas Anooz[2],
Mohammed Ghassan Abdulkareem[3], Musatafa Abbas Abbood Albadr[4]*,
Fahad Taha AL-Dhief[5]*, Yaqdhan Mahmood Hussein[6], Hatem Oday Hanoosh[7], Mohammed Hasan Mutar[8]
Department of Petroleum Project Management-College of Industrial Management of Oil and Gas,
Basrah University for Oil and Gas, Al-Basrah, Iraq[1, 4]
Technical Engineering College, Al-Furat Al-Awsat Technical University (ATU), Kufa, 54003, Iraq[2]
Department of Oil and Gas Management and Marketing-College of Industrial Management of Oil and Gas,
Basra University for Oil and Gas, Al-Basrah, Iraq[3]
Faculty of Engineering-School of Electrical Engineering, Universiti Teknologi Malaysia (UTM), Johor Bahru 81310, Malaysia[5]
Department of Electronic and Communication-College of Engineering, Al Muthanna University, Iraq[6, 7]
Department of Computer Technical Engineering-College of Information Technology,
Imam Ja'afar Al-Sadiq University, Al-Muthanna 66002, Iraq[8]

***Abstract*—Now-a-days, the network has significant roles in transferring data and knowledge quickly and accurately from sender to receiver. However, the data is still not secure enough to transfer quite confidentially. Data protection is considered as one of the principal challenges in information sharing over communication. So, steganography techniques were proposed which are the art of hiding information that prevents secret text message detection from intruders. Nevertheless, most steganography methods use low bits number of secret messages. Moreover, these methods applied a single logic gate for encrypting the secret message. Therefore, this paper proposes a new method for the encryption of secret messages based on the Huffman technique to reduce the secret message dimensions. In addition, the proposed method uses two different logic gates namely XOR and XNOR for increasing the message security. The RGB Lena image is used as the cover image of the secret message. There are six different experiments conducted with respect to various lengths of the secret messages in bits. The experimental results show that when using the highest number of bits (i.e., 66288), the proposed method achieved 0.0233 MSE, 64.4589 PSNR, 0.9999998 SSIM, and 8.2383 encryption time. The proposed method has the ability to encrypt the secret message with a high number of bits.**

Keywords*—Steganography techniques; color images; XOR gate; NOR gate; huffman technique

I. INTRODUCTION

Nowadays, communication is quite necessary for transmitting information quickly and accurately from the sender to the receiver [1]. Meanwhile, the internet in this modern era provides high convenience in transferring big amounts of data in several parts of the world. Everyone needs safety and secrecy in communicating data [2]. In our daily life, there are many secure pathways that we use such as internet or telephone for sharing and transmitting the information. But unfortunately, these pathways still not safe at a particular level [3]. Consequently, there are two common techniques which are widely used in hiding information and then sharing it safety. These techniques are cryptography and steganography [4, 5]. In the cryptography technique, the message or the text is adjusted in an encrypted form and the encryption key is known only to both sender and receiver. However, the transmission of an encrypted message in such a type of technique may lead to easily excite the attacker's suspicion, and hence this encrypted message will be intercepted, attacked and then decrypted [6]. Therefore, steganography techniques have been proposed and developed in order to overcome the insufficiencies of the cryptographic technique [7].

Steganography is the science of communicating in such a method that it covers and hides the presence of the communication [8]. Thus, there is no one that can detect the existence of a message because the steganography technique hides its presence. On other words, the steganography technique is hiding the message inside multimedia content such as video, audio and image files, where the message will be embedded with one of these multimedia contents [9]. Steganography technique is consisting of two main terms which are the data or the message and cover image [10]. The message is the secret information that needs to be hidden. While the cover image is referred to the carrier that hides or covers the message. Fig. 1 shows the steganography diagram. Steganography's word has been taken from Greek words, where "stegos" indicates to "cover" and "grafia" indicates to "writing" and these two words are defining together as "covered writing" [11]. There are many different techniques of steganography such as spatial domain methods, spread spectrum technique, statistical technique, transform domain technique, and distortion technique [12]. Also, there are different measurements which determine the performance efficiency of steganography techniques such as Peak Signal to Noise Ratio (PSNR), Mean Square Error (MSE), Structural Similarity Index Measure (SSIM) and Signal to Noise Ratio (SNR) [13, 14]. Steganography technique is quite valuable and it can be applied in many domains such as communication and secret data storing, e-commerce, database systems, data alteration protection, and media [15]. As we mentioned previously that the steganography technique has different

categories for hiding information which are embedding the information in text, images, audios, videos, or protocol. The images are considered the most common cover objects which have used for steganography technique [16]. Due to the digital images are broadly proliferated on the Internet and also due to provide a large number of excessive bits in the digital image. The picture steganography is considered as a method for secret and ambiguity correspondence that intends to transfer many of mystery information. Generally, to the cover picture extent among conveying parties. Besides, it aims to avoid the suspicion for non-conveying gatherings in such type of correspondence [17].

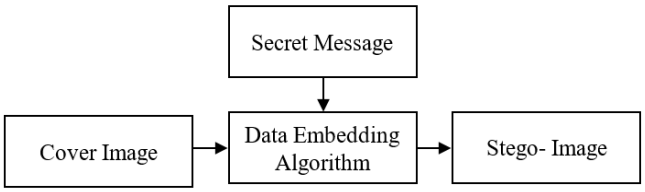


Fig. 1. The steganography diagram.

The image in the computer is a set of numbers that form various light intensities in different image areas. This numeric representation is constituted individual points and a grid which are referred to as pixels [18]. On the Internet, most images are consisting of a rectangular map for the image's pixels which are represented as bits. These pixels are presented horizontally and displayed row by row. The bits number in a color image named the bit depth that belongs to the bits number utilized for every pixel. In present color images, the smallest bit depth is 8 (i.e., there are 8 bits used for representing each pixel color) [19]. Grayscale and monochrome images use 8 bits for every pixel and these pixels are able to present 256 various colors or grey shades.

The digital color images are usually stored in files of 24-bit and use the pattern of RGB color that is known as true color as well. All colors for pixels with 24-bit picture are derived from three essential colors which are Red, Green and Blue (RGB). Each color is produced with 8 bits. Hence, in every pixel, there are 256 different amount of RGB colors as well as to more than 16-million combinations that lead to producing more than 16-million colors in the image [20]. With regards to images, hiding the message in the image is performed by taking the cover object as an image that is indicated as image steganography. The pixel intensities in image steganography are used in order to hide the message [21]. In steganographic images, the Least Significant Bit (LSB) is considered a widely known technique due to its high advantages in encrypting texts in images [22-24]. However, the conventional LSB technique still needs more enhancements in the steganographic technology [25, 26].

Moreover, recent steganography techniques used in the image are yet suffering from the small amount of data that required to be hidden in the cover image. Also, these techniques are showing some negative effects and wasting some of the hidden data. Moreover, there is an urgent need to find a method that is able to hide data without distinguishing between the hidden data and the original covering data. Therefore, the aims of this paper are as follow:

- In this work, we propose the Huffman technique in order to reduce the dimensionality of the secret message.
- Propose two different logic gates called XOR and XNOR for increasing the encryption security of the secret message.
- The proposed method is performed based on six different experiments with respect to various lengths of the secret messages in bits.
- The performance of the proposed method is evaluated in terms of MSE, PSNR, and execution time.

The remainder of this paper is organized as follows: Sections II shows the related works in the steganography technique for hiding information in images. Section III presents the proposed method in terms of the Huffman technique and two logic gates. Section IV discusses the experimental results. Finally, Section V presents the conclusion of this paper.

II. Related Work

Recently, the steganography techniques for hiding text in images have beheld a huge significance by researchers and developers due to the importance such these techniques in terms of hiding data. Furthermore, recent steganography methods of hiding data such as XOR and XNOR logic gates worked on providing a high level of security, where the intended user can only access to the secret data. In other words, the unauthorized user has no ability to detect hidden data, where this is an extremely critical issue in order to protect the sensitivity and confidentiality of information and messages being sent. Here, we will review the up-to-date techniques used in the steganography field. Besides, Table I summarizes the related works used for the encryption of secret messages in images. A steganography technique is proposed in [27] to protect information transported from attackers. This method is worked on the encryption of secret information by using the XNOR gate and the encryption key. The information that required to be encrypted is hidden in a color image by applying Least Significant Bit (LSB) algorithm. Furthermore, this method relies on chromatic channels extraction of 3 RGB channels for every pixel and 2 bits of LSB bits and then determining the channel that will hide the encryption message bit. The second LSB bit is used as an indicator that determines the channel. Meantime, the first LSB bit is replaced with the encrypted message bit, where all the encrypted message bits will be hidden in the cover image. Four different types of images have used as a carrier file which are Airplane, Peppers, Lena, and Baboon. The dimension of all these images is 512x512. Also, different amounts of information were used as secret messages to analyze and evaluate the method, where the smallest amount was 4700 bits and the largest amount was 24250 bits. The performance of this method is evaluated in terms of two measurements which are namely Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE). According to the best-achieved results of the largest amount of secret information are shown that the highest PSNR and lowest MSE are 53.65 and 0.0339 respectively which have obtained from Lina image. However, the amount of hidden information is still not encouraging and a bit small.

The authors in [28] have presented a method to hide text in a color images in order for keeping the information from intruders. This method has two main stages. In the first stage, the steganography technique is used to hide the information in the transmission, where it has used the XOR operation gate for hiding the text bit in the image bit and specifies the position of the image bits according to a random key (the random key has the same length of the secret text). While in the second stage, it has used Triple data encryption standard (3DES) algorithm. This algorithm is used to encrypt the resulting image and the key. Subsequently, the key and image are sent to the receiver. In this method, two types of images have used which are Lena image (256x256 pixels) and Tree image (250x360 pixels). The results of this method have shown that Lena image has 52.0235 PSNR and 0.10024 MSE, while tree image has better performance with 55.0016 PSNR and 0.03602 MSE. However, this method has ignored the length of the confidential text in bit, where it has not taken into account. Furthermore, the steganography technique is applied in [29] to hide the text messages in 24-bit color images. In this method, two schemes have proposed which are embedding data scheme and extraction scheme. In embedding data, the text messages will be hidden in the cover image by using the LSB algorithm. In the extraction scheme, it has used XOR operation to Most Significant Bits (MSB) to recover the secret text from a stego-image. The MSB bits identify the object shape in the image. These two schemes are requiring a key that is created randomly. The experiments have carried out on six different types of images which are Toy (3264 x 2448), Mosque (767x619), Cat (645x533), Flower (551x451), View (32644x2448), and Sunset (1632x1224). The statistical analysis of this method is show that the highest value of PSNR is 63.738 that has obtained by View image, and its MSE value was 0.0026. In this regard, the LSB algorithm is used also in [30] for hiding and protecting secret text message in Image. In this work, the LSB algorithm is combined with XOR encryption techniques in order to increase the security of the text message. Four images have used to embed text messages, these images are Barbara, F16, Lena, and Soccer. All these images have a size of 256x256 pixels and it has used three different sizes (i.e., 1 KB, 2 KB, and 4 KB) of the text messages which are needed to be hidden. The experimental results have shown that the highest achieved PSNR for 1 KB, 2 KB, and 4 KB is 63.5195 (Lena image), 60.3883 (Soccer image), and 57.3182 (Soccer image), respectively. Besides, the lowest MSE for 1 KB, 2 KB, and 4 KB is 0.0289 (Lena image), 0.0595 (Soccer image), and 0.1206 (Soccer image), respectively. Also, the extracted text messages have tested by Character Error Rate (CER) and the value was 0, which proves that text messages have extracted totally. However, the embedded text messages in the cover image are small and limited. A secure model that has embedded text messages for reliable communication was proposed in [31]. Furthermore, this model has used the LSB algorithm for embedding text message bits. In this model, the LSB is based on the secret key and logistic map, this is a spatial domain technique to embed more information in color image without deteriorating the quality of the image. The logistic map method is used for embedding the message bits randomly in the image. The image type used in this work is the Lena image. According to the results, the maximum capacity text messages stored in the cover image is 29127 bytes (233016 bits) and PSNR has been achieved 55.91. A secure method for embedding secret text messages in color images is proposed in [32]. This method is used Integer Wavelet Transform (IWT) technique that is based on the LSB algorithm. The secret text messages are hidden in the LSB algorithm, and the inverse IWT is used to form the stego-image. The secret information is hidden in the approximation coefficient in the components of blue and green colors. While the actual length of the secret data and the sender signature are embedded in the LSB algorithm in the component of the red color of the cover image. In this method, six different types of images have used which are Lena, Baboon, Pepper, Airplane, House, and Tiffany. The size of each image of all these color images is 512×512 pixels. The experimental results are shown that the hybrid IWT-LSB can be embedded secret data with a size of 24 576 bits, where the highest PSNR was 55.5622 that achieved by Baboon image, and the MSE was 0.1807. However, from the studies mentioned above, we can observe some limitations that can be summarized as follow:

- Most models and methods which are used steganography techniques for embedding text messages are yet suffering from a low amount of embedding text messages.
- Majority of these methods are worked on one logic gate in the encryption of secret messages in which results in a low-security level.
- Finally, the execution time of experiments is mostly ignored.

TABLE I. THE SUMMARY OF RELATED WORKS

Years	Techniques	Images	Secret Message Size	PSNR	MSE	Ref.
2020	XNOR and LSB	Airplane, Peppers, Lena, and Baboon	24250 bits	53.65	0.0339	[27]
2020	XOR and 3DES	Lena and Tree	-	55.0016	0.03602	[28]
2019	XOR and MSB	Toy, Mosque, Cat, Flower, View, and Sunset	24 bits	63.738	0.0026	[29]
2019	XOR and LSB	Barbara, F16, Lena, and Soccer	1 KB, 2 KB, and 4 KB	63.5195, 60.3883, and 57.3182	0.0289, 0.0595, and 0.1206	[30]
2018	LSB and logistic map	Lena	29127 bytes	55.91	-	[31]
2018	IWT and LSB	Lena, Baboon, Pepper, Airplane, House, and Tiffany	24 576 bits	55.5622	0.1807	[32]

III. PROPOSED METHOD

A. Encryption Algorithm

In the proposed method, there are three main phases. Fig. 2 shows the steps of these three phases. The first phase includes four steps which are read the secret message, convert the secret message from text to decimal, implement the Huffman to compress the data (reduce the dimensionality), and convert the data from decimal to binary. While in the second phase, the secret message will be encrypted by using the following steps:

- Enter the encryption key value in binary with 8 bits.
- Implement the XNOR logic gate on every other 4 bits (i.e., first 4, third 4, fifth 4 and so on) of the secret message with the first 4 bits of the encryption key.
- Implement the XOR logic gate on every other 4 bits (i.e., second 4, fourth 4, sixth 4 and so on) of the secret message with the second 4 bits of the encryption key.

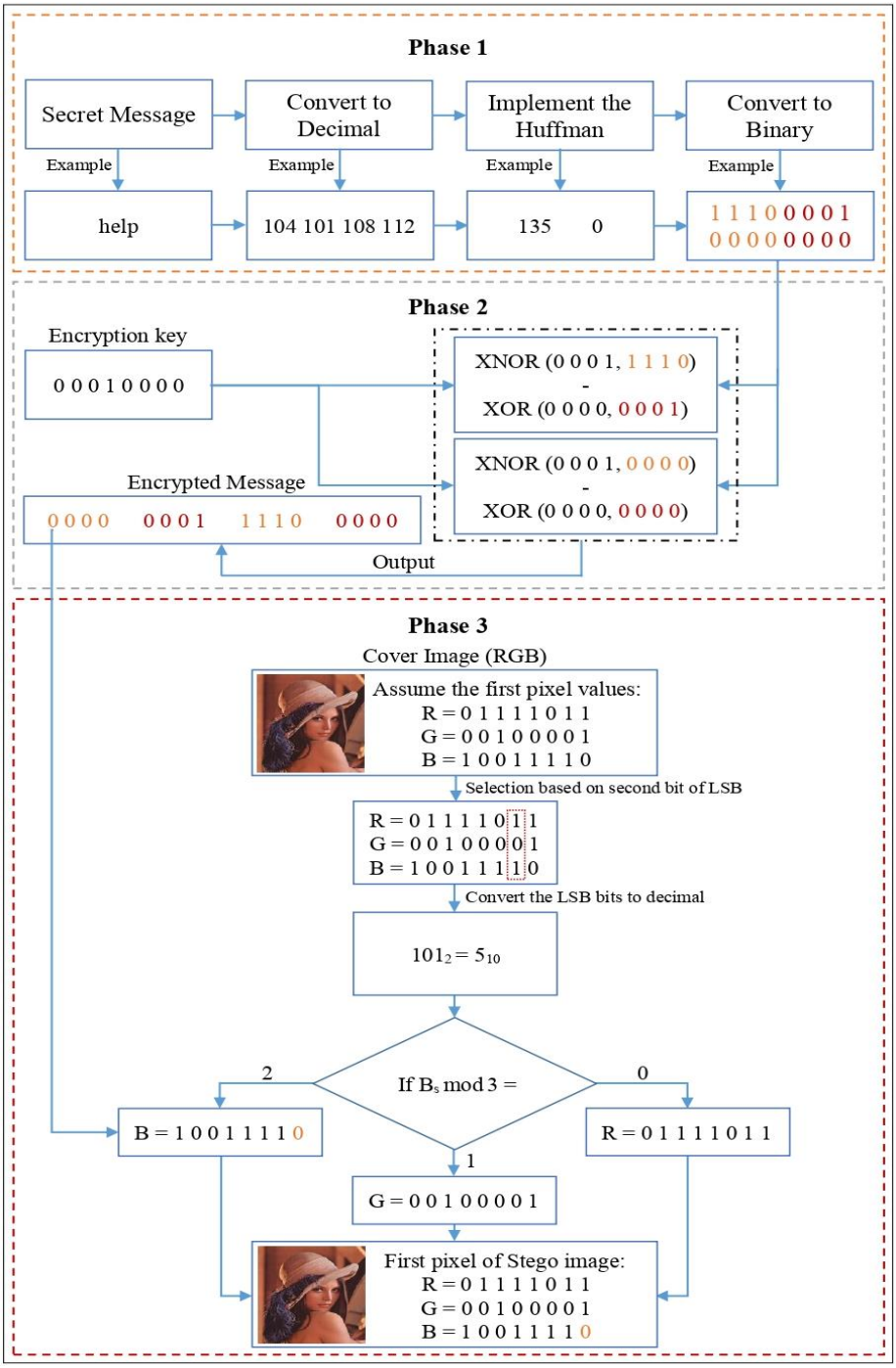


Fig. 2. The three phases of the proposed method.

Furthermore, the third phase includes the following steps:

1. Read the cover image.
2. Selection of one of three RGB channels based on second bit of LSB bits of three channels in decimal and return in B_s.

a. If (B_s mod 3) = 0

Then Red channel is selected.

b. If (Bs mod 3) = 1

Then Green channel is selected.

c. If (Bs mod 3) = 2

Then Blue channel is selected.

3. Apply the LSB algorithm to the selected channel and coding by storing the encrypted bit instead of the first bit of LSB bits.
4. Stego image is achieved.

An example is provided below to encrypt the word "help" as a secret message into the cover image. The steps of this example as follow:

1. Secret message = help.
2. Secret message in decimal = 104 101 108 112.
3. Huffman of the secret message = 135 0.
4. Huffman of the secret message in binary with 8 bits = 1 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0.
5. Encryption key = 0 0 0 1 0 0 0 0.
6. First 4 bits of the Huffman's secret message in binary = 1 1 1 0.
7. Second 4 bits of the Huffman's secret message in binary = 0 0 0 1.
8. Third 4 bits of the Huffman's secret message in binary = 0 0 0 0.
9. Fourth 4 bits of the Huffman's secret message in binary = 0 0 0 0.
10. First 4 bits of the encryption key = 0 0 0 1.
11. Last 4 bits of the encryption key = 0 0 0 0.
12. The first 4 bits of the encrypted message = XNOR (Step 6, Step 10) = 0 0 0 0.
13. The second 4 bits of the encrypted message = XOR (Step 7, Step 11) = 0 0 0 1.
14. The third 4 bits of the encrypted message = XNOR (Step 8, Step 10) = 1 1 1 0.
15. The fourth 4 bits of the encrypted message = XOR (Step 9, Step 11) = 0 0 0 0.
16. The encrypted message bits = 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0.
17. Read the cover image. Table II shows the selection and encryption processes to create four pixels of the stego image.

TABLE II. THE SELECTION AND ENCRYPTION PROCESSES OF FOUR PIXELS IN THE STEGO IMAGE

Pixel	Channel	Value of the channels in binary (cover image)	Second LSB of the three channels	Selected channel	Message bit	Value of the channels in binary (stego image)
1st	R	01111011	$101_2 = 5_{10}$	2 B	0	01111011
	G	00100001				00100001
	B	10011110				10011110
2nd	R	01111011	$101_2 = 5_{10}$	2 B	0	01111011
	G	10100001				10100001
	B	10011111				10011110
3rd	R	01000111	$100_2 = 1_{10}$	1 G	0	01000111
	G	10010001				10010000
	B	10000001				10000001
4th	R	01111011	$110_2 = 3_{10}$	0 R	0	01111010
	G	00111110				00111110
	B	10001100				10001100

B. Decryption Algorithm

This section will present the processes for extracting the secret message from stego image. These processes are summarized below. In addition, Table III shows the determination of RGB channel based on second bit of LSB.

1. Get stego image.
2. Selection of one of three RGB channels based on second bit of LSB bits of three channels in decimal and return in Bs.
   a. If (B_s mod 3) = 0

   Then Red channel is selected.

   b. If (Bs mod 3) = 1

   Then Green channel is selected.

   c. If (Bs mod 3) = 2

   Then Blue channel is selected.
3. Apply the LSB algorithm on the selected channel to get the encrypted message bits (0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0).
4. Input the encryption key = (0 0 0 1 0 0 0 0) and apply the following steps:
   a. XNOR decoding on the first 4 bits of the encrypted message and the first 4 bits of the encryption key. Therefore, XNOR (0 0 0 0, 0 0 0 1) = 1 1 1 0.
   b. XOR decoding on the second 4 bits of the encrypted message and the last 4 bits of the encryption key. Therefore, XOR (0 0 0 1, 0 0 0 0) = 0 0 0 1.
   c. XNOR decoding on the third 4 bits of the encrypted message and the first 4 bits of the encryption key. Therefore, XNOR (1 1 1 0, 0 0 0 1) = 0 0 0 0.
   d. XOR decoding on the fourth 4 bits of the encrypted message and the last 4 bits of the encryption key. Therefore, XOR (0 0 0 0, 0 0 0 0) = 0 0 0 0.
5. The Huffman's secret message in binary is obtained = 1 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0.
6. Convert each 8 bits of the Huffman's secret message to decimal = 135 0.
7. Apply the Huffman to normal which decodes the Huffman's secret message in decimal to get the normal secret message in decimal = 104 101 108 112.
8. Convert the secret message from decimal to character to get the secret message = help.

TABLE III. THE DETERMINATION OF RGB CHANNEL BASED ON SECOND BIT OF LSB

Pixel	Channel	Value of the channels in binary (stego image)	Second LSB of the three channels	Selected channel	Message bit
1st	R	01111011	$101_2 = 5_{10}$	2 B	0
	G	00100001			
	B	10011110			
2nd	R	01111011	$101_2 = 5_{10}$	2 B	0
	G	10100001			
	B	10011110			
3rd	R	01000111	$100_2 = 1_{10}$	1 G	0
	G	10010000			
	B	10000001			
4th	R	01111010	$110_2 = 3_{10}$	0 R	0
	G	00111110			
	B	10001100			

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this study, the Lena color image (RGB) with (512 × 512 × 3) dimensionality was used in order to conduct six different experiments. The six different experiments were implemented based on various lengths of the secret message in bits (i.e., 3640, 6872, 26224, 33128, 40032, and 66288 bits). In this proposed method, we have used Huffman method in order to reduce the dimensionality of the data (i.e., secret message). Furthermore, there are two logic gates which are XNOR and XOR have been used for the purpose of encrypting the Huffman of the secret message with the encryption key. It is worth mention that all experiments have been implemented in MATLAB R2019a programming language over a PC Core i7 of 3.20 GHz with 16 GB RAM and SSD 1 TB (Windows 10).

The proposed method has been evaluated in terms of the most common performance measurements used in the steganography techniques which are Mean Square Error (MSE), Structural Similarity Index Measure (SSIM), and Peak Signal-to-Noise Ratio (PSNR). Besides, the proposed method has been evaluated in terms of the execution time. The MSE, PSNR and the execution time can be calculated as the following equations:

$$MSE = \sum_{m=0}^{M-1} \sum_{f=0}^{F-1} \|A(m,f) - T(m,f)\|^2 \quad (1)$$

$$PSNR = 10log\,10\left(\frac{255^2}{MSE}\right) \quad (2)$$

$$Execution\ Time\ (T) = T_{End} - T_{Start} \quad (3)$$

where, $A(m, f)$ refers to the pixel value of the cover image, $T(m, f)$ refers to the pixel value of the stego image, and M and F refer to the height and width of the images, respectively. In the six different experiments, the minimum and maximum numbers of the secret message bits are 3640 and 66288, respectively. Based on the experiments results, when the minimum number of the secret message bits is used, the proposed method is achieved 0.0013 MSE, 77.1531 PSNR, and it has taken 0.9543 sec for the execution time. Meanwhile, the proposed method has been obtained 0.0140 MSE, 64.4589 PSNR, and 8.2383 sec by using the maximum number of the secret message bits.

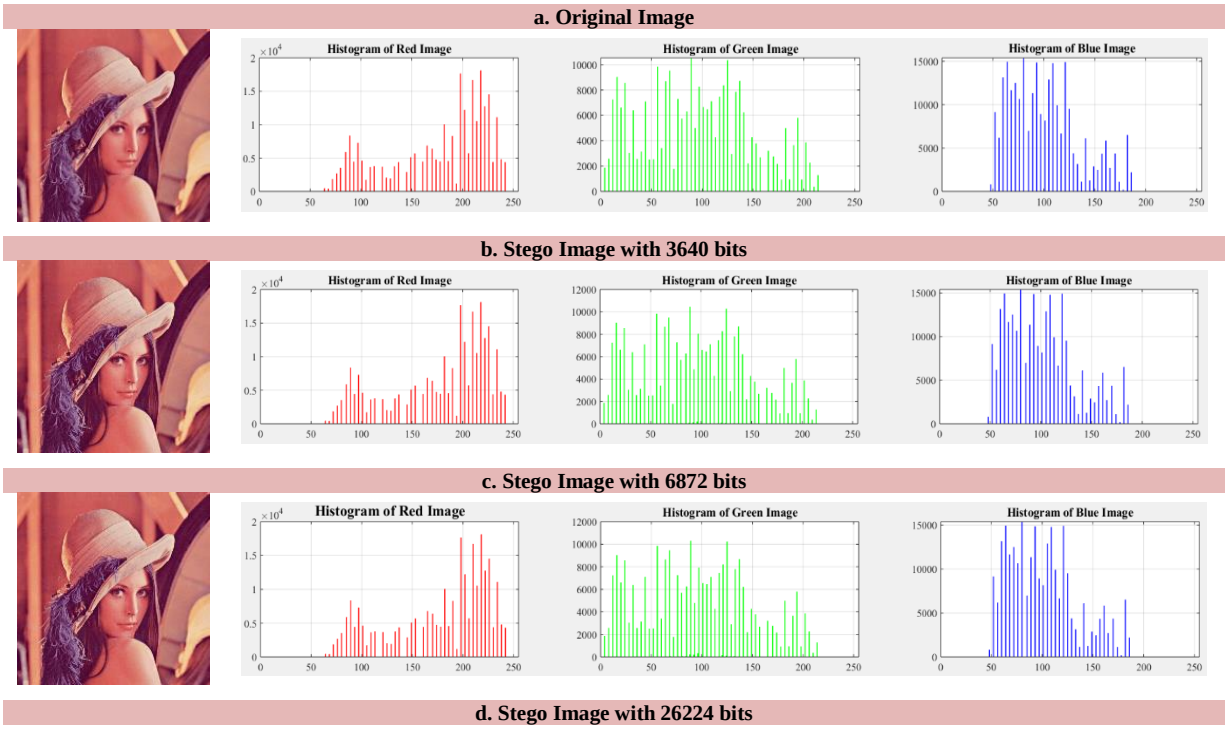
Furthermore, Table IV illustrates the full results of the six different experiments in terms of MSE, PSNR, SSIM, and encryption time in seconds. Fig. 3 shows the histograms of the three colors (red, green, and blue) separately for the original image (cover image) and the stgo images with different lengths of secret messages in bits. Moreover, Fig. 4 demonstrates all the color histograms (red, green, and blue) for the original image (cover image) and the stgo images with different lengths of secret messages in bits. Based on Fig. 3 and Fig. 4, the proposed method has not been influenced the three channels of the original image colors (red, green, and blue). In other words, the proposed method has the ability to encrypt high numbers of secret message bits without affecting the quality of the image.

Moreover, the proposed method has been compared with other methods using Lena image [30, 33-37] as shown in Table V. The experimental results showed that the proposed method has been outperformed the other methods in terms of number of bits, MSE, SSIM and PSNR. Although the proposed method has been shown the high ability to encrypt a high number of bits in RGB image, there are some limitations in the proposed method which can be summarized as follow:

- All experiments of the proposed method have been conducted using one image only. In other words, the proposed method is evaluated using the Lena image only.
- The proposed method has been evaluated based on the standard image size (512 × 512 × 3). Whilst other evaluations based on varying the image sizes may lead to obtaining different results.

TABLE IV. THE RESULTS OF THE SIX DIFFERENT EXPERIMENTS

Image	Number of bits	MSE	PSNR	SSIM	Encryption Time (s)
Lena (512 × 512 × 3)	3640	0.0013	77.1531	0.9999998	0.9543
	6872	0.0023	74.4227	0.9999996	1.0754
	26224	0.0091	68.5274	0.9999983	3.2982
	33128	0.0116	67.480	0.9999976	4.4995
	40032	0.0140	66.6573	0.9999969	4.8424
	66288	0.0233	64.4589	0.9999934	8.2383



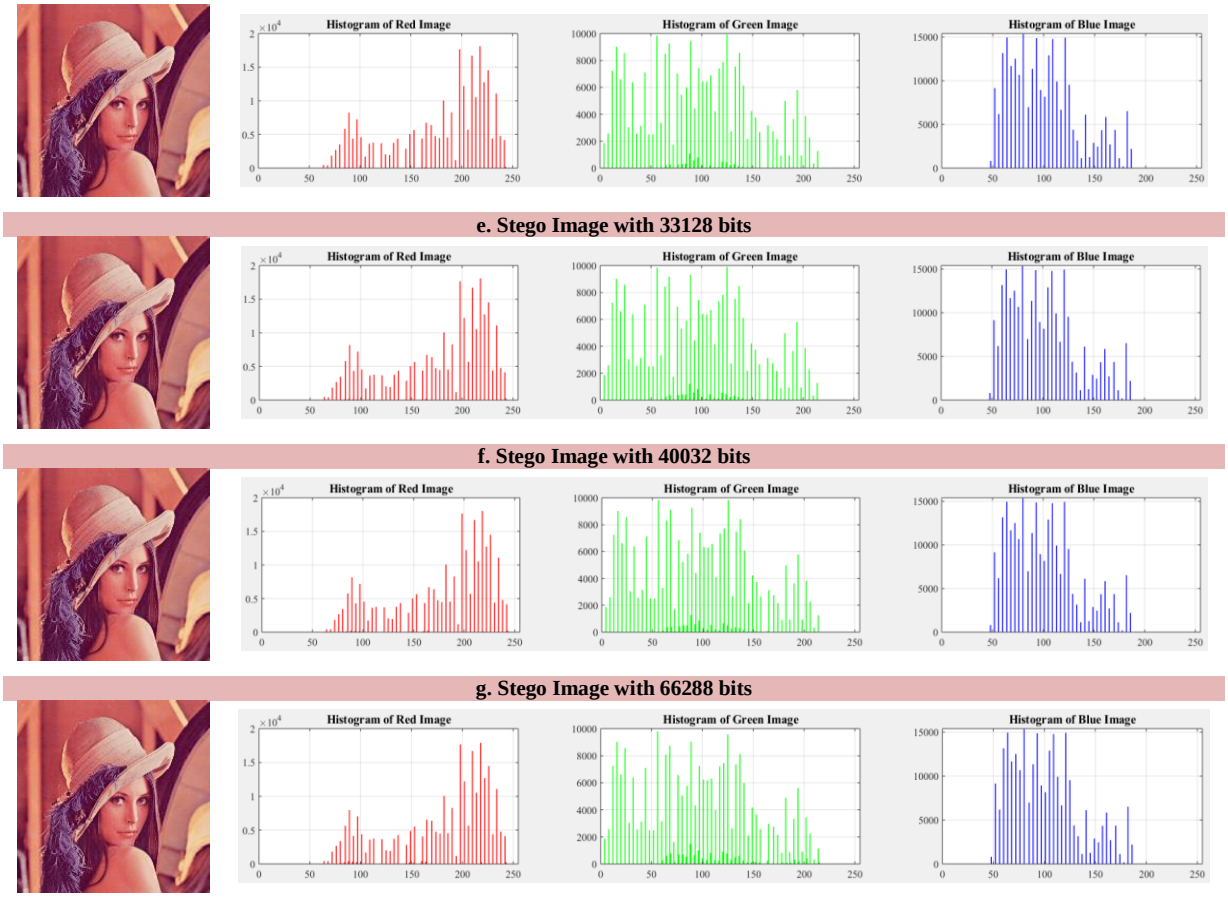
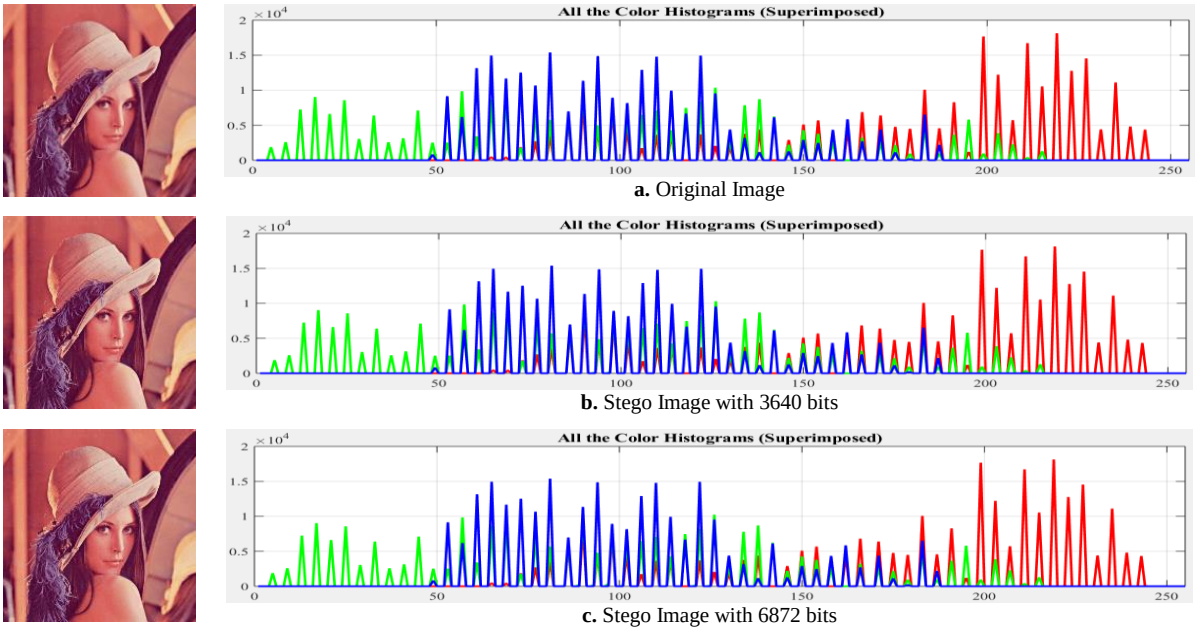


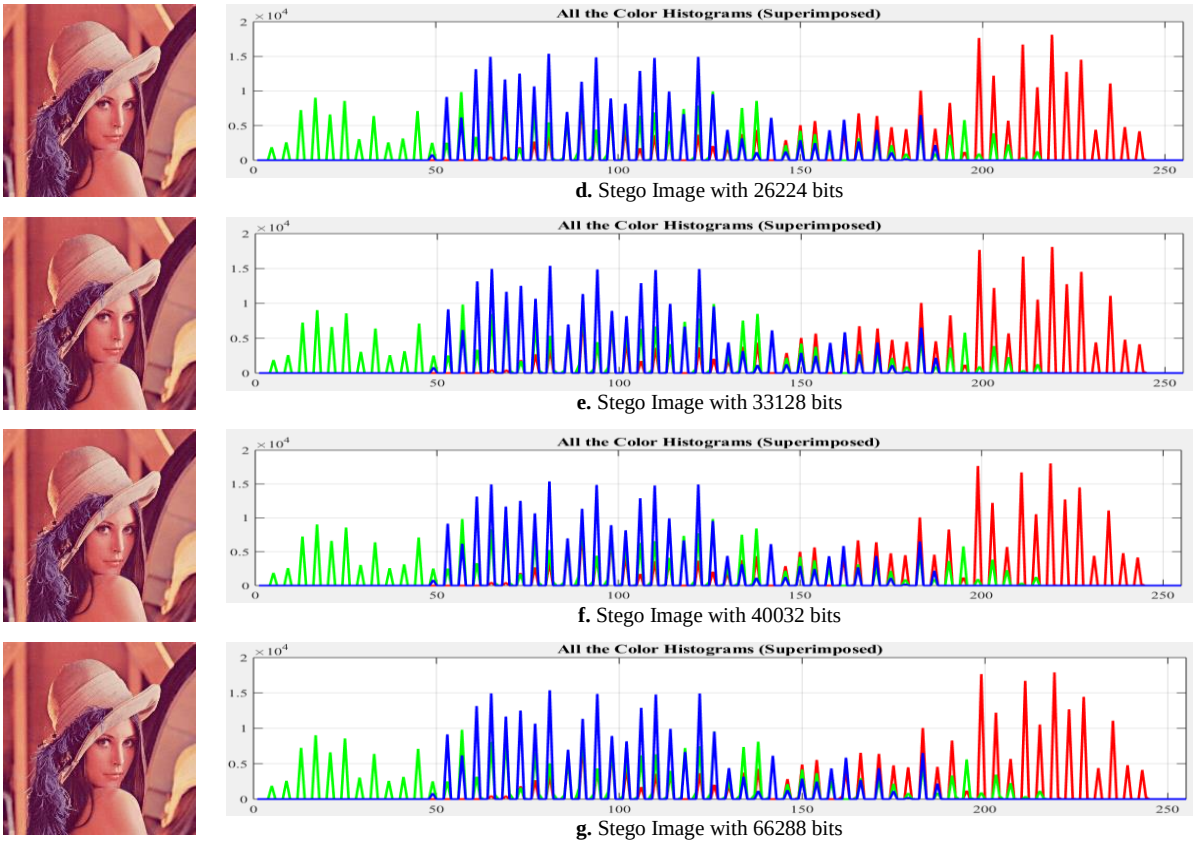
Fig. 3. The separated histograms of three colors for the original image and stego images with different lengths of secret messages bits.



a. Original Image

b. Stego Image with 3640 bits

c. Stego Image with 6872 bits



d. Stego Image with 26224 bits

e. Stego Image with 33128 bits

f. Stego Image with 40032 bits

g. Stego Image with 66288 bits

Fig. 4. All the color histograms for the original image and stego images with different lengths of secret messages bits.

TABLE V. COMPARISON RESULTS BETWEEN METHODS USING LENA IMAGE

Method	Number of bits	MSE	SSIM	PSNR
[33]	24500	-	0.997242	52.9817
[34]	960	5.73706	0.99374	40.54391
[35]	22248	0.042486	0.999919	57.079979
[36]	1200	0.0038	-	72.242
[37]	14357	-	-	41.72
[30]	32000	0.1290	-	57.0260
Proposed Method	**3640**	**0.0013**	**0.9999998**	**77.1531**
	66288	**0.0233**	**0.9999934**	**64.4589**

V. CONCLUSION

In this paper, we have presented a new method in the encryption of secret messages in the RGB image. The proposed method is used the Huffman technique in order to reduce the secret message dimensionality. In addition, there are two logic gates (i.e., XNOR and XOR) have been used in order to encrypt the Huffman of the secret message with the encryption key. The proposed method has been implemented based on six different experiments with respect to various lengths of the secret messages in bits (i.e., 3640, 6872, 26224, 33128, 40032, and 66288 bits). All experiments have been performed using the RGB Lena image with the size of (512 × 512 × 3). The experimental results are showed that the proposed method has been achieved 0.0013 MSE, 0.9999998 SSIM, 77.1531 PSNR, and it has taken 0.9543 sec when the number of the secret message was 3640 bits. Meanwhile, the proposed method achieved 0.0233 MSE, 0.9999934 SSIM, 64.4589 PSNR, and 8.2383 sec when the number of the secret message was 66288 bits. Based on the results, the proposed method is able to encrypt the secret message with a high number of bits efficiently. Future work can include using different RGB images such as Airplane, Peppers, and Baboon with varying sizes and other measurements such as Bits Per Pixel (BPP).

REFERENCES

[1] T. Limbasiya and D. Das, "Lightweight secure message broadcasting protocol for vehicle-to-vehicle communication," IEEE Systems Journal, vol. 14, no. 1, pp. 520-529, 2019.

[2] M. Elhoseny et al., "Security and privacy issues in medical internet of things: overview, countermeasures, challenges and future directions," Sustainability, vol. 13, no. 21, p. 11645, 2021.

[3] T. Naqash, A. Iqbal, and S. H. Shah, "Review on Safe Reversible Image Data Hiding," in 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), 2019: IEEE, pp. 0929-0932.

[4] B. Swathi, K. Shalini, and K. N. Prasanthi, "A review on steganography using images," Asian Journal of Computer Science and Information Technology, vol. 2, no. 8, 2012.

[5] M. K. I. Rahmani, K. Arora, and N. Pal, "A crypto-steganography: A survey," International Journal of Advanced computer science and applications, vol. 5, no. 7, 2014.

[6] M. Mittal, S. Gupta, P. K. Keserwani, and M. C. Govil, "Security Enhancement using Vectoring, Cryptography and Steganography," in 2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC), 2023: IEEE, pp. 1-8.

[7] J. Kour and D. Verma, "Steganography techniques–A review paper," International Journal of Emerging Research in Management & Technology, vol. 3, no. 5, pp. 132-135, 2014.

[8] F. Sharmin and M. I. Khan, "Image steganography using combined nearest and farthest neighbors methods," International Journal of Advanced Computer Science and Applications, vol. 10, no. 11, 2019.

[9] N. Singh, "Survey paper on steganography," International Refereed Journal of Engineering and Science (IRJES), vol. 6, no. 1, pp. 68-71, 2017.

[10] M. K. Abed, M. M. Kareem, R. K. Ibrahim, M. M. Hashim, S. Kurnaz, and A. H. Ali, "Secure medical image steganography method based on pixels variance value and eight neighbors," in 2021 International Conference on Advanced Computer Applications (ACA), 2021: IEEE, pp. 199-205.

[11] P. W. Wong and N. Memon, "Secret and public key image watermarking schemes for image authentication and ownership verification," IEEE transactions on image processing, vol. 10, no. 10, pp. 1593-1601, 2001.

[12] F. Q. A. Alyousuf, R. Din, and A. J. Qasim, "Analysis review on spatial and transform domain technique in digital steganography," Bulletin of Electrical Engineering and Informatics, vol. 9, no. 2, pp. 573-581, 2020.

[13] A. O. Modupe, A. E. Adedoyin, and A. O. Titilayo, "A Comparative Analysis of LSB, MSB and PVD Based Image Steganography," Int. J. Res. Rev, vol. 8, no. 9, pp. 373-377, 2021.

[14] S. Rustad, P. N. Andono, and G. F. Shidik, "Digital image steganography survey and investigation (goal, assessment, method, development, and dataset)," Signal Processing, p. 108908, 2022.

[15] S. Rahman et al., "A Comprehensive Study of Digital Image Steganographic Techniques," IEEE Access, vol. 11, pp. 6770-6791, 2023.

[16] R. A. Watheq, F. Almasalha, and M. H. Qutqut, "A new steganography technique using JPEG images," International Journal of Advanced Computer Science and Applications, vol. 9, no. 11, 2018.

[17] V. Lakshminarayanan and I. Bhattacharya, "Advances in Optical Science and Engineering," Springer Proceedings in Physics, vol. 166, pp. 533-539, 2014.

[18] G. Paul, I. Davidson, I. Mukherjee, and S. Ravi, "Keyless dynamic optimal multi-bit image steganography using energetic pixels," Multimedia tools and applications, vol. 76, no. 5, pp. 7445-7471, 2017.

[19] O. I. I. Al-Farraji, "New technique of steganography based on locations of LSB," International Journal of Information Research and Review, vol. 4, no. 01, pp. 3549-3553, 2017.

[20] M. M. Hashim, M. S. M. Rahim, F. A. Johi, M. S. Taha, and H. S. Hamad, "Performance evaluation measurement of image steganography techniques with analysis of LSB based on variation image formats," International Journal of Engineering & Technology, vol. 7, no. 4, pp. 3505-3514, 2018.

[21] H. L. Hussein, A. Abbass, S. Naji, S. Alaugby, and J. Lafta, "Hiding text in gray image using mapping technique," in Journal of Physics: Conference Series, 2018, vol. 1003, p. 012032.

[22] Y. P. Astuti, E. H. Rachmawanto, and C. A. Sari, "Simple and secure image steganography using LSB and triple XOR operation on MSB," in 2018 International Conference on Information and Communications Technology (ICOIACT), 2018: IEEE, pp. 191-195.

[23] D. R. I. M. Setiadi, S. Rustad, P. N. Andono, and G. F. Shidik, "Graded fuzzy edge detection for imperceptibility optimization of image steganography," The Imaging Science Journal, pp. 1-13, 2023.

[24] H. Hiary, K. E. Sabri, M. S. Mohammed, and A. Al-Dhamari, "A hybrid steganography system based on LSB matching and replacement," International Journal of Advanced Computer Science and Applications, vol. 7, no. 9, 2016.

[25] C. A. Sari, G. Ardiansyah, and E. H. Rachmawanto, "An improved security and message capacity using AES and Huffman coding on image steganography," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 17, no. 5, pp. 2400-2409, 2019.

[26] S. Ghoul, R. Sulaiman, and Z. Shukur, "A Review on Security Techniques in Image Steganography," International Journal of Advanced Computer Science and Applications, vol. 14, no. 6, 2023.

[27] R. M. Neamah, J. A. Abed, and E. A. Abbood, "Hide text depending on the three channels of pixels in color images using the modified LSB algorithm," International Journal of Electrical and Computer Engineering, vol. 10, no. 1, p. 809, 2020.

[28] H. R. Kareem, H. H. Madhi, and K. A.-A. Mutlaq, "Hiding encrypted text in image steganography," Periodicals of Engineering and Natural Sciences, vol. 8, no. 2, pp. 703-707, 2020.

[29] D. Ratnasari and A. S. Aji, "Text to Color Image Steganography Using LSB Technique and XOR Operations," International Journal of Applied Business and Information Systems, vol. 3, no. 2, pp. 59-65, 2019.

[30] A. Setyono, "Securing and hiding secret message in image using xor transposition encryption and LSB method," in Journal of Physics: Conference Series, 2019, vol. 1196, no. 1: IOP Publishing, p. 012039.

[31] M. Ulker and B. Arslan, "A novel secure model: Image steganography with logistic map and secret key," in 2018 6th International Symposium on Digital Forensic and Security (ISDFS), 2018: IEEE, pp. 1-5.

[32] E. Emad, A. Safey, A. Refaat, Z. Osama, E. Sayed, and E. Mohamed, "A secure image steganography algorithm based on least significant bit and integer wavelet transform," Journal of Systems Engineering and Electronics, vol. 29, no. 3, pp. 639-649, 2018.

[33] E. A. Abbood, R. M. Neamah, and S. Abdulkadhm, "Text in Image Hiding using Developed LSB and Random Method," International Journal of Electrical & Computer Engineering (2088-8708), vol. 8, no. 4, 2018.

[34] S. N. Abd-Alwahab, M. K. WAli, and M. F. Bonneya, "Text Hiding in Coded Image Based on Quantization Level Modification and Chaotic Function," International Journal of Integrated Engineering, vol. 13, no. 1, pp. 148-158, 2021.

[35] Ö. Çataltaş and K. Tütüncü, "Comparison of LSB image steganography technique in different color spaces," in 2017 international artificial intelligence and data processing symposium (IDAP), 2017: IEEE, pp. 1-6.

[36] S. A. Mahdi and A. K. Maisa'a, "An Improved Method for Combine (LSB and MSB) Based on Color Image RGB," Engineering and Technology Journal, vol. 39, no. 1B, pp. 231-242, 2021.

[37] S. N. Mali, P. M. Patil, and R. M. Jalnekar, "Robust and secured image-adaptive data hiding," Digital Signal Processing, vol. 22, no. 2, pp. 314-323, 2012.

Bidirectional Long Short-Term Memory for Analysis of Public Opinion Sentiment on Government Policy During the COVID-19 Pandemic

Intan Nurma Yulita[1*], Ahmad Faaiz Al-Auza'i[2], Anton Satria Prabuwono[3], Asep Sholahuddin[4], Firman Ardiansyah[5], Indra Sarathan[6], Yusa Djuyandi[7]

Department of Computer Science, Universitas Padjadjaran, Sumedang, Indonesia[1, 2, 4]
Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Rabigh, Saudi Arabia[3]
Magister of Management, Institut Teknologi dan Bisnis Ahmad Dahlan Lamongan, Indonesia[5]
Faculty of Cultural Sciences, Universitas Padjadjaran, Sumedang, Indonesia[6]
Faculty of Social and Political Science, Universitas Padjadjaran, Sumedang, Indonesia[7]

***Abstract*—One of the initiatives adopted by the Indonesian government to combat the development of COVID-19 in Indonesia is Community Activities Restrictions Enforcement. Many public opinions emerged, both for and against this policy. There are so many comments every second that it is certainly not easy to analyze them by reading each one by one. This task necessitates computer applications. Therefore, this study was conducted to produce an application that can help analyze public sentiment on the policy through social media, namely Twitter, into three classes: positive, neutral, and negative. The method used in this research is bidirectional long short-term memory (BiLSTM), one of the algorithms of deep learning. This study trains the model using the dataset, which consists of 10,486 tweets. The model receives an f1-score of 76.67 %. Thus, the model can be used to analyze public sentiment when the same policy is enforced. It can determine public acceptance of this policy. Thus, the system created in this research can be used as evaluation material for the government to review the policy when it is implemented in the future. However, this study concentrates on how to develop the sentiment analysis system and does not examine how the community responds to government policy.**

***Keywords*—Sentiment analysis; COVID-19; BiLSTM; deep learning; government policy**

I. INTRODUCTION

There are several obstacles that must be overcome in order for Indonesia's political system to function effectively given the country's status as a growing democracy with a sizable population [1, 2]. The political choices that are made by the government have significant repercussions, both directly and indirectly, on the daily lives of the people who live in Indonesia. This encompasses fundamental concerns like as the distribution of wealth, economic equality, human rights, the maintenance of a healthy environment, and the satisfaction of fundamental needs like education and health care. As a result, the leaders and policymakers in Indonesia have a significant challenge in terms of preserving political stability and developing appropriate policies.

The public mood regarding these policies is a reflection of the sentiments, opinions, and perspectives of many individuals and groups [3]. The vast majority of these sensations may be freely communicated through social media, online news platforms, or online discussion forums, which are increasingly reflecting people's emotions in real-time [4]. This point of view should not be ignored by the government, which is the organization that is responsible for formulating and enforcing policies. They should, rather, view this sentiment data as a useful source of knowledge that may assist them in better directing their policies and should do so by capitalizing on it.

The political discourse and polarization that exists in Indonesia have a significant impact on the decision-making processes there as well [5, 6]. Analysis of public sentiment can assist in detecting recurring patterns of support for or opposition to a certain program or leader. It gives an overview of the extent to which diverse political and social ideas exist in society, which is vital for establishing an effective communication strategy and defusing political tensions.

The influence of sentiment analysis can also have an effect on how the public perceives the government in today's digital and social media-driven environment, which links the public at large [7,8]. It is much easier for governments to earn the support and trust of their constituents when they react rapidly to the sentiments and viewpoints expressed by the populace. On the other hand, apathy for public feelings or a refusal to respond to them might result in demonstrations.

As a result, sentiment analysis is not only a tool for analyzing public mood, but it is also a critical factor in the effective crafting of policies and sustaining the political stability that is required in Indonesia's ever-shifting process of democratization. The use of technology such as machine learning in sentiment analysis provides governments with more powerful and efficient tools to analyze large amounts of data and respond in a timely way. This can have a beneficial influence on the government's ability to be effective and responsive. As a result, having a grasp of and making use of sentiment analysis in the context of political policy in Indonesia is an absolute must in order to successfully navigate the political and social dynamics that will continue to emerge in the years to come.

In order to address this issue, the present study used sentiment analysis to evaluate the political strategies of the Indonesian government in response to the COVID-19 pandemic. The Indonesian government published the implementation of restrictions for community activities (PPKM) policy on July 3, 202. It is one of the government's measures to regulate community activities in order to combat the rise in COVID-19 cases in Indonesia. From July 3 until July 20, this policy—originally known as the PPKM Java-Bali emergency—was formally in effect. Then, on July 20, the president announced the extension of PPKM until the 25th and modified the designation of PPKM to level 4 during a virtual press conference. It was used to stop the spread of COVID-19 and minimize the requirement for inpatient care. The 44 areas and cities that made up the Java and Bali regions at first were the only ones to which it was applied. On July 12, however, the administration decided that it would also apply to non-Java and non-Bali areas after deliberating for more than a week. 15 cities and regencies in regions with the highest COVID-19 spread rates were covered by it outside of Java and Bali. The government set restrictions on people's mobility during the era. Residents, for example, must bring a minimum vaccination letter and the results of a PCR swab test or an antigen test indicating a negative result for COVID-19 when traveling by any mode of transportation [9].

The enactment of this policy caused controversy among the people of Indonesia. Some people supported this policy, but not a few people opposed it for certain reasons. Of course, the government issued this policy for the welfare of society. If it is discovered that the policy causes social unrest, the government must respond. The public's opinion of a policy can be seen through the comments that appear in the mass media [10]. One of them is Twitter, a social media site. There are so many comments every second that it is certainly not easy to analyze them by reading each one by one. This task necessitates computer applications. This application can facilitate the analysis of the sentiment that appears in all Twitter comments on this policy automatically and quickly. The computer analysis of textual opinions, attitudes, and emotions is known as sentiment analysis, also referred to as opinion mining [11]. Natural language processing, text analysis, computational linguistics, and biometrics are used to systematically identify, extract, quantify, and study emotional states and subjective information [12]. It is widely used for textual data to help businesses or organizations understand customers' needs based on their opinions of a product or service. The government needs this study's sentiment analysis to help it understand how the general public feels about PPKM policies.

Different techniques can be used to analyze sentiment [13]. One of the deep learning algorithms that are most frequently used is long short-term memory (LSTM) [14]. It is more accurate than the recurrent neural network (RNN) approach for problems involving long texts. It was developed to solve the long-term memory issue brought on by fading gradients while using RNNs to process massive sequential data [15]. When a gradient has a very low value or is extremely close to zero, the network weight remains constant, indicating that the training process cannot be sustained. However, the LSTM algorithm still has some drawbacks. The standard LSTM only considers the previous context. Despite the fact that text identification is seen not only in one direction from the previous context but also in the next context [16]. To address this issue, this work employs the bidirectional long short-term memory (BiLSTM) method. Based on the foregoing, it develops the BiLSTM model to automatically categorize the public sentiment toward PPKM policies during the COVID-19 pandemic. The data came from social media users' tweets on PPKM. This model is a fundamental component of computer algorithms that automatically categorize data tweets into three groups. Thus, there is no need to read each remark individually. Since every second there are so many comments. This categorization is reliant on language hence the success of an application cannot be directly applied if a different language is used. This application's outcomes are supposed to let the government observe the response of the Indonesian people to policies created more quickly and readily, so that they may be examined and utilized as material to consider when establishing a policy in the future. This study concentrates on how to develop the system and does not examine how the community responds to government policy. A second contribution of this study is the generation of a novel dataset that is unique to the Indonesian language for the PPKM topic.

II. Related Works

Sentiments are the underlying feelings, attitudes, assessments, or emotions underpinning an individual's viewpoint. Sentiment orientation might be positive, negative, or neutral. A sentiment target sometimes referred to as an opinion target, is the entity or feature of the entity about which the feeling is conveyed. The computational study of attitudes, opinions, judgments, and feelings concerning things and their features that are expressed in text is known as sentiment analysis. In general, sentiment analysis commonly uses two approaches: the rule-based/lexicon-based approach [17] and the machine learning/supervised learning approach [18]. By concentrating on the structure and features of the social network, sentiment analysis, and dangers including spam, bots, false news, and hate speech, Antonakaki et al. provide an effort to map the current study themes in Twitter [19]. Best practices for data sampling and access are also presented, in addition to Twitter's fundamental data model. This overview provides a foundation for the application of computational methods like Graph Sampling, Natural Language Processing, and Machine Learning in various fields.

COVID-19-related sentiment analysis research has also been extensively explored in recent years. Ridwan et al. collected English tweets that talked about COVID-19 and were geolocated as "Singapore." [20]. The system is based on the VADER lexicon-based classifier and the emotions from the pre-trained recurrent neural networks to find correlations between real-world events and changes in sentiment over the whole time period. The results of the sentiment analysis showed that about half of the tweets in the dataset were positive, and about a quarter of the tweets were negative or neutral. Topic modeling also showed that most of the talk about COVID-19 on Twitter was about staying home and that these conversations were mostly positive, which contributed to the overall positive mood during the study period. Overall, the results showed that the community supported the steps taken

by the Singapore government during COVID-19. This was clear from the positive comments on Twitter.

Sattar et al. use Twitter data to figure out how people feel about different COVID-19 vaccines [21]. It used the Twitter API and the Python library Tweepy to get about 1.2 million tweets. The tweets were collected from April 10 to May 17, 2021. It only got English tweets, and we used NLTK to do more data analysis. It builds the forecasting model classifier with well-known machine learning regression algorithms: Support Vector Machine (SVM) for regression, k-Nearest Neighbor (KNN), Linear Regression (LR), Random Forest (RF), M5 model tree, Gaussian process for regression, and Multi-layer Perceptron (MLP). Even though some of the vaccines have side effects, we find that the general public is more positive than negative about them. The model for predicting vaccinations says that by the end of July 2021, 62.44% of the population will have had at least one dose of vaccine and 48% will be fully vaccinated. The prediction model says that 73.53% of adults will be partially vaccinated on Independence Day (4 July 2021). The findings give a way to measure how the public talks about the COVID-19 vaccination and how to live a healthy life during the pandemic. It classifies tweets into different emotions like inspired, happy, annoyed, sad, angry, afraid, etc.

Many other studies related to sentiment analysis in English have been carried out, such as in English [22], Arabic [23], Italian [24], and Hindi [25]. Moreover, Intan NY et al. did COVID-19-related research in Indonesia. The research addresses the COVID-19 problem in Indonesia, for policies related to vacations [26]. Government initiatives regarding this vacation program have resulted in a diversity of societal perspectives. Using a bidirectional encoder representation from transformers (BERT) technique, they apply sentiment analysis to this government policy. This research produced a new dataset on this subject. The data was taken from the YouTube comments area and classified into three categories: favorable, neutral, and negative. This study generated an F-score of 84.33 percent. Using Twitter data, more COVID-related studies conducted in Indonesia have been conducted [27, 28].

III. METHODOLOGY

Our case study deals with the implementation of government regulations to address the increase in COVID-19 cases in Indonesia. PPKM has gained support from a range of segments of Indonesian society. Twitter is a social media platform where the general public may express their ideas. This research was conducted through stages that included data collection from Twitter social media, data labeling, text preprocessing, data splitting, feature selection, feature extraction, modeling, and evaluation of the model that has been built. The proposed framework is illustrated in Fig. 1.

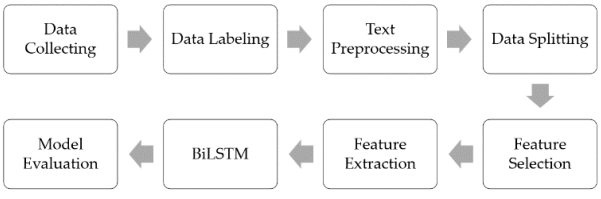


Fig. 1. Proposed framework.

A. Data Collecting

Data obtained from Twitter about the PPKM policy served as the study's subject. The unstructured, heterogeneous input is then categorized into attitudes that are positive, negative, or neutral. The procedure of data collection involved tweets posted by Twitter users between July 1 and August 19, 2021. 10,486 tweets, or 200 tweets each day for 50 days in this period make up the research data. Nevertheless, the obtained data for this study were only collected in Indonesian. Using the Python module snscrape, the scraping operation is carried out to retrieve the data. Government socialization of this policy began on July 1 and was completed on July 3, 2021. Fig. 2 depicts the peak of the topic discussion trend in Indonesia from July 4 through 10, which was the first week that this policy was put into effect.

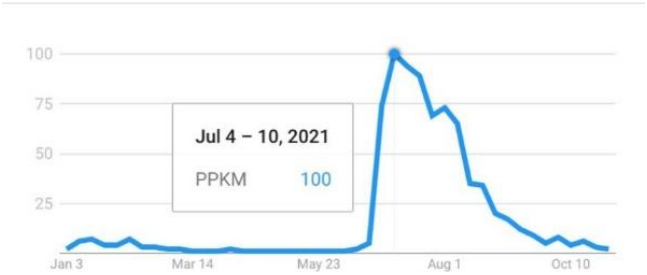


Fig. 2. Trends in PPKM policy tweets in Indonesia.

B. Data Labeling

The labeling of the collected tweet data was done manually, one tweet at a time. Several members of the Faculty of Cultural Sciences of Universitas Padjadjaran were implicated. The findings served as the foundation for this study. Negative tweets comprise profanity, expressions of hostility, criticism, wrath, fear, and despair, as well as expressions of disagreement with government-issued PPKM policies. In the meanwhile, tweets can be classified as positive if they contain terms with positive connotations, give support, demonstrate delight, encourage compliance, and comply with other PPKM criteria. If a tweet does not fall into the good or negative categories, it is labeled as neutral. In this phase, tweets that were deemed to be noise or unrelated to the opinions of Indonesians were also deleted. Its process involved manually deleting tweet data obtained from news accounts, government entities, nonprofits, and businesses. Additionally, tweets with promotions, ads, and tweets in regional languages were removed.

C. Text Preprocessing

Before the text data was processed to the next step, this stage cleaned the data of noise or modified the data's format to make the data processing easier [29]. Each language has its own characteristics hence the procedures and methodologies might differ between languages. The actions taken in this study, i.e.

- Case folding: This step transforms a word's individual letters into a single, consistent shape that may take the form of lowercase or uppercase letters [30].
- Noise reduction: During this step, the text was cleared of distracting elements including mentions, hashtags, links, numbers, symbols, punctuation, and emoticons.

- Tokenizing: This method divided the text into tokens, which were utilized to create practical features. Words in their entirety serve as the stage's tokens.
- Normalizing: In this step, the tokens were normalized to a standardized form. The Indonesian dictionary updated them with standard forms that replaced abbreviations, slang, and other non-standard forms.
- Stemming: By locating and eliminating any affixes present, this step reduced the word to its most basic form.
- Stop-word removal: In this phase, terms that were used excessively or whose definitions were irrelevant to learning are eliminated.

D. Data Splitting

This study used a 7:3 ratio to divide the dataset into training and testing data. 7,340 tweets were used for the training data, whereas 3,146 were used for the testing data. A 10 k-fold cross-validation procedure was then used to separate the training data from the validation data.

E. Feature Selection

Experiments on the minimum count threshold value were conducted during the feature selection step. The limit utilized to filter the existing characteristics based on a word's frequency of occurrence across the entire corpus was the min count value. The characteristics in question in sentiment analysis are the tokens or words that make up a phrase. At this point, irrelevant tokens were eliminated, namely those with too few occurrences

F. Feature Extraction

The process of feature extraction involves numerous initial feature changes that result in the production of new, more important features. It can be used in this situation to simplify the data representation and minimize complexity by treating each variable as a linear combination of the original input variables. The process of obtaining word lists from text data and turning them into a collection of features that the classifier may use is known as feature extraction on text [31]. Word2Vec is a method for feature extraction that appears to be an improvement over the drawbacks of the bag of words (BOW) method, which can create inaccurate models because it doesn't take into account word order or context. To create the appropriate context, it is important to take into account the arrangement of words in a document. When the same type and number of words are used in a document, the word order might yield distinct interpretations, which will impact the sentiment. Another strategy, word embedding, can be applied to get around this. It is a method of representing text in which similar-sounding words are represented similarly in a vector space [32]. In other words, it displays similar terms based on how closely connected the corpus is. It can often be divided into two categories: pre-diction-based embedding and frequency-based embedding. The count vector, TF-IDF, and co-occurrence vector techniques are the ones most frequently employed in frequency-based embedding. As for the frequently employed prediction-based approach, it is Word2Vec. It uses the context of the words around it to quantitatively identify word similarities. It makes vectors whose numerical representations are based on things about words, like how they are used in a sentence. Also, it trains a corpus of text as input data and generates a vector list of words (embedding) as output from a model. The resulting word meanings and word embedding relationships have beneficial properties like vector arithmetic and are spatially represented [33]. It is a self-supervised learning system that gains knowledge from unlabeled input, like a group of texts. The learning technique makes use of conditional probabilities to anticipate certain phrases using some of the words they are surrounded by in a corpus of text. Word pairs that are close in proximity to one another in the text and were one-hot encoded (input, label) make up the training set.

Skip-gram architecture was used in this study since, according to research; it performs better [34]. During training, the skip-gram algorithm could be used to predict the word context of the input word by examining the words nearby. The size of the dimensions and the size of the window serve as the Word2vec hyperparameters that would be optimized. 50, 150, and 300 are the study's test dimensions, whereas 2, 5, 10, and 14 are the study's test window sizes.

G. BiLSTM

The RNN technique, which is utilized in the field of deep learning, has undergone development in the form of LSTM. It was intended to alleviate the vanishing and exploding gradient problems that occur when RNNs are employed to handle extensive sequential input. The cells can learn to recognize substantial input with the aid of the input gate, store it for a long time with the aid of the forget gate, learn to hold onto it for however long is necessary with the assistance of the forget gate, and then learn to delete it when needed. This explains why the algorithm is capable of picking up persistent patterns in time series, lengthy texts, and other data [35].

A memory cell with three gates requires the following gates to control the flow of information:

- Forget gate to decide what data should be removed from the memory cell or not,
- Input gate, which decides what data should be entered into the memory cell,
- Output gates are used to decide the output based on the input and memory cell.

LSTM's first step is to decide what data will be eliminated from the cell state. The choice is made by the sigmoid activation function that takes place in the forget gate layer. The forget gate will take input from ht-1 and xt and output a value between 0 and 1. While 0 indicates that the information will be stored from Ct-1, and 1 indicates that it will be removed.

The information that will be kept in the cell state must then be decided. It is split into two sections. The input gate layer, also known as the sigmoid activation function, first decides what data will be updated. Following that, the vector Ct of new candidate values that can be added to the state is created by the tanh activation function. The outcomes of the sigmoid and tanh activation functions will be merged in the following step, and their values will be utilized to update the state. The subsequent step is to transform the prior cell state, Ct-1, into the new cell

state, Ct. At this point, the value of the new candidate will be added to the prior state after being multiplied by the output from the input gate, and the output from the forget gate. At this stage, processes such as the removal of unimportant information and the addition of significant actual information into the cell state, as defined in the preceding step, occur [36]. The filtered cell state determines the output outcomes. The sigmoid activation function will initially mix input and data from the preceding hidden layer to determine which portion of the cell state should be created. The output from the sigmoid is then multiplied by the output from the tanh activation function (to obtain a value between -1 and 1), yielding only the preset information. Each time step of the LSTM cell's overall process provides two outputs: Ot, which serves as the time step's output and is used as a hidden state, and Ct, which creates a new cell state and is passed on to the following time step.

BiLSTM, a variant of the LSTM algorithm, tries to observe a certain sequence from front to back or back to front. The network creates a context for each character in the text by utilizing their past and future. The goal is to maximize the utilization of an input sequence by stepping through the input step time both forward and backward. In order to create two layers side by side using this design, the first layer is duplicated repeatedly. The input sequence is then fed into the first layer, and the second layer receives a reverse copy of the input sequence. This strategy was created a while back as a general strategy to enhance the functionality of recurrent neural networks [37]. the forward layer and backward layer's information flow. It is typically employed when the order of the data is a factor. It is justified in the context of voice recognition since research suggests that rather than a one-way linear interpretation, humans perceive what is spoken in the context of complete speech. Although it was initially designed for speech recognition, the bidirectional is now a crucial component of LSTM sequence prediction as a method to enhance model performance.

The BiLSTM model neural network layer in this study can be illustrated in Fig. 3. This study's BiLSTM model was developed using the Keras module, which may be obtained at https://keras.io (accessed on 18 February 2020). The first layer was the embedding layer whose embedding matrix values were obtained from the Word2vec model. The embedding matrix resulting from Word2vec was used as the initialization value of the embedding layer's weight which was then updated during the training process. Input on this layer was in the form of tweets which were represented in word index sequences. The sequence had a maximum length of 65 words. Tweet data consisting of less than 65 words was a padded 0 in front of it, while data tweets that were more than 65 words will be truncated so that the length of each tweet sequence is the same. When a shape tuple has a none dimension, it indicates that the network is open to inputs of any dimension.

A 65 x 150-pixel matrix representing each piece of data was the embedding layer's output which is from the feature extraction stage. The maximum length for each data point was 65, and the maximum length for each word's vector representation was 150. The matrix served as the BiLSTM layer's input. The forward LSTM layer and the back-ward LSTM layer are the two LSTM layers that make up the BiLSTM layer, which is made up of one layer altogether. The output of the BiLSTM layer has a size of 65 32 thanks to the combination of the 16 units from each of these LSTM layers with a dropout of 0.5. The GlobalMaxPool1D layer was then given the output from the BiLSTM layer. By taking the largest value in the temporal dimension, this layer was able to down-sample two-dimensional data into one dimension, producing an output with 32 dimensions. It then moves on to the layer for batch normalization. It altered the input to make it uniform. Each input variable's statistics were tracked by the layer during training and utilized to normalize the data. Through a straightforward regularization effect, it can expedite the training process and, in some situations, improve model performance.

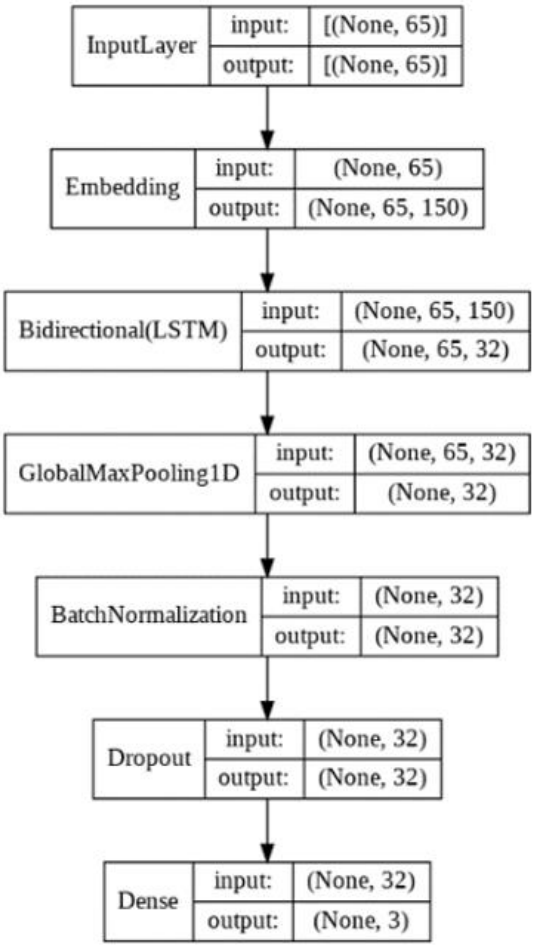


Fig. 3. BiLSTM model architecture.

TABLE I. HYPERPARAMETER VALUES

Hyperparameter	Value
Optimizer	Adam
Merge mode	Concat
Dimension size	150
Activation	Softmax
Max sequence length	65
Max epoch	50

Thus, the layer based on the dropout was a safeguard against overfitting [38], and the resulting layer was dense, fully connected, and had three units. The layer served as the output layer for the Softmax activation function, which created class probabilities from the input with values ranging from 0 419 to 1 for each class. The prediction of input data had the three classes' greatest value making the following step reasonable. It trained the model's trainable parameters defined and assembled earlier. Early stopping was used to avoid overfitting during the model training process using data by assessing the hyperparameter batch size with epoch 50. Table I displays the hyperparameter values used in this investigation. The grid search method in finding optimal hyperparameters was used, in which every combination of predetermined hyperparameters was tried to select the combination that produced the best f1-score value.

H. Model Evaluation

After the BiLSTM model had been successfully created, the next step was to evaluate the model using the confusion matrix values. The metric in this study was the f1-score. It is one of the assessment measures that is often used in sentiment analysis scores. This metric is used to quantify how well a model is able to categorize comments into various sentiment categories, such as positive, negative, or neutral. The F1-score is a metric that offers an overall view of the quality of the model classification since it combines both precision and recall into a single value. Because it strikes a good balance between precision and recall, it is an appropriate tool for use in situations in which the ratio of false positives to false negatives is a significant consideration in sentiment analysis. When identifying sentiment, a high F1 score shows that the model has a high level of precision as well as recall.

IV. Results

The hyperparameters that were analyzed to create the optimal BiLSTM model were the number of BiLSTM layers, the number of BiLSTM units, batch size, dropout rate, and learning rate.

A. Analysis of the Number of Layers

With dropout rate=0.1, batch size=16, and learning rate=0.01, one and BiLSTM layers one and two were stacked against one another and evaluated in this study. Table II shows that using one BiLSTM layer as opposed to two leads in a higher f1-score. The findings of this experiment demonstrate that employing one layer of BiLSTM instead of two layers led to a higher f1-score. It is clear that the model with two layers of BiLSTM could not ensure that the model's performance would be enhanced. Because it yielded improved f1-scores, utilizing just a BiLSTM layer was sufficient for this study. The capacity of the model to learn can theoretically be increased by adding layers to the neural network. However, the model can extract more intricate patterns from the supplied data because of its expanded learning capacity. In general, this is advantageous because the model learns more and becomes better at properly predicting data. On the other side, there is a chance that the created model will be overfitted as a result. According to this criterion, the model performed well when predicting data that has been thoroughly researched but poorly when predicting data that has never been observed.

TABLE II. Analysis of the Number of Layers

Number of layer	LSTM Unit	F1 score
1	16	74.49%
1	64	74.27%
2	16	72.72%
2	64	73.40%

B. Analysis of BiLSTM Unit

Table III demonstrates that an f1-score with a declining trend was obtained by in-creasing the hidden units in the LSTM layer. The model with 16 hidden units generated the model with the highest f1-score, 74.49 %. In this study, 16, 64, and 128 BiLSTM units were examined. Table III gives an overview of the test outcomes that have been evaluated. The test results demonstrate that an f1-score value with a falling trend will be produced by increasing the number of hidden units in the BiLSTM layer. The model with 16 hidden units yields the model with the highest f1-score, or 74.49 %, compared to the other models. It demonstrated that utilizing 16 hidden units in each BiLSTM layer—or 32 units in the BiLSTM layer—was sufficient for the model to extract patterns from the supplied dataset. A higher hidden unit count also did not result in a higher f1-score. It is due to the fact that a model with an excessively large number of hidden units may get overfitted, have a low f1-score in the validation data, and be overly complex.

TABLE III. Results of The BiLSTM Unit

Number of layers	BiLSTM Unit	F1 score
1	16	74.49%
1	64	74.27%
1	128	73.46%

C. Analysis of Dropout

The best outcomes are displayed in Table IV when a dropout of 0.5 was used. It suggests that the optimal model would consist of a BiLSTM layer, 16 LSTM units, and a dropout of 0.5, which would effectively reduce overfitting, raise the average f1-score value, and offer the best model. The model's ability to handle overfitting was less ideal due to the dropout value being too small. Because too many neurons were eliminated during the training phase due to an excessively high dropout number, the final model was subpar.

TABLE IV. Analysis of the Dropout Rate

Dropout rate	LSTM Unit	F1 score
0.1	16	74.49%
0.3	16	74.08%
0.5	16	74.57%

D. Analysis of Batch Size

According to Table V, batch size 32 produced the highest f1-score when compared to the other sizes. The batch size value was about right—it wasn't either too big or too tiny. Larger data sets result in faster model convergence, but since there are so many of them, the model finds it challenging to

identify patterns in the data. Otherwise, if it is too little, the algorithm might not take into account the true level of variance in the sampled distribution, which could lead to a noisy training process.

TABLE V. ANALYSIS OF THE BATCH SIZE

Dropout rate	Batch size	F1 score
0.5	16	74.57%
0.5	32	75.37%
0.5	64	75.08%

E. Analysis of Learning Rate

Table VI shows that when the learning rate was set to 0.001, the model performed at its best. If the learning rate was excessively sluggish and required more repetitions, convergence happened more slowly. If the learning rate was too high, the convergence would occur quickly but the step size would be enormous and the optimal point would be exceeded. The table shows that a model with a learning rate of 0.0001 has a substantially lower f1-score.

TABLE VI. ANALYSIS OF THE LEARNING RATE

Batch size	Learning rate	F1 score
32	0,01	75.37%
32	0,001	76.34%
32	0,0001	70.40%

F. Model Evaluation

The tests with a 10-fold cross-validation technique is presented in Sub-chapters 4.1 to 4.5. The model was then applied to the test data that had been previously separated. 3,146 data points total—including 1,510 negative class data, 1,236 neutral class data, and 400 positive class data—were utilized to evaluate the final model. Fig. 4 demonstrates that in epoch 8, the model obtained the best f1-score. The overfitting that caused the early termination mechanism to end the training process automatically in epoch 9 can be seen. It also demonstrates that there are not many differences in the curves between the training and validation sets of data, indicating that the developed model performs reasonably well. The results of the final model test on the test data shown in the confusion matrix are shown in Fig. 5.

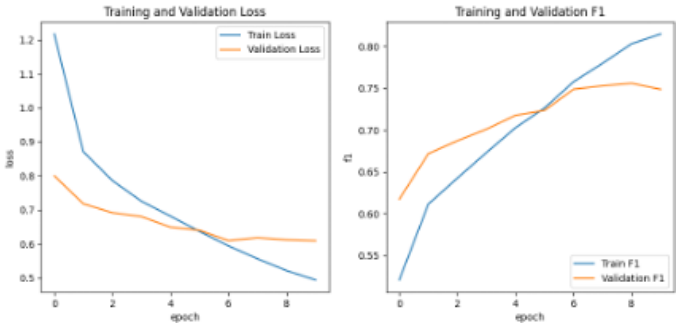


Fig. 4. Graph of F1-Score and loss.

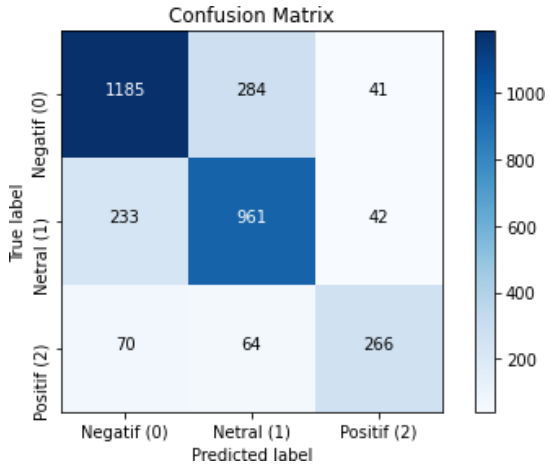


Fig. 5. Confusion matrix.

Also, Table VII displays the precision, recall, and f1-score for each class. The classification result shows metrics for each negative, neutral, and positive class. It is clear that negative classes frequently have metrics that are superior to those of other classes. The built-in model's performance receives a f1-score of 76.67 %.

TABLE VII. MODEL EVALUATION RESULTS

Metric	Negative	Neutral	Positive
Precision	79.64%	73.41%	76.22%
Recall	78.48%	77.75%	66.50%
F1-score	79.05%	75.52%	71.03%

Fig. 6. PPKM trends in November 2021.

G. Sentiment Prediction

This research focuses on the development of computer applications for PPKM-related sentiment analysis in Indonesia. The BiLSTM model was derived from the collected ground truth data in this study. By this model, various sentiment analysis-related data, specifically PPKM-related data, may be automatically classified. An example of test data is PPKM comments data collected in November 2021. In order to avoid an upsurge in COVID-19 cases in Indonesia before the year-end holidays, the government raised the PPKM level once again at the end of 2021. As a result, some Indonesians became quite interested in the PPKM subject. Fig. 6 displays a Google Trends graph of the PPKM keyword's search volume in November 2021. The graph demonstrates that on November 18, the PPKM again declared a peak. On November 17, 2021, the government declared that PPKM level 3 would once again be implemented. 31,788 tweets in total were collected. Based

on this research, there are 10,370, 13,836, and 7.58, respectively, fell into the negative, neutral, and positive categories. In November 2021, a neutral class of 43.5 percent dominated the public's opinion on the PPKM policy. It was followed by a negative class of 32.6 percent, while a small percentage of 23.9 percent expressed a favorable attitude. The majority of tweets that were categorized as neutral provide news or information about PPKM policy. The tweet content that was labeled as unfavorable typically reflects the public's dissatisfaction and frustration with the effects of PPKM's implementation. It demonstrates that the majority of Indonesians disapprove of the PPKM policy, which the government can utilize as evaluation information for reviewing current and future policies.

V. CONCLUSION

The topic of the study was data acquired from Twitter on the PPKM policy. The input is then classified as positive, negative, or neutral. The technique for collecting data involves tweets written by Twitter users between July 1 and August 19, 2021. The research data consists of 10,486 tweets, or 200 tweets every day over 50 days. This is the study's contribution. It is the creation of a new dataset that is unique to the Indonesian language in relation to the PPKM issue.

Also, this study creates the BiLSTM model to automatically classify public opinion on PPKM policies during the COVID-19 epidemic. The best model, with a final f1-score of 76.67 %. Reading each comment individually is unnecessary. Since every second, so many comments are posted. This categorization is dependent on language; hence, the success of an application cannot be directly applied when a different language is utilized. The results of this application are intended to allow the government to monitor the response of the Indonesian people to newly formed policies more quickly and easily, so that they may be analyzed and used as future policy-making material. This research focuses on developing the system and does not investigate how the community reacts to government policies. Last, this article only discusses one of the deep learning algorithms, specifically the BiLSTM method. With the rapid growth of deep learning, however, research may be conducted on the most recent algorithms with the goal of improving performance.

The focus of future work will be on developing more sophisticated pre-processing methods to improve the accuracy of public opinion sentiment on government policy during the COVID-19 pandemic, as well as exploring the application of transfer learning techniques to strengthen understanding of complex patterns in Indonesian language text data.

ACKNOWLEDGMENT

The funding for this project was provided by the Library and Online Data Research Grant from Universitas Padjadjaran, under contract number 1959/UN6.3.1/PT.00/2021.

REFERENCES

[1] E. Warburton, “Deepening Polarization and Democratic Decline in Indonesia,” in Political Polarization in South and Southest Asia Old Divisions, New Dangers, 2020.

[2] E. Warburton and E. Aspinall, “Explaining Indonesia’s Democratic Regression,” Southeast Asia, vol. 41, no. 2, 2019.

[3] X. Dong and Y. Lian, “A review of social media-based public opinion analyses: Challenges and recommendations,” Technol Soc, vol. 67, 2021, doi: 10.1016/j.techsoc.2021.101724.

[4] Y. Lin, “Social media for collaborative planning: A typology of support functions and challenges,” Cities, vol. 125, 2022, doi: 10.1016/j.cities.2022.103641.

[5] E. E. Supriyanto and J. Saputra, “Big Data and Artificial Intelligence in Policy Making: A Mini-Review Approach,” International Journal of Advances in Social Sciences and Humanities, vol. 1, no. 2, 2022, doi: 10.56225/ijassh.v1i2.40.

[6] S. Partelow and A. O. Manlosa, “Commoning the governance: a review of literature and the integration of power,” Sustain Sci, vol. 18, no. 1, 2023, doi: 10.1007/s11625-022-01191-2.

[7] P. Grover, A. K. Kar, and Y. Dwivedi, “The evolution of social media influence - A literature review and research agenda,” International Journal of Information Management Data Insights, vol. 2, no. 2, 2022, doi: 10.1016/j.jjimei.2022.100116.

[8] E. Kubin and C. von Sikorski, “The role of (social) media in political polarization: a systematic review,” Ann Int Commun Assoc, vol. 45, no. 3, 2021, doi: 10.1080/23808985.2021.1976070.

[9] F. Sanjaya, “A Critical Review of the RT-PCR Test Policy and Antigen Swab in Indonesia,” in Proceedings of the International Conference on Communication, Policy and Social Science (InCCluSi 2022), 2022. doi: 10.2991/978-2-494069-07-7_48.

[10] S. Hansun, A. Suryadibrata, R. Nurhasanah, and J. Fitra, “Tweets Sentiment on Ppkm Policy as a Covid-19 Response in Indonesia,” Indian Journal of Computer Science and Engineering, vol. 13, no. 1, 2022, doi: 10.21817/indjcse/2022/v13i1/221301302.

[11] A. Yadav and D. K. Vishwakarma, “Sentiment analysis using deep learning architectures: a review,” Artif Intell Rev, vol. 53, no. 6, 2020, doi: 10.1007/s10462-019-09794-5.

[12] M. R. Hasan, M. Maliha, and M. Arifuzzaman, “Sentiment Analysis with NLP on Twitter Data,” in 5th International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering, IC4ME2 2019, 2019. doi: 10.1109/IC4ME247184.2019.9036670.

[13] K. Afifah, I. N. Yulita, and I. Sarathan, “Sentiment Analysis on Telemedicine App Reviews using XGBoost Classifier,” in 2021 International Conference on Artificial Intelligence and Big Data Analytics, ICAIBDA 2021, 2021. doi: 10.1109/ICAIBDA53487.2021.9689762.

[14] Z. Jin, Y. Yang, and Y. Liu, “Stock closing price prediction based on sentiment analysis and LSTM,” Neural Comput Appl, vol. 32, no. 13, 2020, doi: 10.1007/s00521-019-04504-2.

[15] A. Shewalkar, D. nyavanandi, and S. A. Ludwig, “Performance Evaluation of Deep neural networks Applied to Speech Recognition: Rnn, LSTM and GRU,” Journal of Artificial Intelligence and Soft Computing Research, vol. 9, no. 4, 2019, doi: 10.2478/jaiscr-2019-0006.

[16] I. N. Yulita, M. I. Fanany, and A. M. Arymurthy, “Combining deep belief networks and bidirectional long short-term memory case study: Sleep stage classification,” in International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2017. doi: 10.11591/eecsi.4.1051.

[17] A. Mitra, “Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset),” Journal of Ubiquitous Computing and Communication Technologies, vol. 2, no. 3, 2020, doi: 10.36548/jucct.2020.3.004.

[18] P. K. Jain, R. Pamula, and G. Srivastava, “A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews,” Computer Science Review, vol. 41. 2021. doi: 10.1016/j.cosrev.2021.100413.

[19] D. Antonakaki, P. Fragopoulou, and S. Ioannidis, “A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks,” Expert Syst Appl, vol. 164, 2021, doi: 10.1016/j.eswa.2020.114006.

[20] K. Mohamed Ridhwan and C. A. Hargreaves, “Leveraging Twitter data to understand public sentiment for the COVID-19 outbreak in Singapore,” International Journal of Information Management Data Insights, vol. 1, no. 2, 2021, doi: 10.1016/j.jjimei.2021.100021.

[21] N. S. Sattar and S. Arifuzzaman, "Covid-19 vaccination awareness and aftermath: Public sentiment analysis on twitter data and vaccinated population prediction in the usa," Applied Sciences (Switzerland), vol. 11, no. 13, 2021, doi: 10.3390/app11136128.

[22] C. Baden, C. Pipal, M. Schoonvelde, and M. A. C. G. van der Velden, "Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda," Commun Methods Meas, vol. 16, no. 1, 2022, doi: 10.1080/19312458.2021.2015574.

[23] L. Abualigah, H. E. Alfar, M. Shehab, and A. M. A. Hussein, "Sentiment Analysis in Healthcare: A Brief Review," in Studies in Computational Intelligence, vol. 874, 2020. doi: 10.1007/978-3-030-34614-0_7.

[24] R. Catelli, S. Pelosi, and M. Esposito, "Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian," Electronics (Switzerland), vol. 11, no. 3, 2022, doi: 10.3390/electronics11030374.

[25] P. Chauhan, N. Sharma, and G. Sikka, "The emergence of social media data and sentiment analysis in election prediction," J Ambient Intell Humaniz Comput, vol. 12, no. 2, 2021, doi: 10.1007/s12652-020-02423-y.

[26] I. N. Yulita, V. Wijaya, R. Rosadi, I. Sarathan, Y. Djuyandi, and A. S. Prabuwono, "Analysis of Government Policy Sentiment Regarding Vacation during the COVID-19 Pandemic Using the Bidirectional Encoder Representation from Transformers (BERT)," Data (Basel), vol. 8, no. 3, 2023, doi: 10.3390/data8030046.

[27] S. H. Sahir, R. S. Ayu Ramadhana, M. F. Romadhon Marpaung, S. R. Munthe, and R. Watrianthos, "Online learning sentiment analysis during the covid-19 Indonesia pandemic using twitter data," IOP Conf Ser Mater Sci Eng, vol. 1156, no. 1, 2021, doi: 10.1088/1757-899x/1156/1/012011.

[28] M. Rahardi, A. Aminuddin, F. F. Abdulloh, and R. A. Nugroho, "Sentiment Analysis of Covid-19 Vaccination using Support Vector Machine in Indonesia," International Journal of Advanced Computer Science and Applications, vol. 13, no. 6, 2022, doi: 10.14569/IJACSA.2022.0130665.

[29] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," Organ Res Methods, vol. 25, no. 1, 2022, doi: 10.1177/1094428120971683.

[30] M. F. R. Abu Bakar, N. Idris, L. Shuib, and N. Khamis, "Sentiment Analysis of Noisy Malay Text: State of Art, Challenges and Future Work," IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2968955.

[31] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, "An ensemble machine learning approach through effective feature extraction to classify fake news," Future Generation Computer Systems, vol. 117, 2021, doi: 10.1016/j.future.2020.11.022.

[32] P. G. Shivakumar and P. Georgiou, "Confusion2Vec: Towards enriching vector space word representations with representational ambiguities," PeerJ Comput Sci, vol. 2019, no. 6, 2019, doi: 10.7717/peerj-cs.195.

[33] A. Hernandez-Suarez et al., "Using twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation," Sensors (Switzerland), vol. 19, no. 7, 2019, doi: 10.3390/s19071746.

[34] F. A. O. Santos, H. T. Macedo, T. D. Bispo, and C. Zanchettin, "Morphological skip-gram: Replacing fasttext characters n-gram with morphological knowledge," Inteligencia Artificial, vol. 24, no. 67, 2021, doi: 10.4114/intartif.vol24iss67pp1-17.

[35] K. Khalil, O. Eldash, A. Kumar, and M. Bayoumi, "Economic LSTM Approach for Recurrent Neural Networks," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 66, no. 11, 2019, doi: 10.1109/TCSII.2019.2924663.

[36] F. Landi, L. Baraldi, M. Cornia, and R. Cucchiara, "Working Memory Connections for LSTM," Neural Networks, vol. 144, 2021, doi: 10.1016/j.neunet.2021.08.030.

[37] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," Physica D, vol. 404, 2020, doi: 10.1016/j.physd.2019.132306.

[38] M. M. Agüero-Torales, J. I. Abreu Salas, and A. G. López-Herrera, "Deep learning and multilingual sentiment analysis on social media data: An overview," Appl Soft Comput, vol. 107, 2021, doi: 10.1016/j.asoc.2021.107373.

New AHP Improvement using COMET Method Characteristic to Eliminate Rank Reversal Phenomenon

Yulistia[1], Ermatita[2*], Samsuryadi[3], Abdiansah[4]
Doctoral Program in Engineering Science, Universitas Sriwijaya, Palembang Indonesia[1]
Faculty of Computer Science, Universitas Sriwijaya,Indralaya, Indonesia[2, 3, 4]
Faculty of Computer Science and Engineering, Universitas Multi Data Palembang, Palembang, Indonesia[1]

***Abstract*—Rank Reversal in Multi-Criteria Decision Making (MCDM) is a phenomenon that occurs when an alternative is added or deleted because of a change in the order in which the result is ranked. The evaluation of the weight of criteria, which are established based on whether a decision maker considers them important, impacts the alternative ranking result in MCDM. Changes in decision result ranking called rank reversal cannot be acceptable. Many researchers have done lots of research and created new methods for eliminating rank reversal, but until now there is still research that denies these new methods are free from rank reversal. The Analytical Hierarchy Process Method (AHP), the oldest Decision support Method has an advantage in the decision according to the Decision Maker's (DM's) preference. Still, it is vulnerable to the rank reversal phenomenon. While Characteristic Object Method (COMET) is a method claimed to be free of rank reversal phenomenon. This paper will discuss how the integration of COMET to AHP especially in the phase of generating characteristic value and characteristic objects is added to the AHP phase, which will have an impact on digital marketing strategy decision-making for private Universities in Indonesia, especially the city of Palembang. The combination of COMET and AHP in this paper is tested with several testing tools; they are case study testing, accuracy testing, and sensitivity analysis testing. The result of the combination of COMET and AHP will be named C-AHP, which is a consideration of DM's preference for the criteria weight, and the generation of alternative comparison based on criteria, or any other attributes makes AHP free from rank reversal.**

***Keywords*—Method; combination; C-AHP; rank reversal; elimination**

I. INTRODUCTION

Multi-Criteria Decision Making (MCDM) Alternative ranking is affected by the weight given by the DMs (Decision Makers) to the criteria. The main methods of MCDM such as TOPSIS (Technique for Order Preference by Similarity to Ideal Solution), ELECTRE (Elimination Et Choix Tradusiant la REalite), PROMETHEE (Preference Ranking Organization Method for Enrichment Evaluation), AHP (Analytical Hierarchy Process), and their combination, have been criticized in accordance with the occurrence of a problem called Rank Reversal Phenomenon (RRP) [1]. Rank reversal is a phenomenon where the alternative's order of preference is altered when a new alternative is added, or an existing alternative is deleted from a decision problem. A rank reversal occurs when a new alternative has been added or an old alternative eliminated from decisions, and the order of preference for other options is changed. In 1980 Belton and Gear first observed a change in ranking in the AHP [2]. One of the most important criteria for selecting the MCDM method is the phenomenon of rank reversal. There has been no answer to the question of a shift in rank as far as MCDM is concerned. Therefore, to obtain genuine results, a DMM (Decision-Making Method) using MCD (Multi-Criteria Decision) methods must know the problems that arise because of rank reversal. The rank reversal issue has not been solved yet in the MCDM context. Consequently, to obtain a valid result, DMs using MCDM methods should be familiar with rank reversal phenomenon challenges. Although a lot of researchers declare that RRP is a natural feature of the decision-making process, RRP is undesirable and unwanted because it indicates unreliability in the MCDM approach, in the research about sustainable material selection, the research result showed that there was no way to confirm whether the number of options and criteria had any effect on rank reversal [3].

There have been many studies regarding RRP, which tried to eliminate rank reversal using various methods for the past 10 years. In 2014, a framework for the experiments to determine the cause of rank reversal in an MCDM Method was done, the result is a modification of a method with a robust combination in it [4]. In addition to the RRP research carried out in 2017 with a reciprocal fuzzy preference relationship based on additive consistency for addressing RRP, there were also new methods that use proximity-indexed values and have demonstrated their accuracy compared to existing MCDM methods [1] [2]. An RRP investigation into potential causes of rank reversal was carried out in 2018, which indicates that preference followed by ranking score aggregation is the primary cause of RRP because of a lack of information in other research papers. A method for aggregation of scores has been proposed in some research to describe and illustrate the phenomenon of rank reversal using numerical examples. Compared to other tested methods, the results are better. According to the literature on decision-making, several methods suffer from this phenomenon, and one of them is AHP [5]. From 2020 era until 2023 the RRP research focused on the new method and/or the enhancement of the old method in MCDM with the additional Fuzzy to eliminate RRP such as AHP and TOPSIS method [6][7][8].

AHP as the method that applied for the last 25 years in many MCDM decision-making, has been used in lots of decisions in various fields. The pairwise comparison is the basic way of breaking down the problem into a hierarchy of subproblems making AHP a method with its advantages [9][10]. The numeric value assigned to each variable in AHP helps decision-makers defend a cohesive paradigm by deriving the relative weight of each component of the hierarchy, criteria, and alternatives [11]. For the management of qualitative and quantitative multicriteria elements in decisions, as a powerful and efficient tool, For the following criteria and benchmarks, AHP is capable of applying sensitivity analysis. Because of combined comparisons, judgment and calculation are easy in AHP. Moreover, AHP provides proof of compatibility and incompatibility decisions which is the compensation for the multicriteria decision making. With the combination of mathematics and expert judgment, AHP helps decision makers to make better choices both about tangible criteria and intangible criteria [12][13][14]. Besides its advantages, AHP also has disadvantages; RRP is neither a fatal flaw of the AHP nor a desirable property of it. It is a symptom of inherent problems with the AHP [15]. [16] When a decision problem is broken down into smaller problems, ranking irregularities occur in AHP each comprising two alternative solutions and a set of identical criteria as the first problem, which is called the Rank Reversal Phenomenon (RRP) in AHP.

Continuing the first version of the AHP method affected by RRP which was founded by Saaty & Vargas in 1984, Belton and Gear in 1983-1985, and Schoner & Wedley in 1989 who brought out a new version of AHP called the 'ReferenceAHP', to avoid RRP. This means that the weighting of the criteria is changed on each occasion when a different criterion is introduced or deleted. The underlying mathematical justification for this phenomenon is rank reversal local preferences are normalized, in the relative measurement mode, with a reduced level of hierarchy so that they add up to 1. As new alternatives are inserted or removed, Changes in local preferences will be influenced by other alternatives the result may therefore be a change in the end ranking of alternatives. Belton & Gear agrees with this statement as well [17]. In 2012, scholarly papers from an assessment of 61 scholarly papers on AHP methodologies and ranking reversals were carried out in 18 journals [18]. [19]A few studies conclude that the findings do not cover all the areas, the establishment of a new hybrid method for implementing and evaluating the results of studies will be proposed in future work, together with its proposal to evaluate every completed study. By using different methods that will be used in future work, a hybrid method will be integrated and used.

By using different methods that will be used in future work, a hybrid method will be integrated and used. Unlike conventional methods, COMET (Characteristic Object Method) uses a different approach, during initial investigations the phenomena of rank reversal have not been observed. To determine the measurement standard with a fuzzy reference model, a constant set of specially selected characteristic objects distinguishes the COMET model which is independent of the alternatives. The ability of COMET to detect multiple criteria models for the decision-makers, including a nonlinear multicriteria model, is its greatest advantage. [20] The COMET method was primarily designed to deal with actual value data in the first place. In several cases, it has proved difficult to define precisely the exact attributes of decision criteria. Therefore, in complicated decisions with data uncertainties, the use of intervals or fuzzy numbers should be substituted for numerical values.

In this research, as the solution to the AHP in the RRP problem, the advantage of COMET in analyzing characteristics through the categorization of characteristics objects with characteristic value and the classification of criteria and the consistent criteria will be added to AHP and combined with the pairwise comparison in AHP. The purpose of this research is to avoid and eliminate RRP in AHP and the impact on the case of digital marketing strategy decisions in private higher education in Palembang City, Indonesia.

II. Literature Review

A. The Correlation COMET and AHP

COMET and AHP have similarities in the pairwise comparison step. The unstable solution obtained in AHP is caused by the only discrete value of priority for all alternatives from the space of problem which COMET can provide the solution for it [21]. AHP is the best way to prioritize as it delivers highly accurate results for decisions that need to be made. Conventional AHP has scalability problems that can be solved with the modification of the AHP method structure [22]. AHP is also the method to have a coherent result, where to calculate the weights of the criteria, first, we need to establish the criteria' weight by comparing them in pairs [23]. With this strength of AHP combined with the advantage of COMET, AHP can be more powerful in avoiding and eliminating the RRP.

B. COMET and AHP Combination

To create a free RRP, the COMET method needs to be done step by step, the following Fig. 1 is the step of the COMET method [24][25].

Meanwhile, two main steps are carried out in the AHP method, Setting up the problem's hierarchy structure is an initial step, The next step is to give each grade of the hierarchy a nominal value, then set up a pairwise comparison matrix [12][26] The AHP hierarchy structure, and the form of the structure is as presented in Fig. 2.

From the COMET framework steps, and AHP, can be stated that COMET has the biggest contribution in the criteria modification which the adding of a characteristic value and characteristic object to the Criteria is the main key to eliminate the rank reversal. Before adding the numerical example, let's look at the combination of COMET and AHP named the C-AHP method which is depicted in the following Fig. 3.

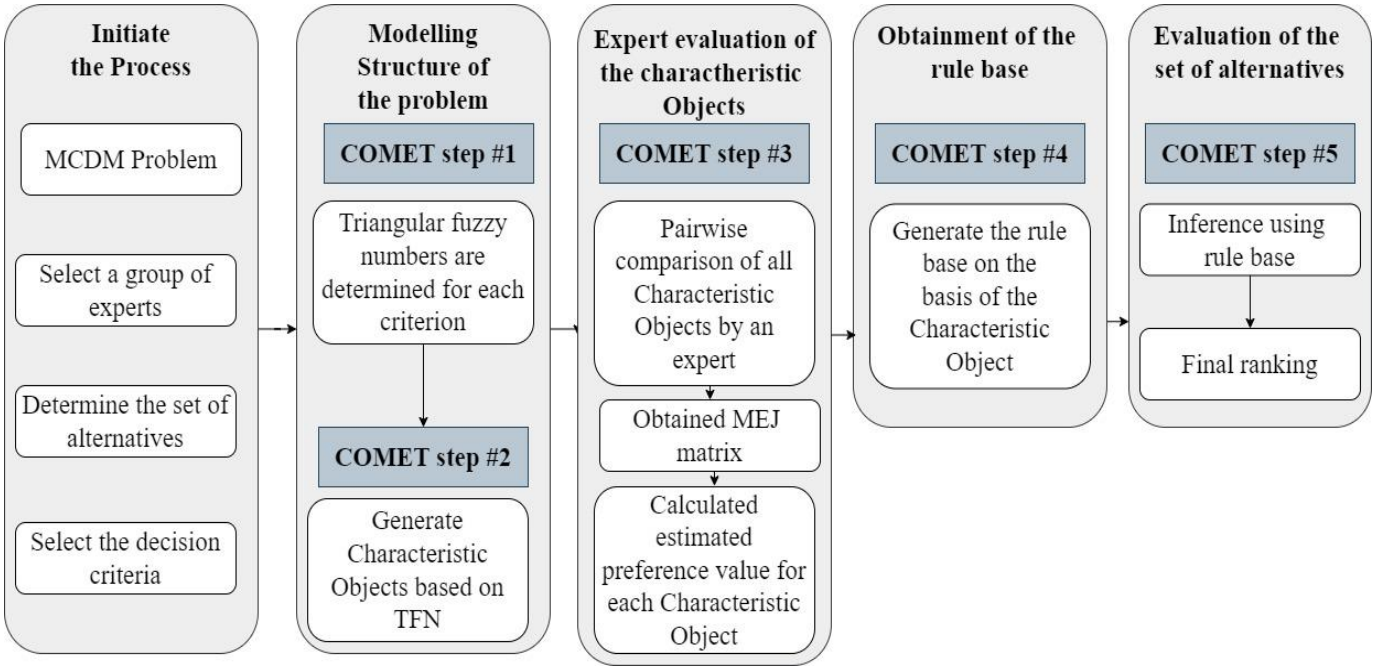


Fig. 1. COMET method model.

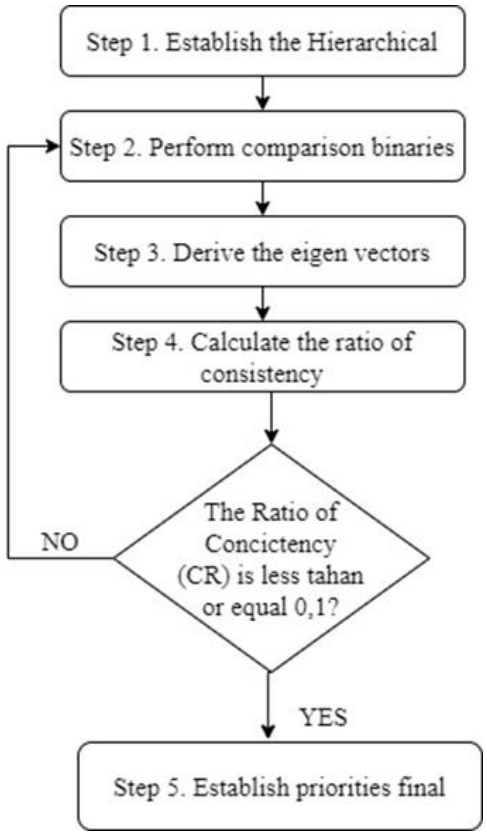


Fig. 2. The AHP method model.

C. Characteristic Object and Value

The characteristic object method (COMET) is based on the idea that a characteristic object is obtained as a combination of values characteristic to individual criteria, points that are regularly distributed in the problem state. By combining fuzzy theory set elements to define a decision model in the area of a problem, which is mainly designed for dealing with real-valued data [23][27]. In COMET, it is the first step to determine how many variants are associated with certain linguistic values that describe criteria, then, a characteristic variable is generated from values of vertices with particular fuzzy numbers [28]. Before the pairwise comparison is conducted, based on the distance from the nearest characteristic objects and their values, preference is given to each alternative [29][30]. For all these criteria, the domains and fuzzy numbers are determined. As a combination of the crisp values of all the fuzzy numbers, the characteristic objects are obtained.

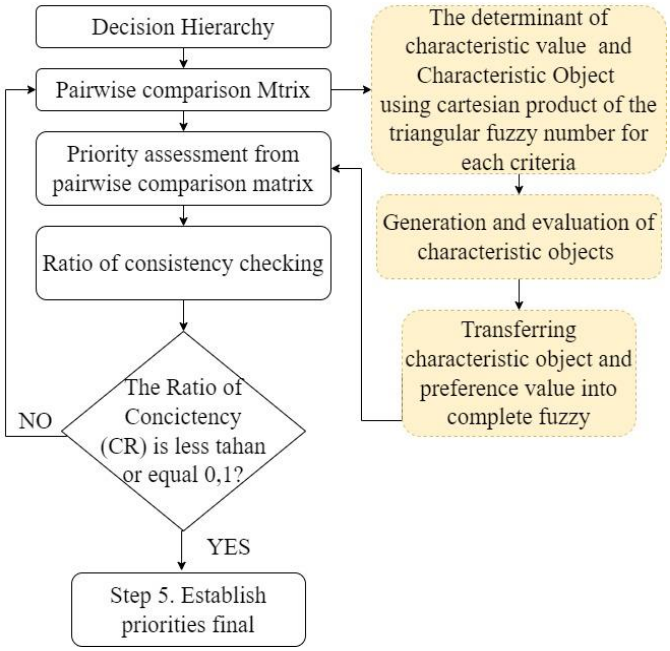


Fig. 3. The C-AHP method framework combination.

III. RESEARCH METHODOLOGY

The methodology of this research as shown in Fig. 4, started with the COMET method component analysis that gives the biggest part to reduce or remove rank reversal in decision-making. Formulation of COMET and AHP focusing on the characteristic value and characteristic object exist in the COMET method, and how its triangular fuzzy number affects the weight of the criteria in AHP. The combination of COMET and AHP is then generated into one framework combination form of those two methods which is later called C-AHP. The new framework, C-AHP, will be implemented in the numeric case, it is digital marketing for private higher education.

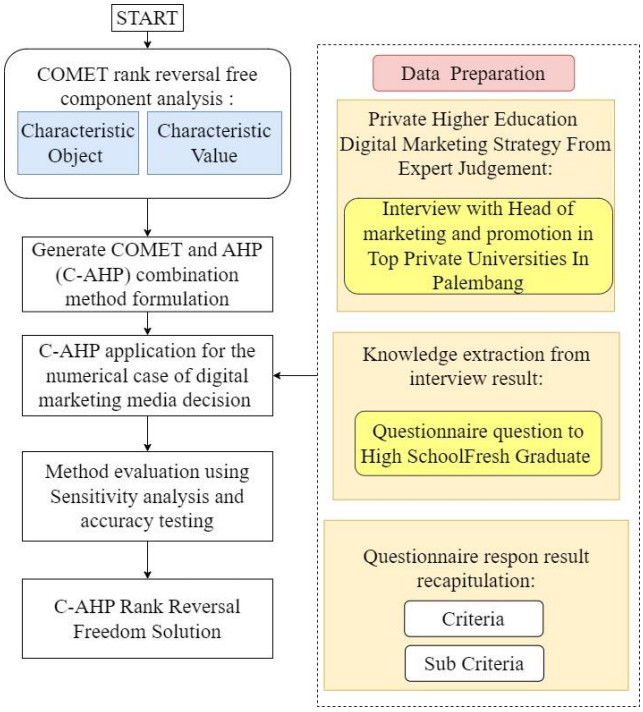


Fig. 4. Framework for C-AHP combination.

Data preparation for the higher educational marketing strategy started with interviewing the head of marketing and public relations in a private University, to collect the digital marketing strategy and criteria according to their wants and needs. After the interview result is a recap and the resulting knowledge about the digital marketing strategy and criteria, the next step is the question questionnaire will be made, which refers to the interview result and will be given to newly graduated high school student, to check whether the criteria is the same between those two parties. After the questionnaire is given to high school graduate students, the result will be a recap and the final criteria and sub-criteria for digital marketing will be established. To prove the rank reversal paradox avoidance in the C-AHP Method, the data from the questionnaire and interview results will be integrated and used in the C-AHP method for the numerical calculation. An additional step is added at the beginning of the C-AHP step after the criteria are determined, then will be building the characteristic value for each criterion. After defining the characteristic value, a characteristic object is created according to its characteristic value, using the formula as follows:

A. Characteristic Value of Criteria

In this concept, by using the number r of criteria to define the size of the problem and establish its dimensionality,

B. $C_1\ C_2\ ,...,C_r$.

The selection of the triangular fuzzy numbers for each criterion is made.

C_i, as example $C_{i1}\ C_{i2}\ ,...C_{ic}$

The result of this is as follows::

$$C_1\{C_{11},C_{12}\ ,...,C_{1c1}\}$$

$$C_2\{C_{11},C_{12}\ ,...,C_{1c1}\}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots..$$

$$C_r\{Cr_1,Cr_2\ ,...,C_{rcr}\}$$

C1, C2,..., Cr for all criteria, are numbers of fuzzy numbers. After obtaining the triangular fuzzy number, and then a linguistic model has been created.

C. Generating Characteristic Object (CO)

CO $C(C_1)$ x $C(C_2)$ x...x$C(C_r)$

The next step is to define all of the set values for a given characteristic object as a result.

$$CO_1 = \{C(C_{11}),C(C_{21}),\ldots,C(C_{r1})\}$$

$$CO_2 = \{C(C_{11}),C(C_{21}),\ldots,C(C_{r2})\}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$CO_t = \{C(C_{1c1}),C(C_{2C2}),\ldots,C(C_{rcr})\}$$

t is a number of CO [31]

$$t = \prod_{i=1}^{r} c_i$$

D. Pairwise Comparison of Characteristic Objects

Before we carry out an evaluation and a final ranking, a comparison of characteristic Objects between Characteristic Object (CO) for each criterion will be implemented. With the scale of:

$$\alpha \text{ is } = f(COi, COj) = \begin{cases} 0{,}0,\ f_{exp}(COi) < f_{exp}(COj) \\ 0{,}5,\ f_{exp}(COi) = f_{exp}(COj) \\ 1{,}0,\ f_{exp}(COi) > f_{exp}(COj) \end{cases}$$

This step is the normalization step that will make sure that each alternative and criteria have the same consistent numeric calculation even though there is a new alternative added or deleted.

E. Previous Related COMET and AHP Research

The previous works regarding to COMET and AHP, or one of the method, which related to the combination of those method, analyzed in this paper. The result of latest work related to this research is started in 2016 in the research about the rank reversal paradox in management decisions: the comparison of the AHP and COMET methods, the result is COMET is free of

rank reversal in compare to AHP on the Problem of Selecting Providers. COMET method is very easy to use, and the simplest models given reliable results. in 2020, intuitionistic fuzzy sets in multi-criteria group decision making problems using the characteristic objects method, resulting the case is academic problem of selection of the best mobile company using fuzzy set. In this research, on 2023, the case in digital marketing strategy decision for private higher education in South Sumatra, Palembang, Indonesia, combining COMET and AHP for the Rank Reversal Phenomenon free.

IV. C-AHP Implementation in Digital Marketing Private Higher Education Decision

A. Digital Marketing Private Higher Education Decision

Data sample for digital marketing private higher education is taken in two ways; first interview with the head of marketing and/or public relations in private higher education, University in Palembang City is done. The interview questions about digital marketing strategy are Table I:

TABLE I. Interview Questions

Question	Answers
Is there a specific target or niche targeted at the prospective student market segment?	SMA & SMK
Is the marketing strategy model used the same for all study programs/faculties?	All the same
How big is the comparison between the amount of conventional/offline marketing compared to digital / online at your university?	50% offline - 50% Online
Is there a strategy for when to use conventional marketing media and when to use digital marketing at your university?	No
What are the criteria or conditions to be considered when determining which digital marketing media to use? In meaning, what are the criteria for choosing marketing media at your university?	No
What are the strategies and/or media used for digital marketing strategies at your current university? is one of these lists - SEO - Content marketing - Paid search - Social media - Marketplaces - Paid advertising media, banner ads - Media chat - Emails - Influencer - Affiliate marketing	SEO FB IG TIKTOcK WA Blast Marketplace Paid Advertising
Which do you use most of all the digital marketing media mentioned earlier?	
Which digital marketing media has an impact in terms of: - Easily measurable results - Increasing customer loyalty, which ensures the most retention of customers - Cost-effectiveness - Conversion rate: the percentage of your website's visitors who do business on your company's page, actions they take during their visit and steps that lead them to make transactions	All the same
You agree that the needs and benefits of digital marketing in the field of education should be summarised as follows: t's measurable IEasy to Access impactful Instant Feedback System Brand awareness reach	Yes
Demographics, age or other characteristics are usually the criteria which you will be considering when it comes to a particular Digital Marketing.?	No we come to every school which accept us

As shown in the methodology, from the interview results, knowledge about the targeted and planned digital marketing strategy of a private higher education institution's Marketing Director or Public Relations and promotion Department is created. The next step is to collect data from the questionnaire with the question using the Likert scale, where the data of the questionnaire is taken from 325 respondents, who are high school graduate students in Palembang city. Using the Slovin formula, this research uses a population of 420 with the error of tolerance e=0,05 and using the Slovin formula:

$$n = \frac{N}{1+Ne^2} \quad (1)$$

The sample from the formula is 204 respondents, but from the questionnaire responses, 325 respondents fill and answer the questionnaire. So minimum sample is n=325. From this sample result, the validity and reliability test are conducted, and the result for validity with the tolerance of error is 0,05, resulting in the r count being bigger than the r table for every aspect of the question which means all the value is valid. The reliability test for the graduated high school student is shown in Table II and Table III:

TABLE II. Reliability Test Result

		N	%
Cases	Valid	325	100,0
	Excluded[a]	0	0,0
	Total	325	100,0

TABLE III. Reliability Cronbach Result

Cronbach's Alpha	N of Items
0,732	39

To the answers of the interviewed person, it is concluded that the criteria and sub-criteria for the digital marketing strategy in the decision of which media to choose, is as shown in Table IV:

TABLE IV. Criteria and Sub Criteria

Criteria	Name	Sub Criteria
C1	Advertising Type	Video, Photo, short video
C2	Advertising content	Program explanation, campus facilities, tuition fee and scholarship, campus performance
C3	Marketing Period	Anytime, April-june, October-December, Januari-March, July-September

From the questionnaire and interview result comparison, the alternatives for this research are Google ads, Instagram, WhatsApp blast, and YouTube.

The C-AHP method will be used to avoid rank reversal paradox in this digital marketing for private higher education decision, following the C-AHP method framework combination in Fig. 4, with these steps:

1) Problem and decision hierarchy: As the first step in C-AHP still follows the AHP stages, the problem domain definition starts with the hierarchical structure format, as shown in Fig. 8.

2) Determination of characteristic value: To generate a characteristic object, characteristic value is the first thing to settle. Characteristic value is different for each criterion it depends on the data and assumption. In the agreed and very agreed choice, use the triangular fuzzy number and the number of answers chosen by the respondent, The characteristic value of the criteria is set out in this case as presented in Table V:

TABLE V. CHARACTERISTIC VALUE

Criteria	Sub Criteria	Value
C_1 (Advertising type)	Video	65
	Photo	92
	Short Video	193
C_2(Advertising content)	Campus Performance	47
	Scholarship	48
	Tuition Fee	61
	Campus Facilities	67
	Program Explanation	70
C_3 (Marketing Period)	July-September	11
	October-December	17
	Januari-March	18
	April-june	21
	Anytime	32

There is a minimum possible value and a maximum possible value for each characteristic value. For instance, video 65 is the minimum possible value for advertising type and the highest possible value can be the shortest video 193, and a medium value is photo 92. After the characteristic value Triangular Fuzzy number for each criterion is conducted next step is to result in the linguistic model of the characteristic value which shown in Fig. 5, Fig. 6, and Fig. 7.

C_1 Advertising type, linguistic model

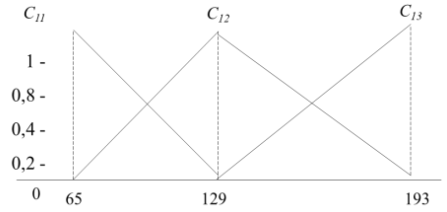


Fig. 5. Characteristic value type.

C_2 Advertising content, linguistic model

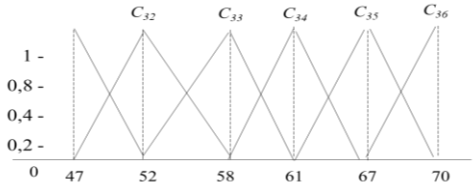


Fig. 6. Characteristic value advertising content.

C_3 Period, linguistic model

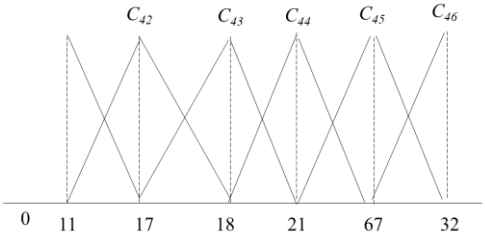


Fig. 7. Characteristic value period.

B. Generate Characteristic Object Value

From the characteristic value, the characteristic object can be generated using the formula to generate, and then from the characteristic object, with the limitations of maximum possible and minimum possible it will limit the additional aspect of criteria, and when a ranking result of digital marketing is obtaining, and another alternative in example TikTok occurred, the rank result difference will not be too far because the limitations of the character of the criteria are works in C-AHP model since the comparison of the criteria in pairs with another criterion. After the Criterion value is determined, the pairwise comparison with the generated score 0, 0,5, and 1 is entered in the criteria and sub-criteria value and the result of the final ranking is obtained.

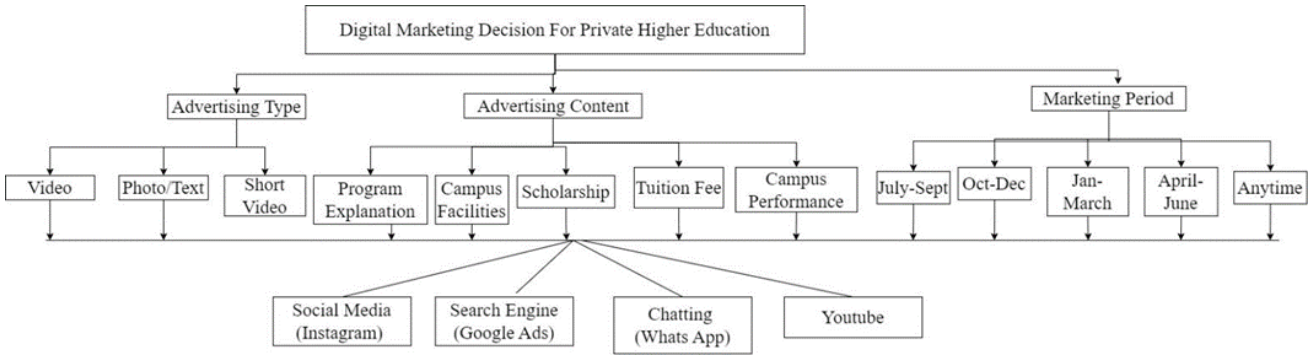


Fig. 8. Decision hierarchy digital marketing strategy.

V. Result and Discussion

From the C-AHP numeric calculation, with additional characteristic value and object in the criteria formulation, in the digital marketing for higher education case study, the Consistency Ratio (CR) results for criteria and sub-criteria is consistent with the score of Criteria is -0,66 and sub-criteria is -1,00 for C1, -0,50 for C2, and -0,67 for C3. This CR result shows because the result is under 1,00 shows that it is acceptable to implement. After all, it is stable and reliable. This research result has shown that C-AHP is reliable in keeping the criteria value stable even when the changes of alternatives are done. The addition of the same alternative and the reducing alternatives is done, and the result is the rank not changing for A1, A2, A3 and A4 as the alternative for the digital marketing higher education strategy case. The rank of the alternative remains the same. The case study has proven that C-AHP in the digital marketing strategy is affected by the characteristic object and value generation. This statement is proven by the accuracy and sensitivity analysis conducted on the variable chosen as the characteristic object in the criteria and sub-criteria, which changed with the adding of one point, and deduction of one point and the result is the sensitivity analysis acceptable. Adding a point to this research in the future will be the implementation of C-AHP in another case outside digital marketing for higher education strategy. As the same with the previous research conducted, the Rank Reversal phenomenon is eliminated for the comparison with other free rank reversal methods already conducted but there is no test to the result in the previous research using accuracy and sensitivity analysis.

VI. Conclusion

The rank reversal paradox in AHP with the combination of COMET and C-AHP was able to reduce the RRP in the decision to the digital marketing strategy for private higher education in Palembang city. The key combination for C-AHP is characteristic value and characteristic object which brought new weight to the criteria and sub-criteria in decision making, and also when generated it brings limitations to any addition or changes of the alternatives in the digital marketing strategy for private higher education.

Future studies can be conducted for the specific possibility of C-AHP implementation in another case study and can also be conducted for the same case with changes in the weight of the criteria from characteristic objects and values. The decision in this research is conducted by an individual; the future research can use C-AHP for the group decision support system in the group decision maker.

References

[1] R. F. de F. Aires dan L. Ferreira, “The rank reversal problem in multi-criteria decision making: A literature review,” Pesqui. Operacional, vol. 38, no. 2, hal. 331–362, 2018, doi: 10.1590/0101-7438.2018.038.02.0331.

[2] S. Mufazzal dan S. M. Muzakkir, “A new multi-criterion decision making (MCDM) method based on proximity indexed value for minimizing rank reversals,” Comput. Ind. Eng., vol. 119, no. Mcdm, hal. 427–438, 2018, doi: 10.1016/j.cie.2018.03.045.

[3] S. H. Mousavi-Nasab dan A. Sotoudeh-Anvari, “A new multi-criteria decision making approach for sustainable material selection problem: A critical study on rank reversal problem,” J. Clean. Prod., vol. 182, hal. 466–484, 2018, doi: 10.1016/j.jclepro.2018.02.062.

[4] R. F. de F. Aires dan L. Ferreira, “A new approach to avoid rank reversal cases in the TOPSIS method,” Comput. Ind. Eng., vol. 132, no. August 2018, hal. 84–97, 2019, doi: 10.1016/j.cie.2019.04.023.

[5] A. A. Al Salem dan A. Awasthi, “Investigating rank reversal in reciprocal fuzzy preference relation based on additive consistency: Causes and solutions,” Comput. Ind. Eng., vol. 115, hal. 573–581, 2018, doi: 10.1016/j.cie.2017.11.027.

[6] A. Majumdar, M. K. Tiwari, A. Agarwal, dan K. Prajapat, “A new case of rank reversal in analytic hierarchy process due to aggregation of cost and benefit criteria,” Oper. Res. Perspect., vol. 8, no. March, hal. 100185, 2021, doi: 10.1016/j.orp.2021.100185.

[7] K. Singh, N. Naicker, dan M. Rajkoomar, “Selection of Learning Apps to Promote Critical Thinking in Programming Students using Fuzzy TOPSIS,” Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 10, hal. 383–392, 2021, doi: 10.14569/IJACSA.2021.0121042.

[8] A. S. Abdelaziz, H. Harb, A. Zaghloul, dan A. Salem, “An Enhanced MCDM Model for Cloud Service Provider Selection,” Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 2, hal. 70–77, 2023, doi: 10.14569/IJACSA.2023.0140209.

[9] J. A. Alonso dan M. T. Lamata, “Consistency in the analytic hierarchy process: A new approach,” Int. J. Uncertainty, Fuzziness Knowldege-Based Syst., vol. 14, no. 4, hal. 445–459, 2006, doi: 10.1142/S0218488506004114.

[10] H. M. Osman, M. M. Singh, A. R. M. Shariff, A. A. Bakar, dan M. S. Plasencia, “Enhanced Analytical Hierarchy process for U-Learning with Near Field Communication (NFC) technology,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 12, hal. 281–290, 2018, doi: 10.14569/IJACSA.2018.091241.

[11] T. L. Saaty, “Analytic Hierarchy Process,” Encycl. Biostat., 2005, doi: 10.1002/0470011815.b2a4a002.

[12] H. Business dan S. Sdn, “Decision Making Using the Analytic Hierarchy Process (AHP); A Step by Step Approach,” vol. 2, 2017.

[13] S. Kaur, Y. Singh, dan N. Kaur, “Applications of Multi-criteria Decision Making in Software Engineering,” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 7, 2016, doi: 10.14569/ijacsa.2016.070765.

[14] A. Anggrawan, Mayadi, C. Satria, dan L. G. R. Putra, “Scholarship Recipients Recommendation System Using AHP and Moora Methods,” Int. J. Intell. Eng. Syst., vol. 15, no. 2, hal. 260–275, 2022, doi: 10.22266/ijies2022.0430.24.

[15] J. Barzilai dan B. Golany, “Ahp Rank Reversal , Normalization And Aggregation Rules,” vol. 5986, no. February, 2017, doi: 10.1080/03155986.1994.11732238.

[16] E. Triantaphyllou, “Two new cases of rank reversals when the AHP and some of its additive variants are used that do not occur with the multiplicative AHP,” J. Multi-Criteria Decis. Anal., vol. 10, no. 1, hal. 11–25, 2001, doi: 10.1002/mcda.284.

[17] S. Schenkerman, “Avoiding rank reversal in AHP decision-support models,” Eur. J. Oper. Res., vol. 74, no. 3, hal. 407–419, 1994, doi: 10.1016/0377-2217(94)90220-8.

[18] H. Maleki dan S. Zahir, “A comprehensive literature review of the rank reversal phenomenon in the analytic hierarchy process,” J. Multi-Criteria Decis. Anal., vol. 20, no. 3–4, hal. 141–155, 2013, doi: 10.1002/mcda.1479.

[19] S. Ramli, H. Mohamed, dan Z. Muda, “Determinants of interface criteria learning technology for disabled learner using analytical hierarchy process,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 1, hal. 518–523, 2020, doi: 10.14569/ijacsa.2020.0110164.

[20] A. Piegat dan W. Sałabun, “Identification of a Multicriteria Decision-Making Model Using the Characteristic Objects Method,” Appl. Comput. Intell. Soft Comput., vol. 2014, hal. 1–14, 2014, doi: 10.1155/2014/536492.

[21] W. Sałabun, P. Ziemba, dan J. Wątróbski, “The rank reversals paradox in management decisions: The comparison of the AHP and COMET methods,” Smart Innov. Syst. Technol., vol. 56, hal. 181–191, 2016, doi: 10.1007/978-3-319-39630-9_15.

[22] A. Darko, A. P. C. Chan, E. E. Ameyaw, E. K. Owusu, E. Pärn, dan D. J. Edwards, “Review of application of analytic hierarchy process (AHP) in construction,” Int. J. Constr. Manag., vol. 19, no. 5, hal. 436–452, 2019, doi: 10.1080/15623599.2018.1452098.

[23] W. Sałabun, A. Karczmarczyk, J. Wątróbski, dan J. Jankowski, "Handling Data Uncertainty in Decision Making with COMET," Proc. 2018 IEEE Symp. Ser. Comput. Intell. SSCI 2018, no. Mcdm, hal. 1478–1484, 2019, doi: 10.1109/SSCI.2018.8628934.

[24] K. Palczewski dan W. Sałabun, "Identification of the football teams assessment model using the COMET method," Procedia Comput. Sci., vol. 159, hal. 2491–2501, 2019, doi: 10.1016/j.procs.2019.09.424.

[25] W. Sałabun dan A. Karczmarczyk, "Using the COMET Method in the sustainable City Transport Problem: An Empirical study of the Electric Powered Cars," Procedia Comput. Sci., vol. 126, hal. 2248–2260, 2018, doi: 10.1016/j.procS.2018.07.224.

[26] V. D. B. Huynh, P. Van Nguyen, Q. L. H. T. T. Nguyen, dan P. T. Nguyen, "Application of Fuzzy Analytical Hierarchy Process based on Geometric Mean Method to prioritize social capital network indicators," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 12, hal. 182–186, 2018, doi: 10.14569/IJACSA.2018.091227.

[27] W. Sałabun, A. Karczmarczyk, J. Wątróbski, dan J. Jankowski, "Handling Data Uncertainty in Decision Making with COMET," Proc. 2018 IEEE Symp. Ser. Comput. Intell. SSCI 2018, hal. 1478–1484, 2019, doi: 10.1109/SSCI.2018.8628934.

[28] W. Sałabun, J. Watróbski, dan A. Shekhovtsov, "Are MCDA methods benchmarkable? A comparative study of TOPSIS, VIKOR, COPRAS, and PROMETHEE II methods," Symmetry (Basel)., vol. 12, no. 9, hal. 1–56, 2020, doi: 10.3390/SYM12091549.

[29] W. Sałabun, "Reduction in the Number of Comparisons Required to Create Matrix of Expert Judgment in the Comet Method," Manag. Prod. Eng. Rev., vol. 5, no. 3, hal. 62–69, 2014, doi: 10.2478/mper-2014-0028.

[30] B. Kizielewicz dan J. Kolodziejczyk, "Effects of the selection of characteristic values on the accuracy of results in the COMET method," Procedia Comput. Sci., vol. 176, hal. 3581–3590, 2020, doi: 10.1016/j.procs.2020.09.028.

[31] W. Sałabun, "The characteristic objects method: A new distance-based approach to multicriteria decision-making problems," J. Multi-Criteria Decis. Anal., vol. 22, no. 1–2, hal. 37–50, 2015, doi: 10.1002/mcda.1525.

Automated Detection and Classification of Soccer Field Objects using YOLOv7 and Computer Vision Techniques

Jafar AbuKhait, Murad Alaqtash, Ahmad Aljaafreh, Waleed Othman
Dept. of Computer and Communications Engineering
Tafila Technical University
Tafila, Jordan

***Abstract*—In the last two decades, many technologies have been deployed and utilized in Soccer games (Football) as a result to the huge investment of Federation of International Football Association (FIFA). These technologies aim to monitor and track all soccer match objects including players and the ball itself in order to measure the player performance, and tracking the players' positions and movements at the field. Latest emerging artificial intelligence and computer vision techniques are being used recently in many systems and deployed in different scenarios. Identifying all field objects automatically has to be the first step in the monitoring process of soccer games. In this paper, we are proposing an automated system that has the ability to detect and track the ball and to detect and classify players and referees on the soccer field. The proposed system implements a detection model using a real-time object detection model YOLOv7 to detect the ball and all humans on the field after building a labeled dataset of 1300 different soccer game frames. It also deploys Improved Color Coherence Vector (ICCV) features to classify all humans on the field to five classes (Team1, Team2, Goalkeeper1, Goalkeeper2, and Referee) using K-Nearest Neighbor algorithm. The proposed system has achieved high accuracy in both the detection and classification modules.**

***Keywords*—Soccer game; football; YOLOv7; human detection and classification; ball detection; improved color coherence vector**

I. INTRODUCTION

In the realm of sports, where every moment carries significance, the automated computer vision detection and recognition of soccer match field objects has emerged as a pivotal technological advancement [1]. Soccer, often referred to as football in many parts of the world, stands as one of the most globally celebrated and passionately played sports. Over the years, it has not only evolved in terms of gameplay but has also embraced technology to analyze and elevate the sport to new heights. The integration of automated computer vision has ushered in a transformative era, fundamentally reshaping how we perceive and comprehend soccer matches.

The soccer field itself serves as a dynamic canvas, where players, the ball, and various other elements such as goalposts, corner flags, and boundary lines converge to create a complex and fast-paced spectacle. Traditionally, the task of monitoring and dissecting these elements fell to human operators, a process fraught with potential errors and subjectivity. However, with the advent of automated computer vision, we have witnessed a profound shift in our ability to capture, process, and leverage data from soccer matches [2]. This technology empowers us to identify and track field objects in real-time, providing invaluable insights to coaches, players, analysts, and fervent fans.

In this context, this paper delves into the profound significance of automated computer vision in the detection and recognition of soccer match field objects. We explore the tangible applications of this technology, its impact on game analysis, player performance evaluation, automatic offside detection, and fan engagement. Furthermore, we delve into how it is poised to redefine the future of soccer as we know it. Through this exploration, it becomes increasingly evident that automated computer vision is not just a tool but a transformative force that is redefining the very essence of soccer analysis and appreciation.

In the past two decades, the world of soccer (or football) has witnessed a profound transformation, fueled by substantial investments from organizations like the Federation of International Football Association (FIFA) [3]. These investments have ushered in a new era of technology-driven enhancements within soccer games, aimed at monitoring and tracking various aspects of the game, including player performance and positional data [4]. Recent advancements in artificial intelligence and computer vision techniques have played a pivotal role in this transformation.

The initial step towards automating the monitoring process of soccer matches involves the automatic detection and classification of all relevant objects on the field. In this context, several Convolutional Neural Network (CNN) architectures were suggested and deployed to detect the ball and the players on the soccer field [5, 6]. In addition, several research works have addressed the detection of soccer events, ball events, actions on the soccer game, and team tactics estimations [7, 8].

In general, several researchers have addressed the detection process of soccer field objects for various applications but, it is also important to classify these objects to ease the monitoring process of each soccer team player. In this context, the class of each human on the soccer field should be determined to enable tracking of individual players and team movements.

In this paper, we present an automated system designed to detect and track the soccer ball and classify players and

referees on the field using state-of-the-art techniques. In this work, we aim to: 1) construct an annotated detection dataset that consists of 1300 soccer game matches' images; 2) detect soccer field objects, ball and humans using YOLOv7; 3) classify every human on the soccer field to Team1, Team2, Goalkeeper1, Goalkeeper2, and Referee using Color Coherence Vector and k-NN classifier; and 4) improve the detection precision and the classification accuracy by implementing a cascaded detection and classification systems. Overall, this system aims to provide an accurate and efficient method for detecting and classifying soccer game objects, which can be beneficial for coaches, broadcasters, and analysts in different computer vision applications such as team performance monitoring, automated offside detection and tactics estimation.

This paper is organized as follows. Section II provides a theoretical background of the techniques being used. Section III demonstrates the architecture of the proposed system. Section IV presents the experimental results and discussion. Finally, conclusions are drawn in Section V.

II. THEORETICAL BACKGROUND

A. YOLO

You Only Look Once (YOLO) is convolutional neural network architecture for object detection in one shot [9, 10]. YOLO partitions the input image into N grids, each with equal dimensions. Each grid is responsible for the detection and localization of the object that it contains. In general, YOLO networks extract features through a backbone. The extracted features are combined and mixed in the neck, and then they are passed along to the head of the network to predict the locations and classes of objects around which bounding boxes should be drawn [11, 12].

The YOLOv7 algorithm, an upgraded version of YOLO object detectors, surpasses all known object detectors in both speed and accuracy [13]. YOLOv7 was trained only on MS COCO dataset from scratch without using any other datasets or pre-trained weights. It improved real time object detection accuracy without increasing the inference cost.

YOLOv7 has extended efficient layer aggregation networks (E-ELAN). It also has model scaling for concatenation-based models. The YOLOv7 algorithm also uses a technique called anchor boxes to improve the accuracy of object detection. Anchor boxes are pre-defined shapes that the algorithm uses to predict the location of objects in an image. By using anchor boxes, the algorithm is able to detect objects of different sizes and shapes with greater accuracy.

B. Improved Color Coherence Vector

Color Coherence Vector (CCV) is a color feature extractor that encodes information about color spatial distribution. It classifies each pixel in the image as either coherent or incoherent. Coherent pixels belong to a big connected component (CC) while incoherent pixels belong to a small connected component. CCV aims to build a low dimensional representation of the image through the following steps [14]:

1) Blur the image by averaging.

2) Quantize the image colors into n distinct colors.

3) Classify each pixel either as coherent or incoherent by:

a) Finding the connected components for each quantized color.

b) Determining the Tau's value which is typically about 1% of image size. Pixels are considered coherent if they belong to any connected component with number of pixels are more than or equal to tau.

4) For each color compute two values: α which is the number of coherent pixels, and β which is the number of incoherent pixels.

Improved Color Coherence Vector (ICCV) has more spatial information with respect to CCV and thus; it is more efficient in comparing image contents [15]. In addition to (α, β) pairs that were computed by CCV, the mean of position coordinates (rows and columns) for the maximum connected region in the coherent pixels (γ) is computed using ICCV. Each quantized color would be described accordingly by four values in the form of (α, β, γ_x, γ_y). The size of feature vector would be the number of quantized colors multiplied by 4.

C. K-Nearest Neighbor Algorithm

The k-nearest neighbor algorithm (k-NN) is one of the simplest machine learning algorithms. Any test object is assigned to its nearest neighbors' class by adopting majority voting. Nearest neighbors are determined by measuring a distance metric which could be the Euclidean distance, the hamming distance, or the correlation. Number of nearest neighbors is defined by K which is a problem dependent parameter [16, 17]. k-NN has two stages; determining the nearest k neighbors and determining the class by majority voting.

III. THE PROPOSED SYSTEM

The proposed system shown in Fig. 1 is composed of three modules:

- Dataset Preparation: in which soccer game field images of on-going games are collected, pre-processed, annotated, and augmented to build the dataset for training, evaluating, and testing the YOLO detection model.
- Human and Ball Detection: in which a YOLOv7 deep learning model is selected and trained on both the training and validation datasets and then, evaluated on the test dataset. The detection model has the capability of detecting the ball and all humans on the soccer field
- Humans' Classification: in which k-NN classifier is being used to classify detected humans to team1, team2, goal keeper, and referee using Color features of detected human. Human templates of team1, team2, goal keeper, and referee are used in the classification based on Color Coherence Vector (CCV).

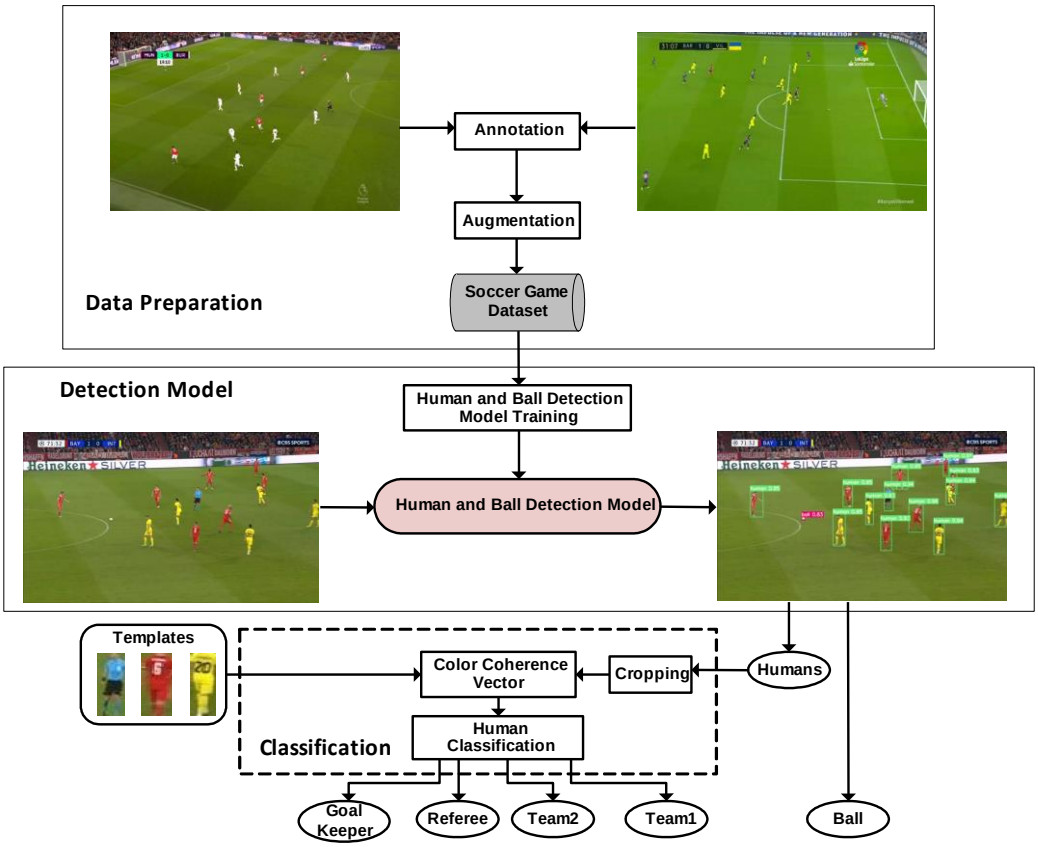


Fig. 1. The proposed system.

A. Dataset Preparation

In this module, a labeled dataset of 1300 different soccer match images was constructed. This dataset will be used for training, validation, and testing the Human and Ball detection model.

1) Image collection: A set of 1300 different images of different soccer matches was constructed. 810 images were obtained from multiple videos of soccer game matches and 490 images were obtained from the online dataset at [18]. All images were captured from a single camera position covering one half of the field. Several images were selected for each soccer game. Fig. 2 shows samples of soccer matches' images.

2) Annotation, preprocessing and augmentation: In this step, the collected images were uploaded to Roboflow platform, where it was labeled into two classes: human and ball using Roboflow's annotation tool. Fig. 3 shows some annotations of ball and humans on the original images.

The annotation task of the original 1300 images has resulted into 25636 different labels of both ball and human labels as shown in Table I.

TABLE I. SUMMARY OF ANNOTATED IMAGES

Number of Images	Annotations	Ball Annotation	Human Annotation
1300	25636	1174	24462

Fig. 2. Sample images of soccer game matches.

Fig. 3. Sample images of soccer game matches

After labelling, some preprocessing and augmentation tasks of the images were applied as follows:

- Preprocessing:
  - Auto Orient: Applied
  - Resize: Fit within 640x640
- Augmentations:
  - Grayscale: Apply to 25% of images
  - Saturation: Between -25% and +25%
  - Brightness: Between -30% and +30%

After Augmentations, a set of 2203 images were obtained. These images were splitted into 83% (1826), 12% (273), and 5% (104) for training, validation, and testing, respectively.

B. Human and Ball Detection Model

In this module, YOLOv7 algorithm is used for human and object detection. Model training was achieved on Google Colab notebook. Once the model was trained, it was tested on a separate set of validation images to evaluate its performance.

1) Model training and evaluation: The model was trained using YOLOv7. This model is evaluated for detecting two classes: human and ball. Two epoch choices were used to train the model; 100 and 180.

The model detection performance was evaluated using mean average precision (mAP), recall and precision. The evaluation metrics that were used to evaluate the model are explained as follows:

Precision is a measure of a network's ability to accurately identify targets at a single threshold, calculated by:

$$\text{Precision} = \frac{\text{Tp}}{\text{Tp+Fp}} \quad (1)$$

Recall is a measure of the network's ability to detect its target, calculated by:

$$\text{Recall} = \frac{\text{Tp}}{\text{Tp+Fn}} \quad (2)$$

Where:

- Tp: are the Bounding Boxes (BB) that the intersection over union (IoU) with the ground truth (GT) is above 0.5.
- Fp: two cases (a) BB that the IoU with GT is below 0.5 (b) the BB that have IoU with a GT that has already been detected.
- Tn: there are not true negative, the image is expected to contain at least one object.

Fn: images containing an object were the method failed to produce a BB.

Intersection over Union (IoU) is a method used to compare two arbitrary shapes, i.e., object widths, heights, and location of two boxes into the original region. This will evaluate the precision of the object detector on particular dataset [19] as in (3). Fig. 4 shows how IoU is calculated diagrammatically.

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \quad (3)$$

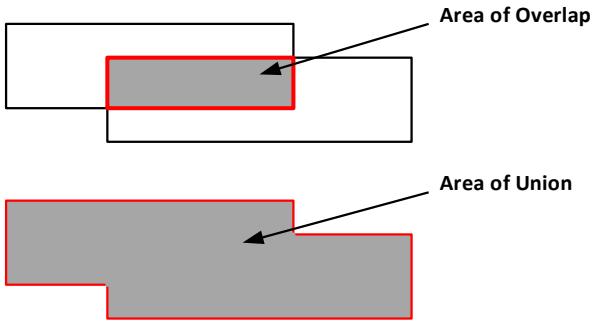


Fig. 4. Diagrammatic example intersection over union (IoU) calculation.

Average precision is a method combining recall and precision for the entire ranking. It is the average of precision in a single ranking [20].

$$AP = \frac{1}{|Class|}\sum_{c \in class} \frac{TP(c)}{TP(c)+FP(c)} \quad (4)$$

Mean average precision (mAP) is the average of precision values at the rank where there is a relevant document [21]. It is calculated from precision, recall and interception over union IOU.

$$mAP = \frac{AP}{Total\ number\ of\ class} \quad (5)$$

C. Humans' Classification

This module classifies the detected humans in the previous module to team1, team2, goal keeper, and referee using shirt

colors of detected human. Color Coherence Vector (CCV) is used to describe the color features and k-NN is deployed for classification. The classification has been achieved by comparing each detected human with stored templates of humans from the same soccer match.

1) Template preparation and preprocessing: Template Preparation and Preprocessing have been done prior to features extraction and classification. At first, manual cropping of humans on the soccer match images was achieved for image group of each distinct soccer match. Five suitable templates of team1, team2, goalkeeper1, goalkeeper2 and referee were selected. Each five templates were obtained from different image frames of the same soccer match to choose the best body orientation of each class. Each single template represents the body area of the human excluding the head and lower part of leg to concentrate on distinct color features for classification. All templates were resized to 25x50. Fig. 5 shows the five templates of humans selected from different images of the same soccer match.

Next, each detected human in the detection module is cropped automatically to exclude most of the soccer field background and body parts that haven't distinct color features. This process has been done by cropping 15% from each side of the detected humans' area as shown in Fig. 6.

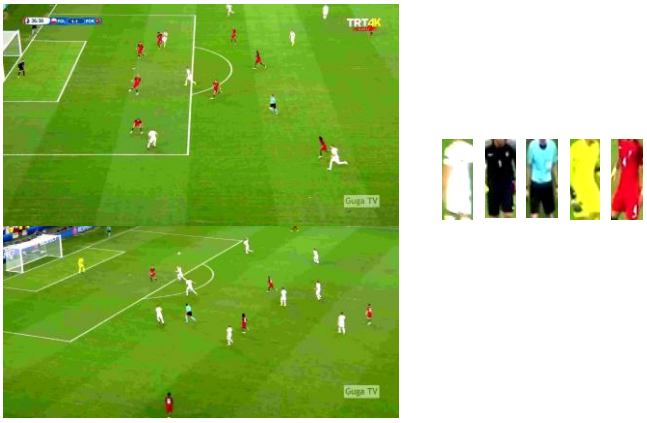


Fig. 5. Example of template preparation.

Fig. 6. Examples of cropping detected humans' images.

2) Color feature extraction using ICCV: Color features of human templates and all cropped images of all detected humans are extracted using Improved Color Coherence Vector (ICCV). The output of this stage is a feature vector of size 64x1 based on the following steps:

- Quantize the color-space into 16 distinct colors.
- Classify each pixel either as coherent or incoherent by finding the connected components for each quantized color and determining the tau's value to be 5% of image's size. Any connected component with number of pixels more than or equal to tau then its pixels are considered coherent otherwise they are incoherent.
- For each color, compute the number of coherent pixels (α), the number of incoherent pixels (β), and the mean of position coordinates (rows and columns) for the maximum connected region in the coherent pixels (γ).

Each quantized color would be described accordingly by four values in the form of (α, β, γx, γy). For the 16 colors, we would gain a feature vector of size 64x1.

3) Classification using k-NN: Each detected human would be classified to team1, team2, goalkeeper1, goalkeeper2 or referee based on 3-NN classifier. Nearest neighbors are determined by measuring the Euclidean distance and determining the class by majority voting.

Classification accuracy can be calculated according to the following formula:

$$\text{Accuracy} = \frac{\text{Tp+Tn}}{\text{All Elements}} \quad (6)$$

Where Tp and Tn are the elements correctly classified by the model.

IV. Results and Discussion

In this section, we demonstrate the experimental results of both the detection and classification modules. We have tested the detection model on 104 images that were picked randomly from various soccer matches. In the classification module, we have achieved the testing on five different matches because we use color features which are different on each single match. A total number of 51 different images have been selected to validate the classification results.

A. Detection Results

In this section, we evaluate the performance of the YOLOv7 detection model, which plays a crucial role in automatically identifying and tracking soccer field objects such as the ball and players. The results provide insights into the model's accuracy and effectiveness in object detection.

Fig. 7 serves as a comprehensive illustration of the YOLOv7 model's performance in identifying and localizing objects of interest, including both players and the ball. This figure utilizes three distinct metrics—Accuracy, Precision, and mAP@0.5—to evaluate the model's precision in object localization and its proficiency in correctly classifying objects.

Fig. 8 presents a valuable snapshot of the quantitative metrics used to evaluate the performance of the detection model throughout the training process. These metrics include precision, recall, and mean average precision (mAP@0.5), which provide insights into the model's effectiveness in detecting objects of interest in soccer matches.

Precision measures the accuracy of positive predictions made by the model. It is a crucial metric for object detection, as it assesses the model's ability to correctly identify objects without generating too many false positives. Recall evaluates the model's ability to detect all relevant objects in the dataset, minimizing false negatives. It is particularly important in ensuring that no objects of interest are missed. mAP@0.5 is a comprehensive metric that combines precision and recall across different object classes. It considers precision-recall trade-offs and provides an aggregate assessment of the model's performance.

The confusion matrix, as shown in Fig. 9, provides a detailed breakdown of the model's performance, including true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each object class (e.g., ball, humans and background).

Fig. 10 offers an insightful comparison between different models' training using varying numbers of training epochs (specifically, 100 and 180). The key takeaway from this figure is that it underscores the importance of training the model beyond 100 epochs for achieving optimal performance.

To perform model testing, we loaded the saved weights of the trained model and passed the test dataset through it. Fig. 11 shows some examples of detection model Inference.

The results of the model testing showed very good accuracy and performance in detecting objects of interest in the images. Overall, the model testing was successful in demonstrating the accuracy and effectiveness of the trained model in detecting objects of interest in the images.

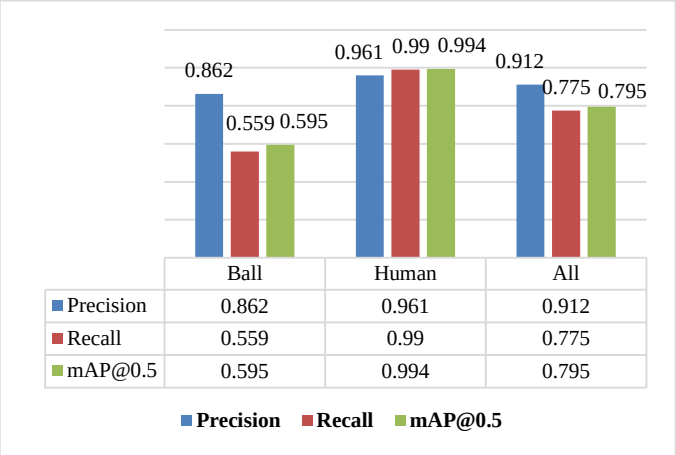
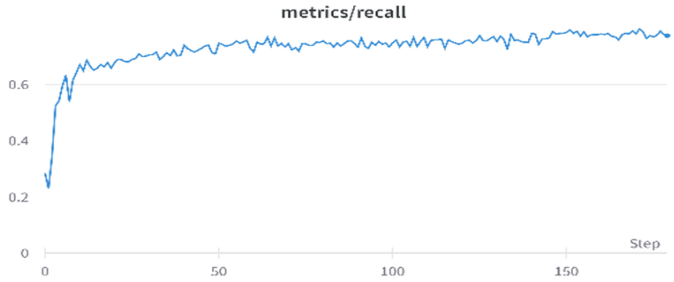
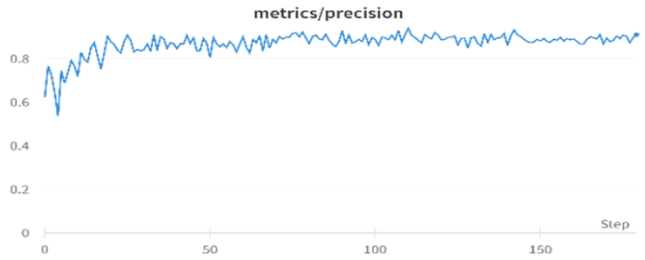


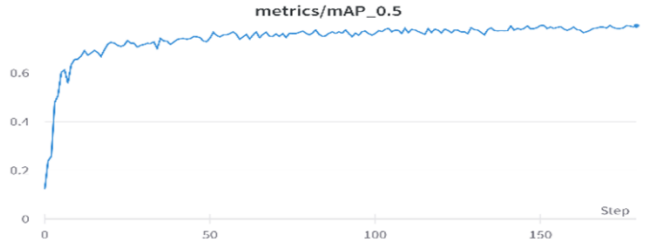
Fig. 7. YOLOv7 model's performance.



(a)



(b)



(c)

Fig. 8. Performance of the detection model throughout the training process: a) recall; b) precision, c) mAP@0.5.

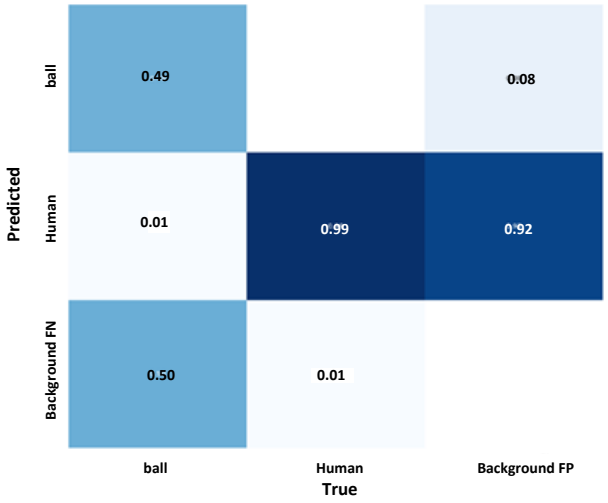


Fig. 9. Confusion Matrix of the detection model.

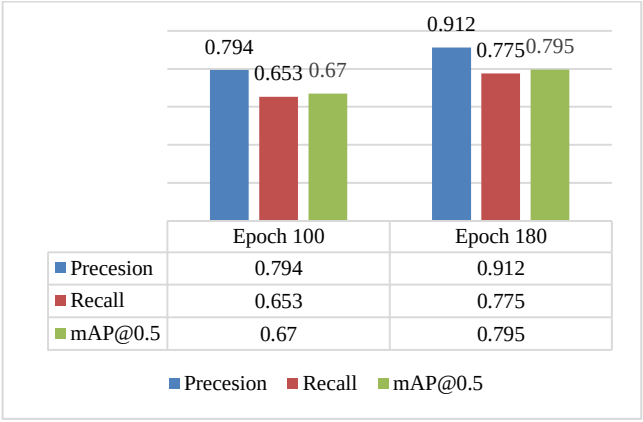


Fig. 10. Comparison between two different models using different epochs.

Fig. 11. Model Inference examples

B. Classification Results

The proposed human classification module has been tested on 5 different soccer matches to validate its performance for different color variances among humans' shirts on each single match. Fig. 12 shows cropped human images for each one of the five matches. These images are used in the discussions of the classification results. For each match, 11 to 14 frames have been selected to obtain 2 human templates for each one of the five human categories (team1, team2, goalkeeper1, goalkeeper2, and referee) and the rest detected humans are used to test the classification accuracy. A total number of 827 humans is existed among all selected soccer matches' frames. Table II shows the exact division of these images cross the five matches. Table III shows the division of the five tested human classes cross the five matches.

Fig. 13(a) to (e) shows the confusion matrices of the five matches. The classification accuracy for each human classes cross the five soccer matches is presented in Table IV.

In Match 1, the classification results exhibit strong accuracy with 94.8%. The model fails to predict correctly some players from Team2. This happens because of the similarity between the shirt color (green) and the background. The prediction of the referee fails in two cases because of the similarity between the Referee colors (black) and both Team1 and Goalkeeper2 colors.

Match 2 demonstrates consistent performance in classification, with an accuracy of 95.5%. The model succeeds in all cases where color variance between the five classes is high. It misclassifies some instances in Team2 and Goalkeeper2 because of the existence of black and dark colors in these different classes.

Fig. 12. Cropped images of all human classes in each Match.

TABLE II. HUMAN DATASET IMAGES' DIVISION CROSS THE FIVE SOCCER MATCHES

Soccer Match	No. of image frames	No. of human templates	No. of tested humans	No. of all existed humans
Match 1	11	10	135	145
Match 2	12	10	154	164
Match 3	11	10	141	151
Match 4	13	10	168	178
Match 5	14	10	179	189
	61	**50**	**777**	**827**

TABLE III. THE NUMBERS OF TESTED IMAGES FOR EACH HUMAN CLASS CROSS THE FIVE SOCCER MATCHES

	humans	Team1 (T1)	Team2 (T2)	Goalkeeper1 (GK1)	Goalkeeper2 (GK2)	Referee (Ref)
Match 1	135	47	55	3	4	26
Match 2	154	64	59	2	5	24
Match 3	141	54	58	2	4	23
Match 4	168	71	63	5	4	25
Match 5	179	77	61	4	5	32

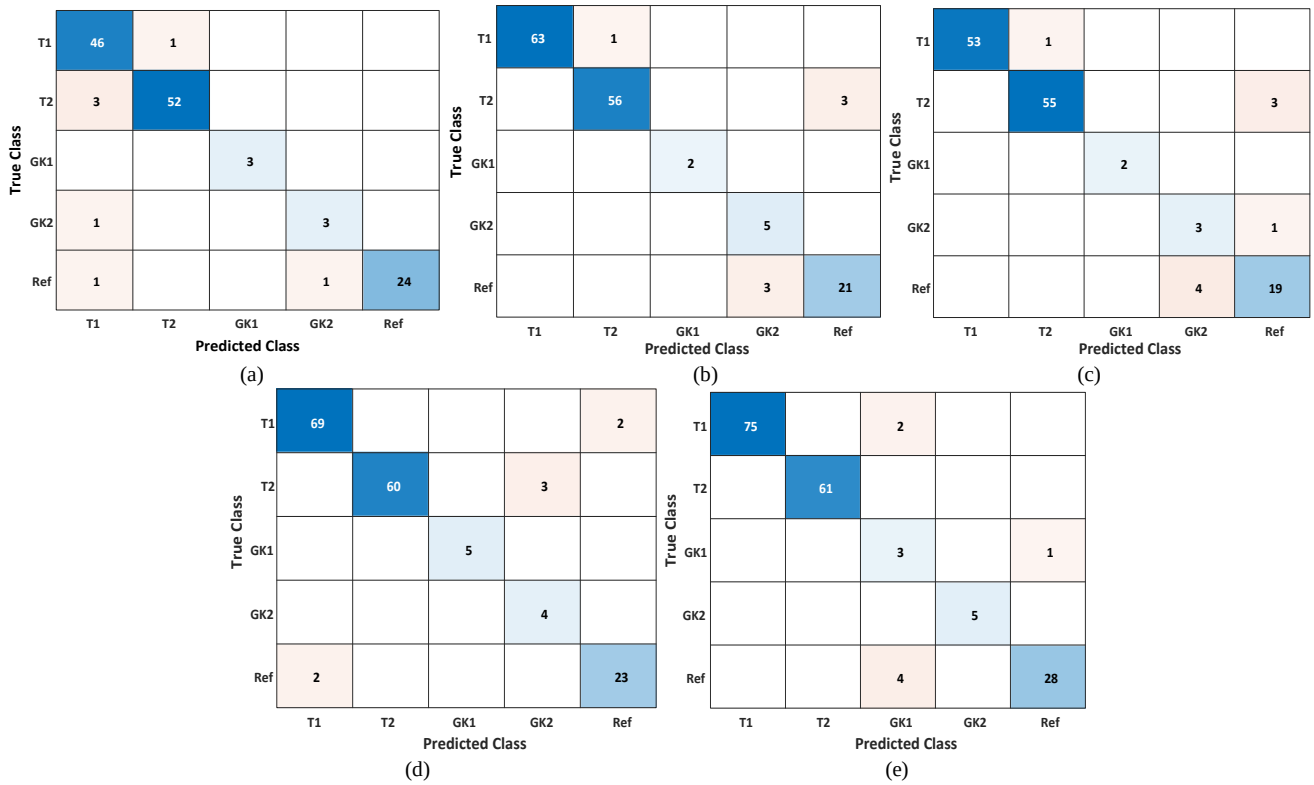


Fig. 13. Confusion matrices of the classification module for: a) Match 1, b) Match 2, c) Match 3, d) Match 4, e) Match 5.

TABLE IV. THE CLASSIFICATION ACCURACY FOR EACH HUMAN CLASSES CROSS THE FIVE SOCCER MATCHES

Soccer Match	Team1 (T1)	Team2 (T2)	Goalkeeper1 (GK1)	Goalkeeper2 (GK2)	Referee (Ref)	**Overall Accuracy**
Match 1	97.9%	94.5%	100%	75.0%	92.3%	**94.8%**
Match 2	98.4%	94.9%	100%	100%	87.5%	**95.5%**
Match 3	98.1%	94.8%	100%	75.0%	82.6%	**93.6%**
Match 4	97.2%	95.2%	100%	100%	92.0%	**95.8%**
Match 5	97.4%	100%	75.0%	100%	87.5%	**96.1%**
	97.8%	**95.9%**	**95.0%**	**90.0%**	**88.4%**	

In Match 3, classification accuracy is slightly lower at 93.6%, but the model still maintains robust performance. Team2 and Referee classes have some misclassifications based on the color similarity. Referee classification is challenging.

Match 4 demonstrates outstanding classification accuracy at 95.8%. Goalkeeper1 and Goalkeeper2 classes perform exceptionally well while Team1 and Team2 show some misclassification due to the background of the cropped images.

In the final match, the model maintains high classification accuracy at 95.4%. The model has some misclassification results especially between Referee and Goalkeeper1 because of the similarity of colors between these two classes.

In general, the classification module performs exceptionally and has excellent classification accuracy with an overall accuracy of 95.2%. The model succeeds to classify 740 different instances from a total of 777 cross the five matches. For some instances, misclassification arises because of color similarities between shirt colors or the effect of background color (soccer field color). Low color variance may lower the classification performance since we use color feature descriptors (ICCV) in this module.

V. CONCLUSIONS

In this paper, we have proposed an automated system that has the ability to detect the soccer ball and classify players and referees on the soccer field using computer vision techniques. The proposed system implements a detection model using YOLOv7 to detect the ball and all humans on the field after building a labeled dataset of 1300 different soccer game frames. It also deploys Improved Color Coherence Vector (ICCV) features to classify all humans on the field to five classes (Team1, Team2, Goalkeeper1, Goalkeeper2, and Referee) using K-Nearest Neighbor algorithm. The proposed system has achieved high efficiency in both the detection and classification modules.

The proposed system can be considered the first phase of any computer vision application in soccer game matches. It can be deployed on game analysis, player performance evaluation, automatic offside detection and fan engagement. Furthermore,

we delve into how it is poised to redefine the future of soccer as we know it. Through this exploration, it becomes increasingly evident that automated computer vision is not just a tool but a transformative force that is redefining the very essence of soccer analysis and appreciation.

REFERENCES

[1] B.T. Naik, M.F. Hashmi, and N.D. Bokde. "A comprehensive review of computer vision in sports: Open issues, future trends and research directions." *Applied Sciences* 12, no. 9 (2022): 4429.

[2] S. Kusmakar, S. Shelyag, Y. Zhu, D. Dwyer, P. Gastin, and M. Angelova. "Machine learning enabled team performance analysis in the dynamical environment of soccer." *IEEE access* 8 (2020): 90266-90279.

[3] FIFA Research Programme, available at https://www.fifa.com/technical/football-technology/research.

[4] P.R. Kamble, A.G. Keskar, K.M. Bhurchandi. "A deep learning ball tracking system in soccer videos." *Opto-Electronics Review* 27, no. 1 (2019): 58-69.

[5] B.T. Naik and M.F. Hashmi. "YOLOv3-SORT: detection and tracking player/ball in soccer sport." *Journal of Electronic Imaging* 32, no. 1 (2023): 011003-011003.

[6] B.T. Naik, M.F. Hashmi, ZW. Geem, ND. Bokde. "DeepPlayer-Track: player and referee tracking with jersey color recognition in soccer." *IEEE Access* 10 (2022): 32494-32509.

[7] G. Suzuki, S. Takahashi, T. Ogawa, M. Haseyama. "Team tactics estimation in soccer videos via deep extreme learning machine based on players formation." In *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, pp. 116-117. IEEE, 2018.

[8] Y. Ganesh, A. Sri Teja, SK. Munnangi, G. Rama Murthy. "A novel framework for fine grained action recognition in soccer." In Advances in Computational Intelligence: 15th International Work-Conference on Artificial Neural Networks, IWANN 2019, Gran Canaria, Spain, June 12-14, 2019, Proceedings, Part II 15, pp. 137-150. Springer International Publishing, 2019.

[9] J. Redmon and A. Farhadi. "YOLO9000: better, faster, stronger." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263-7271. 2017.

[10] J. Du. "Understanding of object detection based on CNN family and YOLO." In *Journal of Physics: Conference Series*, vol. 1004, p. 012029. IOP Publishing, 2018.

[11] J. Nelson, and J.Solawetz,"YOLOv5 is here: State-of-the-art object detection at 140 FPS" (2022). Available at: https://blog.roboflow.com/yolov5-is-here/.

[12] C.Y. Wang, I.H. Yeh, and HYM. Liao. "You only learn one representation: Unified network for multiple tasks." *arXiv preprint* arXiv:2105.04206 (2021).

[13] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7464-7475. 2023.

[14] G. Pass, R. Zabih, and J. Miller. "Comparing images using color coherence vectors." In *Proceedings of the fourth ACM international conference on Multimedia*, pp. 65-73. 1997.

[15] X. Chen, X. Gu, and H. Xu. "An improved color coherence vector method for CBIR". *In Proceedings of the Graduate Students Symposium of Communication and Information Technology Conference*, Beijing. 2007.

[16] P. Cunningham and S. Delany. "k-Nearest neighbour classifiers-A Tutorial." *ACM computing surveys (CSUR)* 54, no. 6 (2021): 1-25.

[17] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Fourth Edition. Academic Press, 2008.

[18] N. Panse and A. Mahabaleshwarkar, "A Dataset & Methodology for Computer Vision Based Offside Detection in Soccer", *Association for Computing Machinery*, NY, USA, 2020, available at https://github.com/Neerajj9/Computer-Vision-based-Offside-Detection-in-Soccer.

[19] H. Rezatofighi, N. Tsoi, J.Y. Gwak, A. Sadeghian, I. Reid, and S. Savarese. "Generalized intersection over union: A metric and a loss for bounding box regression." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658-666. 2019.

[20] M. Everingham, L.V. Gool, C. Williams, J. Winn, and A. Zisserman. "The pascal visual object classes (voc) challenge." *International journal of computer vision* 88 (2010): 303-338.

[21] T.T. Nguyen, K. Vandevoorde, N. Wouters, E. Kayacan, J.G. De Baerdemaeker, and W. Saeys. "Detection of red and bicoloured apples on tree with an RGB-D camera." *Biosystems Engineering* 146 (2016): 33-44.

Quality In-Use of Mobile Geographic Information Systems for Data Collection

Badr El Fhel[1]*, Ali Idri[2]
Mohammed V University in Rabat, Rabat, Morocco[1, 2]
Mohammed VI Polytechnic University[2]

***Abstract*—Mobile Geographic Information Systems (GIS) plays a vital role in data collection, offering diverse functionalities for spatial data handling. Despite advancements, accurately determining the usage environment during development remains challenging. This study uses machine learning and natural language processing to automatically classify user reviews based on the ISO 25010 quality-in-use model. Motivated by the challenge of gauging user experience during development, stakeholders analyze user reviews for insights. An experimental study compares Support Vector Machine (SVM), Random Forest, Logistic Regression, and Naive Bayes classifiers, revealing superior performance by SVM and Random Forest, particularly in efficiency evaluation. Findings underscore the efficacy of SVM in classifying user reviews, emphasizing its effectiveness in evaluating efficiency within mobile GIS applications. Moreover, it provides valuable insights for stakeholders, contributing to the enhancement of software quality of mobile GIS apps.**

***Keywords*—Mobile GIS for data collection; machine learning; software product quality; ISO/IEC 25010; natural language processing; user experience**

I. Introduction

Mobile GIS has known a significant rise in recent years as a method for data acquisition across diverse disciplines including, but not limited to, environmental monitoring [1], urban planning [2], and emergency management [3]. These GISs allow users to efficiently capture, analyze, and store spatial data related to space, resulting in an increase in productivity compared to traditional methods [4]. The implementation of Mobile GISs can provide significant benefits in terms of cost-effectiveness and real-time data acquisition [5]. In fact, mobile GIS is widely considered for data collection purpose, primarily due to the set of sensors supported by mobile devices that enable capturing positions especially Global Positioning System (GPS) and the Global Navigation Satellite System GNSS. In addition, mobile GIS enable orientation measure through the compass sensor [6]. Moreover, from a data quality point of view, mobile GISs functionalities allow controlling data quality during collection activities [7]; thereby aspect of data quality can be ensured. For instance; the accuracy of data is verified by implementing data validation rules that prevent users from inputting data when the positioning system provides values out of tolerance. Another aspect of data quality is the completeness of data which can be achieved by ensuring that all required items are collected. Finally, the verification of data consistency is achieved through the application of spatial constraints. These constraints serve to alert the user when collected data conflicts with information from other data sources. For instance, an area may be collected as a building, whereas in another data source, it is classified as a farm. These functionalities and features have the potential to influence the attractiveness of the application by partially or fully meeting user's needs. In fact, multiple mobile GIS apps, specifically designed for data collection, are currently available for public use in app repositories [8]. These repositories allow users to provide their feedbacks in the form of ratings and reviews, which are crucial for app developers and designers to improve their services and tailor the applications to meet user needs. However, due to the large number of feed backs and the diversity of wording used, reading and analyzing all reviews and ratings is time consuming manually, thus the need for the automation of this process. Moreover, the quality-in-use evaluation of these apps from the user point of view with respect to (International Standardization Organization) ISO 25010 standard [9] can be a tedious and a difficult task.

Besides, recent technological advancements have resulted in the proliferation of frameworks and libraries for natural language processing (NLP) [10], a specific area within the field of computer science and artificial intelligence that focuses on the comprehension, interpretation, and generation of human language by computers. One widely employed technique in NLP is the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, which represents text as numerical vectors [11]. When combined with machine learning (ML) classification methods, this technique enables the automated categorization of natural language into predefined classes.

For software quality, the ISO 25010 model provides two distinct models: The first is a software product quality model, which outlines eight characteristics pertaining to the static and dynamic properties of a given specific system or software product. The second is a quality-in-use model which defines the quality in use as the extent to which a product or system can be used by specific users to meet their needs and achieve specific goals with effectiveness, efficiency, freedom from risk, and satisfaction in specific contexts of use. In addition, the quality in use model defines five quality characteristics: (1) effectiveness, which refers to the accuracy and completeness with which users achieve their specified goals; (2) efficiency, which refers to the resources expended in relation to the accuracy and completeness with which users achieve their goals; (3) satisfaction, which refers to the degree to which user needs are satisfied when a product or system is used in a specified context of use; (4) freedom from risk, which refers to the degree to which a product or system mitigates potential

risks to economic status, human life, health, or the environment; and (5) context coverage, which refers to the degree to which a product or system can be used with effectiveness, efficiency, freedom from risk, and satisfaction in both specified contexts of use and in contexts beyond those initially identified.

This study assesses the quality-in-use of mobile GIS for data collection by employing manual labeling, NLP techniques, and term frequency-inverse TF-IDF as pre-processing steps on collected reviews and ratings. Subsequently, ML classification techniques are applied to the pre-processed reviews through an experimental process to identify the most suitable classifier for the specific domain of mobile GIS data collection. The classification of reviews aligns with the quality-in-use model of the ISO 25010 standard.

The study's novel contributions in the field of mobile GIS for data collection can be summarized as follows:

1) Proposing a novel application of natural language processing techniques, specifically IF-IDF, for analyzing user reviews in the context of mobile GIS. This approach enables the extraction of valuable insights from a large volume of user-generated data.

2) Evaluating the performance of four machine learning techniques - Logistic Regression, Support Vector Machine, Random Forest, and naïve bayes - in classifying user reviews based on the ISO 25010 quality characteristics, with a particular focus on the "efficiency" class (characteristic).

3) Comparing the performance metrics of SVM and Random Forest in identifying reviews belonging to the "efficiency" class, showcasing the superior performance of SVM.

4) Underlining the significance of SVM as a suitable classifier for classifying mobile GIS user reviews according to ISO 25010, offering better performance in accurately categorizing reviews related to "efficiency."

The paper is organized as follow: Section II provides an overview of the related works. Section III presents the method. Section IV outlines the experimental process, and Section V presents the results of the study. Section VI discusses the findings, and Section VII addresses potential threats to validity. Finally, Section VIII encompasses Conclusion and potential future works.

II. Related Work

In order to identify the used approaches for analyzing and classifying user reviews and ratings in mobile GISs for data collection, an analysis of previous relevant studies was conducted, with a focus on the type of study (i.e., review or empirical study, etc.), the scope (i.e., the mobile applications of GIS for data collection, or mobile applications in general, etc.), the quality aspects (i.e., quality attributes from ISO 25010 or others), NLP techniques, and ML techniques.

The aforementioned relevant studies are presented in Table I, which indicates that there have been diverse approaches employed to tackle the issue of software quality for both mobile apps in general and mobile GIS specifically for data collection purposes. For instance, Lew et al. [12] employed a modeling framework, 2Q2U (Internal/External Quality, Quality in Use, Actual Usability, and User Experience), to evaluate the quality of a desktop GIS application. This framework adopts a flexible approach to integrate and establish connections between the usability and user experience in order to evaluate software applications. Rahman et al. [13] conducted a study to validate the reliability and validity of an instrument aimed at assessing the influence of GIS quality and user satisfaction on individual work performance. The researchers drew upon an extensive analysis of existing literature and sought input from experts to develop a comprehensive questionnaire consisting of 68 items specifically related to GIS quality, user satisfaction, and individual work performance. In addition, Moumane et al. [14] conducted an empirical study with the objective of assessing the usability of mobile applications on different mobile operating systems. The study aimed to evaluate a framework specifically designed for mobile environments, based on the usability characteristic outlined in the ISO 9126 Software Quality Standard. Meng et al. [15] conducted an assessment of the usability of a Web-based Public Participatory GIS (Web-PPGIS) in a practical application setting. The researchers administered a questionnaire to participants and discovered notable disparities in system usability. These variations were observed based on the users' levels of experience and education. Other related studies have focused on the quality of data in mobile GIS as part of the system. Wang et al. [7] outlined the open architecture of field-based Mobile GIS and emphasized the importance of spatial data quality considerations. The study further elucidated how spatial data quality issues were tackled within the Mobile GIS context, in accordance with internationally recognized geoinformatics standards like ISO and Open Geospatial Consortium (OGC) standards. Furthermore, in another study by Song et al. a linear evaluation model utilizing Geographical Weighted Regression (GWR) and a nonlinear evaluation model based on random forest (RF) were developed [16]. These models were employed to quantitatively assess the relationship between geographical factors and the positioning bias of mobile phone locations.

With respect to the application of ML classification and NLP, Oyebode et al. [17] used ML classification, NLP, and TF-IDF techniques to evaluate and classify 88,125 user reviews in 104 mental health apps based on predefined classes. Five techniques were involved in this study and they are RF, Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), Logistic Regression (LR), and Stochastic Gradient Descent (SGD). Dos et al. [18] conducted a user feedback classifier based on ML of Decision Tree (DT), Naïve Bayes (NB), LR, RF, and SVM for the classification of reviews on mobile apps across various domains. The classification was performed in accordance with software quality characteristics defined by the ISO 25010 standard. In addition, Dias et al. [19] applied ML techniques and NLP in the context of software requirements classification. The study employed four algorithms: LR, SVM, MNB, and kNN. The results indicated that the use of TF-IDF in conjunction with LR produced the best classification results in differentiating requirements.

TABLE I. RELATED STUDIES

Study ID	Type of study	Scope	Quality aspects	NLP techniques	ML techniques
Lew et al. [12]	Modelling framework	Desktop GIS	Learnability	-	-
Rahman et al. [13]	survey research	GIS	Validity Reliability	-	-
Moumane et al. [14]	Empirical study	Mobile apps	Usability	-	-
Meng et al. [15]	Empirical study	Web GIS	Usability	-	-
Song et al. [16]	Qualitative study	Mobile apps	Spatial accuracy	-	GWR, RF
Oyebode et al. [17]	Comparative study	Mobile health apps	Thematic and sentiment analysis	TF-IDF	SVM, MNB, SGD, LR, RF
Dos et al. [18]	Algorithm development and evaluation study	mobile apps	External quality characteristics of ISO 25010	TF-IDF	NB, LR, DT, RF, and SVM
Dias et al. [19]	Algorithm development and evaluation study	Software Requirements	(Functional of non-functional)	Bag of Words and TF-IDF	LR,SVM,MNB,KNN
Elfhel et al. [20]	Requirements engineering	Mobile GIS for data collection	External quality characteristics of ISO 25010	-	-
Elfhel et al. [21]	Requirements engineering	Mobile GIS for data collection	usability internationalization (i18n) performance efficiency reliability sustainability	-	-

The authors in [20] has presented a measure of the external quality of mobile GIS for data collection by assessing the degree of impact of requirements related to mobile GIS for data collection on each external quality characteristic, aligned with ISO/IEC 25010. In a separate study, the authors in [21] presented a catalog of requirements for mobile GIS data collection, and demonstrated how it can be used to evaluate such applications.

This study diverges from the aforementioned related work by integrating various dimensions. Notably, while prior studies have explored diverse aspects such as Mobile GIS for data collection, algorithm development, and evaluation, the current study uniquely incorporates and merges these facets. Specifically, the investigation delves into the intersection of Mobile GIS for data collection and the application of both machine learning and natural language processing techniques. In contrast to certain previous studies that addressed the scope of Mobile GIS for data collection but refrained from employing machine learning techniques, this study bridges the gap by incorporating advanced methodologies to automatically classify user reviews based on the ISO 25010 quality-in-use model. This integration enables a more comprehensive understanding of the user experience, contributing a novel perspective to the existing body of literature in this domain. Through the integration of the Mobile GIS scope for data collection with the refined application of machine learning techniques, this study presents a distinctive and valuable contribution to the field, laying the foundation for more refined insights and progress in the evaluation of software quality for mobile GIS applications.

To the best of our knowledge, there have been no prior assessments conducted on the quality in use of mobile GIS for data collection using the ISO 25010 standard, natural language processing (NLP), and machine learning (ML) techniques.

III. METHOD

The methodology employed in this study comprises five stages, as illustrated in Fig. 1: data collection, data preprocessing, data labeling, data vectorization, automated classification, and evaluation. The subsequent subsections offer a detailed overview of each step in the methodology:

A. Data Collection

During the data collection step, a two-fold approach is used to gather users' reviews on mobile GIS applications for data collection.

- First, a pre-existing list of apps obtained from [8] was utilized, and specific inclusion criteria were applied to determine their selection. Each app needed to satisfy the following inclusion criteria: (1) relevance to mobile GIS for data collection, (2) an update date of 2020 or later, and (3) a minimum of five user reviews.
- Second, a combination of the Google Play API [22] and a Java program, developed by the research team, was utilized to gather user reviews from the selected applications.

As a result, a set of 19 apps were selected in the data collection step with a total of 8,793 reviews collected from these apps (see Table II) for comprehensive list of the selected applications and detailed of collected reviews).

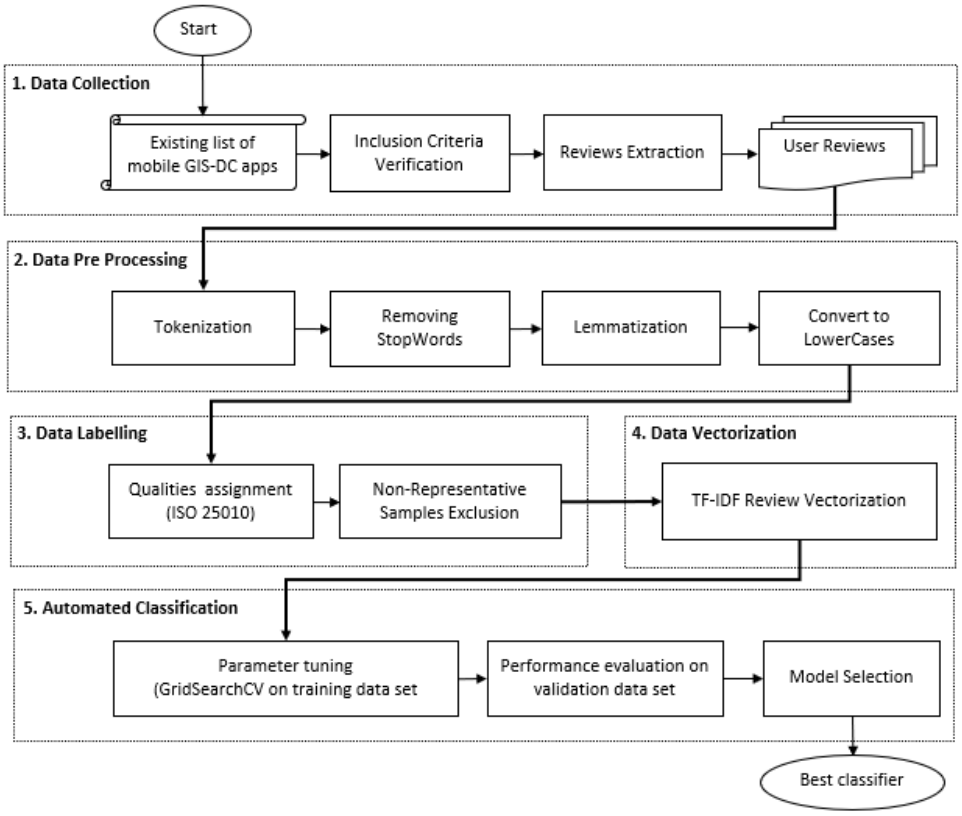


Fig. 1. User reviews classification pipeline.

TABLE II. MOBILE GIS APPS SELECTED

Application Name & Link	Number of Reviews
Mappt: GIS Data Collection	57
QField for QGIS	467
Mobile Topographer GIS	148
My GPS Coordinates	1570
Mobile Data Collection	64
GIS Mapper - Surveying App for	5
GIS Surveyor - Land Survey and	64
Land Map - GPS Land Survey & M	34
GPS Coordinates	3780
Measure map	109
Mapit Spatial - GIS Data Colle	15
Geo Survey	5
NextGIS Mobile	8
Mapit GIS - Map Data Collector	456
MapPad GPS Land Surveys	286
Save Location GPS	1056
Locus GIS offline land survey	179
SW Maps - GIS & Data Collector	388
Epicollect5 Data Collection	102

B. Data Preprocessing

Data preparation is a crucial step in natural language processing (NLP), involving the cleaning and preprocessing of raw text data to eliminate irrelevant information. In order to achieve this, the following well-known steps were followed [23]:

- Tokenization: In NPL, tokenization involves segmenting words into units called tokens based on certain rules such as removing punctuation or capitalization. The resulting tokens are intended to convey a semantic meaning. The tokenization of the collected reviews was achieved by removing punctuation marks, digits, and foreign characters (non-Latin) from the text data.
- Removing stop words: Stop words are commonly occurring words within text data that have little semantic value, such as "the" or "is", and are removed during preprocessing for NLP. The Natural Language Tool Kit (NLTK) package contains a pre-built list of stop words that can be downloaded and used [24]. However, to ensure the inclusion of domain-specific terms in the data analysis, the authors of this study have compiled a list of words related to mobile GIS to prevent them from being removed during preprocessing. This list included for instance "GPS" - a widely-known sensor used for positioning that facilitates data collection via mobile GIS. Other term of “Accuracy" was included in the list as it relates to the precision of positioning, and consequently, the quality of data collected through mobile GIS. Additionally, “Map” was involved in the list as it’s an important component in GIS that allow data presentation.

- Lemmatization in order to reduce words to their base form. For instance, words of "running," "ran," and "run" will be reduced to their base form "run".
- Convert words to lowercase.

The aforementioned steps of data preprocessing were achieved using a python program developed by the authors of this study. For each review in the data set, the program executes successively the operations of Tokenization, removing stop words, Lemmatization and converting to lowercase. The output of these steps is then stored into new column of 'pre-processed-review'.

C. Data Labelling

The data labelling step consists on the classification of the user reviews (resulted from step 2) through a manual process, which was carried out by the primary author, with respect to the quality characteristics specified in the ISO 25010 model for quality-in-use. For each review, the corresponding predefined quality characteristics are affected by the primary author and then validated by the others authors for relevance and consistency. In cases of disagreement, a consensus was achieved through collective discussion among all authors. The manual process was conducted through a web application that was specifically developed by the research team for this purpose. Fig. 2 depicts the interface of this application, which enables users to navigate through reviews and manually assign quality characteristics to each review by clicking the button related to the corresponding quality. At the end of the data labelling, a comma-separated values (CSV) file that contains the pre-processed-review with the corresponding label is generated using the button CSV. It is noteworthy that during the data labeling process, certain reviews were deemed ambiguous due to their unclear meanings or the presence of non-Latin characters that remained from the data preparation stage. As a result, these reviews were excluded from the data set, resulting in a reduction in the total number of reviews from 7322 to 6904. Table III shows the detailed results in term of reviews and quality characteristics.

D. Data Vectorization

This step consists of transforming text reviews into numerical values which can then be utilized as input for machine learning classification algorithms. TF-IDF [25] an extensively utilized technique in natural language processing, facilitates the transformation of text data into numerical vectors with a focus on classifying user reviews. This method computes multiplication of the term frequency (TF) with the inverse document frequency (IDF) for each term present in the review, yielding a numerical representation of the significance and rarity of the terms. This numerical representation enables the detection of patterns and trends within user reviews and the subsequent categorization of these reviews according to specific quality characteristics.

TABLE III. DISTRIBUTION OF REVIEWS ACROSS QUALITY CHARACTERISTICS

Quality Characteristic	Number of reviews
Context Completeness	78
Flexibility	6
Effectiveness	2236
Efficiency	952
Economic Risk Mitigation	25
Environmental Risk Mitigation	2
Health and Safety Risk Mitigation	6
Comfort	1128
Pleasure	1653
Trust	190
Usefulness	628

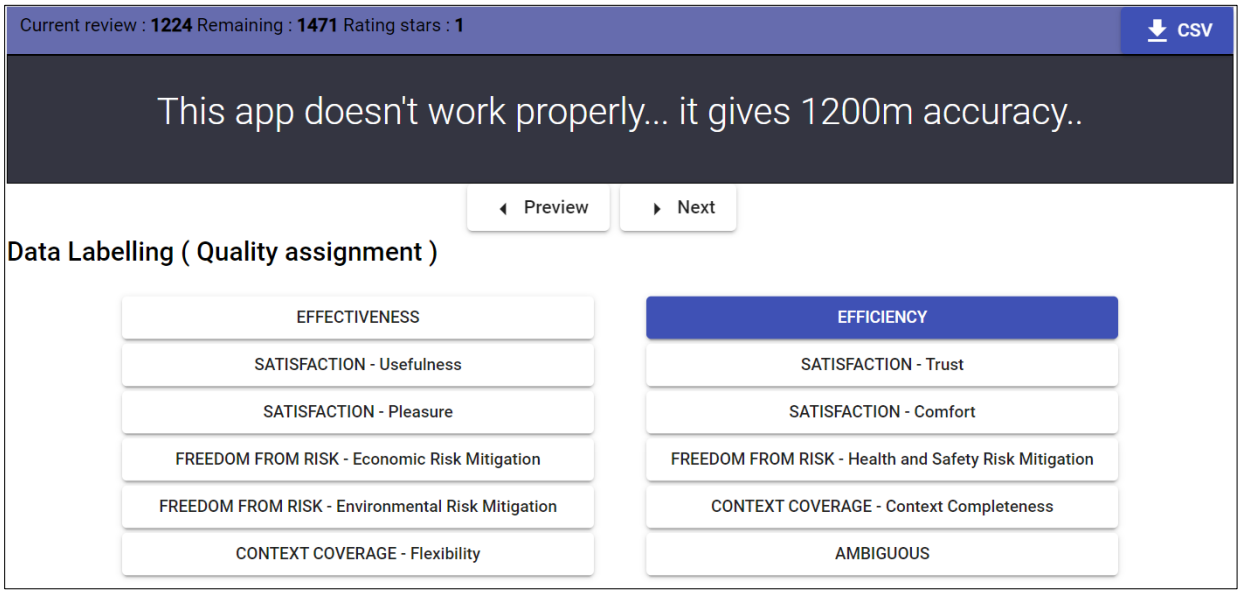


Fig. 2. Screenshot of the data labelling web interface.

In order to apply the TF-IDF vectorization technique on the user reviews, the authors developed a Python script that makes uses of the Scikit-learn [26]. This script reads the CSV file generated during the preceding data labeling phase and computes the frequency of each term in the reviews along with their respective importance scores. The resulting TF-IDF matrix comprises the user reviews in the rows and the overall terms in the columns. Moreover, the script stores the quality characteristic of each review, obtained from the data labeling step, in an additional column labeled "labels". Finally, the output of the script is produced in a new CSV file named "TF-IDF.csv".

E. Automated Classification and Evaluation

The objective of this step is to identify the most suitable machine learning algorithm for classifying user reviews related to mobile GIS for data collection based on quality-in-use characteristics of ISO. To achieve this, the datasets generated through steps 1 to 3 were used as input for the classification methods. Given the impracticality of testing all potential combinations of classification techniques, an experimental study was conducted to automate the testing and evaluation process for each machine learning algorithm's performance.

To summarize, in this study, a dataset of user reviews related to a set of mobile GIS for data collection was obtained. These reviews were subjected to preprocessing utilizing natural language processing methodologies, followed by vectorization utilizing the TF-IDF vectorization technique. A manual labelling process was carried out to classify reviews based on the quality-in-use model of ISO. A dataset with 6904 reviews was obtained and will be used in the experimental study performed in the next section.

IV. Experimental Study

In this section, an experimental study is conducted to explore the application of machine learning (ML) classification techniques on the pre-processed reviews (obtained from steps 1 to 3 in the previous section). The objective is to identify the best classifier for mobile GIS data collection.

A. Dataset Preprocessing

As shown in Table III, A few quality characteristics within the quality-in-use model have limited or insignificant representation due to the small number of available samples. These qualities are: Context Coverage – Flexibility with only six reviews, Freedom from Risk quality with 2, 6, and 25 reviews respectively to Environmental, Health and Safety, and Economic Risk Mitigation. To maintain the validity and reliability of the model, reviews associated with these particular qualities were subsequently excluded from further analysis. Thus, the dataset has undergone a reduction in the total number of samples from 6904 to 6815.

Furthermore, Fig. 3 presents a statistical analysis of the data related to this study, revealing a notable discrepancy in the sample distribution across different quality characteristics. This discrepancy gives rise to an imbalanced data challenge. To mitigate the issue of imbalanced data, the Synthetic Minority Over-sampling Technique (SMOTE) [27] was utilized to generate synthetic samples.

B. Experimental Process

The experimental process steps used is summarized as the following:

- Four ML techniques are used, namely: (1) Support Vector Machine was introduced by (Vapnik and coworkers) as “a training algorithm that maximizes the margin between the training patterns and the decision boundary” [28]. The SVM classifiers can be improved by modifying the kernel functions (Linear, Polynomial…) and its parameters (C: regulation, gamma: kernel coefficient …) [29]. (2) Logistic Regression is a statistical method applied for classification tasks by analyzing the relationship between a binary variable and one or more independent variables using a logistic function. [30]. (3) Naive Bayes is defined as a simple probabilistic model for classification that assumes that the features are conditionally independent given the class label [31]. The method models the probability of each class given the observed features using Bayes' theorem, and selects the class with the highest probability as the predicted class for a given input. (4) Finally, Random Forest is defined as an ensemble learning method that perform classification by aggregating the predictions of multiple decision trees [32].
- A Grid Search [33] tuning parameter method with five-fold cross-validation was employed to identify the optimal set of hyper-parameters for each technique (see Table IV for the values for GS parameters).

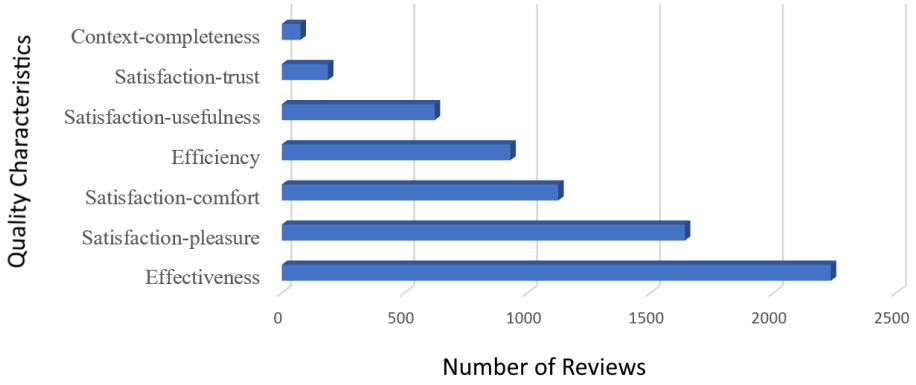


Fig. 3. Distribution of the dataset into quality classes.

TABLE IV. VALUES OF GRID SEARCH PARAMETERS

Model	Parameter	Values
Logistic Regression	Regularization Parameter C	0.1, 1, 10
	Optimization Algorithm	liblinear, lbfgs, saga
Support Vector Machine	Regularization Parameter C	0.1, 1, 10
	Kernel Function	linear, rbf, sigmoid
	Gamma	scale, auto
	Degree	2, 3
Random Forest	Number Of Decision Trees	50, 100, 200
	Criterion	gini, entropy
	Max Depth	None, 10, 20
	Min Samples Split	2, 5
	Min Samples Leaf	1, 2
	Max Features	sqrt, log2
	Bootstrap	True, False
naïve bayes	Alpha	0.1, 0.5, 1.0
	Algorithm	Multinomial Naïve Bayes

- A Python script was developed using the Scikit library to achieve optimal classifier performance. The script implements the algorithm depicted in Algorithm 1 and is available upon email request to the author.

Algorithm 1: Grid Search for ML Algorithms

Initialize a model-params dictionary of the four ML algorithms and their parameters

Create an empty report array

Compute

For each model in model-params

  Create gvv instance of GridSearchCV with model params and five-fols cross-validation

  Fit gcv with the training set to find the best hyper parameters

  Test the fitted model on the test dataset

  Compute the confusion matrix

  Compute the evaluation metrics

  Add the confusion matrix and the evaluation metrics to the report

  End

Display report

- The performance of the four-classifier experimented in this study was evaluated using four commonly used accuracy criteria [34]: (1) Precision, which quantifies the proportion of true positive predictions among all positive predictions made by the classifier. (2) Recall, which quantifies the proportion of true positive predictions among all actual positive instances. (3) Accuracy, which quantifies the proportion of correct predictions made by the classifier among all instances. (4) F-score, which combines precision and recall into a single score.

V. RESULTS ANALYSIS

Table V displays the performance of each classifier with respect to all the utilized performance metrics, along with the corresponding optimal values for the hyperparameters.

The results indicated that:

- The Random Forest classifier achieved a precision of 0.81, indicating that, out of all instances that were predicted as positive, 81% were actually positive. The classifier also achieved a recall of 0.79, indicating that, out of all true positive instances, 79% were correctly identified by the classifier. The overall accuracy of the classifier was found to be 0.79, indicating that 79% of the predictions made by the classifier were correct. The F1-score, which is a harmonic mean of precision and recall, was found to be 0.80, indicating that the precision and recall of the classifier was balanced.
- The SVM classifier obtained scores that were slightly different from those of the Random Forest classifier, with a precision score of 0.79, an accuracy score of 0.80, a recall score of 0.80, and an F1-score of 0.79.
- The Logistic Regression classifier performed slightly worse in terms of accuracy and recall, but obtained 0.81 in precision and 0.79 in F1-score.
- The Naive Bayes classifier had the lowest scores across all accuracy criteria, indicating that it performed less well than the other three classifiers.

Moreover, the confusion matrices scores related to SVM and Random Forest were calculated and presented respectively in Table VI and Table VII. As depicted in the confusion matrices, both models demonstrate strong performance. This is evidenced by the majority of entries being located along the diagonal of the matrices.

TABLE V. GLOBAL CLASSIFICATION SCORES AND HYPER PARAMETERS

Model	Performance scores				Hyper Parameters	
	Precision	*Accuracy*	*Recall*	*F1-score*	*Parameter*	*Value*
Random Forest	0.81	0.79	0.79	0.8	Number Of Decision Trees	200
					Criterion	entropy
					Max Depth	None
					Min Samples Split	2
					Min Samples Leaf	1
					Max Features	log2
					Bootstrap	False
Support Vector Machine	0.79	0.8	0.8	0.79	Regularization Parameter C	10
					Kernel Function	rbf
					Gamma	scale
					Degree	2
Logistic Regression	0.81	0.77	0.77	0.79	Regularization Parameter C	10
					Optimization Algorithm	saga
naïve bayes	0.77	0.73	0.73	0.75	Alpha	0.1
					Algorithm	Multinomial Naïve Bayes

TABLE VI. SVM CONFUSION MATRIX

		Actual Values						
		Context Completeness	**Effectiveness**	**Efficiency**	**Comfort**	**Pleasure**	**Trust**	**Usefulness**
Predicted Values	Context Completeness	2	3	3	3	0	0	3
	Effectiveness	0	402	9	14	7	3	13
	Efficiency	0	14	164	16	6	2	4
	Comfort	0	12	19	178	12	2	10
	Pleasure	1	3	5	19	275	1	4
	Trust	0	4	10	9	7	7	2
	Usefulness	0	10	11	19	10	1	64

TABLE VII. RANDOM FOREST CONFUSION MATRIX

		Actual Values						
		Context Completeness	**Effectiveness**	**Efficiency**	**Comfort**	**Pleasure**	**Trust**	**Usefulness**
Predicted Values	Context Completeness	2	3	1	4	1	3	0
	Effectiveness	0	399	10	10	5	13	11
	Efficiency	0	11	160	10	3	14	8
	Comfort	0	10	18	169	6	19	11
	Pleasure	2	0	11	13	262	12	8
	Trust	1	1	6	4	6	18	3
	Usefulness	1	4	12	12	9	11	66

TABLE VIII. SVM AND RANDOM FOREST QUALITY CLASS SCORES

Classifier	Quality Class	precision	recall	F1_score
SVM	Context Completeness	0.67	0.14	0.24
	Effectiveness	0.9	0.9	0.9
	Efficiency	0.74	0.8	0.77
	Comfort	0.69	0.76	0.73
	Pleasure	0.87	0.89	0.88
	Trust	0.44	0.18	0.25
	Usefulness	0.64	0.56	0.6
Random Forest	Context Completeness	0.33	0.14	0.2
	Effectiveness	0.93	0.89	0.91
	Efficiency	0.73	0.78	0.75
	Comfort	0.76	0.73	0.74
	Pleasure	0.9	0.85	0.87
	Trust	0.2	0.46	0.28
	Usefulness	0.62	0.57	0.59

Furthermore, the performance scores related to SVM and Random Forest were calculated for each quality class and the results are presented in Table VIII. As demonstrated, the precision, recall, and F1-score demonstrate heterogeneity across various categories, providing valuable insights into the classification performance of each algorithm. Subsequently, in the following section, these outcomes will be discussed in the context of the criteria for mobile GIS for data collection to select the best classifier from SVM and RF.

VI. Discussion

Table V illustrates that the accuracy metric for SVM and RF classifiers achieved high values of 0.80 and 0.79, respectively. These results suggest that both classifiers were successful in correctly classifying a high proportion of instances, indicating that the vectorization process utilizing TF-IDF was successful in identifying relevant terms within the corpus of user reviews. Note that TF-IDF was previously identified in research as a strong vectorization method among user reviews [17, 18].

The effectiveness of TF-IDF in mobile GIS for data collection reviews can be explained by its adeptness at capturing term significance through frequency calculations. Within this domain, where reviews frequently incorporate specialized terminology and jargon pertaining to geographic information, mobile devices, and associated technologies, TF-IDF stands out by recognizing and assigning importance to these specific terms based on their frequency. This emphasis on the frequency of domain-specific terms contributes to a more precise representation of the data, aligning with the high accuracy metrics observed in the classifiers' performance as highlighted in Table V.

The SVM classifier and Random Forest classifier were evaluated using precision, accuracy, recall, and F1-score metrics. The results revealed that the Random Forest classifier obtained scores of 0.81, 0.79, 0.79, and 0.80, respectively, while the SVM classifier obtained scores of 0.79, 0.80, 0.80, and 0.79, respectively. These results indicate that the Random Forest classifier performed slightly better in terms of precision and F1-score, while the SVM classifier performed better in terms of accuracy and recall.

The SMOTE technique has been employed to mitigate the issue of class imbalance. however, a detailed analysis of class scores is still necessary to reveal any performance variations of classifiers on specific classes and provide a more comprehensive understanding of their capabilities. Although no significant differences were observed in the four performance scores of the two classifiers, Random Forest and SVM, an exhaustive evaluation of their performance was conducted, taking into account the specific domain of mobile GIS for data collection. In fact, the requirements of mobile GIS for data collection regarding the positioning accuracy is crucial, as it affects directly the quality of collected data [21], which subsequently impacts the overall data collection process. Moreover, a real challenge is associated with GPS positioning accuracy in smartphones [35] and extensive investigations were conducted to identify factors that influence the accuracy of mobile GIS positioning [36-38]. In this light, various solutions have been adopted to enhance the positioning accuracy in mobile mode [39, 40]. Therefore, comparing the performance of Random Forest and SVM classifiers on the class of efficiency can aid in selecting the best classifier for mobile GIS data collection purposes.

Based on the evaluation of the classifiers scores presented in Table VIII, the SVM classifier appears to be a more suitable option for identifying a maximum number of user reviews belonging to the "Efficiency" class in mobile GIS data collection. The SVM classifier exhibits a higher F1-score (0.77), recall score (0.80), and precision (0.74) as compared to the Random Forest classifier (F1-score: 0.75, recall: 0.78, precision: 0.73) for this class. These findings suggest that the SVM classifier has a greater ability to detect positive samples of the "Efficiency" class while maintaining a good balance between precision and recall. Furthermore, the SVM classifier has a higher precision score (0.74) than the Random Forest classifier (0.73) for this class, indicating that the SVM classifier generates fewer false positive predictions. Thus, the SVM classifier may be the optimal choice for this classification task in the mobile GIS data collection domain.

In addition, the complexities inherent in user reviews within the mobile GIS for data collection domain introduce a level of intricacy marked by complex and nonlinear relationships between linguistic expressions and corresponding sentiments. These reviews serve as reflections of nuanced discussions prevailing in this specialized technical domain. Leveraging their unique capacity to define optimal hyperplanes within high-dimensional spaces, SVM exhibit notable proficiency in capturing the nuanced patterns embedded in these reviews. The algorithm's adeptness in recognizing subtle differences and correlations within the technical language of user reviews establishes SVMs as a resilient and effective choice for classifying user-generated content within the intricate realm of mobile GIS for data collection. This underscores their efficacy in addressing the inherent complexities specific to mobile GIS for data collection.

VII. Threats to Validity

Although objectivity was applied during the research process, there may still be limitations to this study:

- In the natural language processing phase, certain terms may have been erroneously categorized as stop words and consequently eliminated from the dataset. This could impact the construct validity of the study. To address this issue, a specialized GIS term dictionary was constructed to ensure that relevant terms are not automatically removed during the data preprocessing stage, thus improving construct validity.
- The automated classification in this study concerned mobile GIS user reviews, which could pose potential challenges to external validity. To address this concern, the set of studied reviews was carefully chosen to ensure a representative sample. This limitation may have slightly affected the performance metrics, but optimism exists that the results may be utilized in forthcoming studies related to mobile GIS.

- User reviews were assigned manual classifications based on the quality-in-use model. However, there is a possibility that a review may belong to more than one class which impact the internal validity. To address this issue, only the clearest classification was considered.

VIII. Conclusion and Future Work

This study involved an experiment aimed at identifying the best classifier for analyzing user reviews of mobile GIS applications in the context of data collection. The process involved five steps: data collection, data preprocessing, data labeling, data vectorization, automated classification, and evaluation.

The evaluation of classifiers unveiled notable performance metrics. The Random Forest classifier showcased balanced performance, exhibiting a precision of 0.81, a recall of 0.79, an accuracy of 0.79, and an F1-score of 0.80. The SVM classifier, with slightly differing yet competitive scores, achieved a precision of 0.79, accuracy of 0.80, recall of 0.80, and an F1-score of 0.79. Likewise, the Logistic Regression classifier demonstrated a precision of 0.81, accuracy of 0.79, recall of 0.79, and an F1-score of 0.79, while the Naive Bayes classifier showed lower scores across accuracy criteria. Notably, when honing in on the "efficiency" class, the SVM classifier outperformed the Random Forest classifier, displaying superior precision (0.74), recall (0.80), and F1-score (0.77) compared to the Random Forest classifier (precision: 0.73, recall: 0.78, F1-score: 0.75). These results underscore the effectiveness of the TF-IDF vectorizer and SVM classifier combination within the specific domain of mobile GIS for data collection, emphasizing the significance of efficiency requirements in this context. The implications of this study extend to developers and designers of mobile GIS applications, providing insights for automatic quality evaluation using the ISO 25010 quality-in-use model.

In future investigations, the aim is to expand the scope of the study by increasing the number of experiments conducted. This expansion will enable a more extensive gathering of relevant and accurate results. Additionally, we intend to investigate the correlation between external quality and the quality-in-use of mobile GIS applications specifically designed for data collection purposes, with the ultimate goal of developing a predictive model for quality-in-use. This may have practical implications for enhancing the user experience and satisfaction of mobile GIS applications for data collection by ensuring that external quality meets the requirements of quality-in-use.

Declaration of Competing Interest

The authors declare that the publication of this article does not involve any conflicts of interest.

References

[1] M. M. Nowak, K. Dziób, Ł. Ludwisiak, and J. Chmiel, "Mobile GIS applications for environmental field surveys: A state of the art," Global Ecology and Conservation, vol. 23, p. e01089, 2020/09/01/ 2020, doi: https://doi.org/10.1016/j.gecco.2020.e01089.

[2] B. Yang, "Developing a Mobile Mapping System for 3D GIS and Smart CityPlanning,"Sustainability,vol.11,no.13,2019,doi:10.3390/su11133713

[3] I. H. El-Gamily, G. Selim, and E. A. Hermas, "Wireless mobile field-based GIS science and technology for crisis management process: A case study of a fire event, Cairo, Egypt," The Egyptian Journal of Remote Sensing and Space Science, vol. 13, no. 1, pp. 21-29, 2010/06/01/ 2010, doi: https://doi.org/10.1016/j.ejrs.2010.07.003.

[4] A. Jayasinghe, N. Sanjaya, and Y. Chemin, "Application of Mobile GIS for Mobility Mapping," 06/01 2014.

[5] F. Döner, "Examination and comparison of mobile GIS technology for real time Geo-data acquisition in the field," Survey Review, vol. 40, no. 309, pp. 221-234, 2008/07/01 2008, doi: 10.1179/003962608X291013.

[6] Z. Ma, Y. Qiao, B. Lee, and E. Fallon, Experimental evaluation of mobile phone sensors. Signals and Systems Conference (ISSC 2013), 24th IET Irish (in en), 2013.

[7] F. Wang and W. Reinhardt, "Spatial data quality concerns for field data collection in mobile GIS," in Proc.SPIE, 2006, vol. 6420, p. 64201C, doi: 10.1117/12.712733.[Online].Available: https://doi.org/10.1117/12.712733.

[8] B. E. Fhel, A. Idri, and L. Sardi, "Free Mobile Geographic Information Apps Functionalities: A Systematic Review," (in en), RACSC, vol. 16, no.3,p.e200722206911,2023,doi:10.2174/2666255816666220720113157.

[9] Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models, I. S. ISO/IEC/IEEE-25010, 2011.

[10] V. N. Gudivada and K. Arbabifard, "Chapter 3 - Open-Source Libraries, Application Frameworks, and Workflow Systems for NLP," in Handbook of Statistics, vol. 38, V. N. Gudivada and C. R. Rao Eds.: Elsevier, 2018, pp. 31-50.

[11] "Data Mining," in Mining of Massive Datasets, A. Rajaraman and J. D. Ullman Eds. Cambridge: Cambridge University Press, 2011, pp. 1-17.

[12] P. Lew, L. Zhang, and L. Olsina, "Usability and user experience as key drivers for evaluating GIS application quality," in 2010 18th International Conference on Geoinformatics, 18-20 June 2010 2010, pp. 1-6, doi: 10.1109/GEOINFORMATICS.2010.5567803.

[13] M. S. Rahman, S. M. Shuhidan, M. N. Masrek, and M. F. Baharuddin, "Validity and Reliability Testing of Geographical Information System (GIS) Quality and User Satisfaction towards Individual Work Performance,"Proceedings,vol.82,no.1,2022,doi:10.3390/proceedings2022082068.

[14] K. Moumane, A. Idri, and A. Abran, "Usability evaluation of mobile applications using ISO 9241 and ISO 25062 standards," SpringerPlus, vol. 5, p. 548,2016, doi: 10.1186/s40064-016-2171-z.

[15] Y. Meng and J. Malczewski, "Usability evaluation for a web-based public participatory GIS: A case study in Canmore, Alberta," cybergeo, 2009/12/17/ 2009, doi: 10.4000/cybergeo.22849.

[16] X. Song, Y. Long, L. Zhang, D. G. Rossiter, F. Liu, and W. Jiang, "Spatial Accuracy Evaluation for Mobile Phone Location Data With Consideration of Geographical Context," IEEE Access, vol. 8, pp. 221176-221190, 2020, doi: 10.1109/ACCESS.2020.3043317.

[17] O. Oyebode, F. Alqahtani, and R. Orji, "Using Machine Learning and Thematic Analysis Methods to Evaluate Mental Health Apps Based on User Reviews," IEEE Access, vol. 8, pp. 111141-111158, 2020, doi: 10.1109/ACCESS.2020.3002176.

[18] R. dos, "A Practical User Feedback Classifier for Software Quality Characteristics," in The 33rd International Conference on Software Engineering and Knowledge Engineering, 2021/07/06/ 2021, pp. 340-345,doi:10.18293/SEKE2021-055.[Online].Available: http://ksiresearch.org/seke/seke21paper/paper055.pdf.

[19] E. Dias Canedo and B. Cordeiro Mendes, "Software Requirements Classification Using Machine Learning Algorithms," (in en), Entropy, vol. 22, no. 9, p. 1057, 2020/09// 2020, doi: 10.3390/e22091057.

[20] B. E. Fhel, L. Sardi, A. Idri, and A. Idri, "Quality Evaluation of Mobile GIS for Data Collection," in 17th International Conference on Evaluation of Novel Approaches to Software Engineering, 2023/01/25/ 2023, pp. 309-316.[Online].Available: https://www.scitepress.org/Link.aspx?doi=10.5220/0011033900003176.

[21] B. El Fhel, L. Sardi, and A. Idri, "A Requirements Catalog of Mobile Geographic Information System for Data Collection," Á. Rocha, H. Adeli, G. Dzemyda, F. Moreira, and A. M. Ramalho Correia, Eds., 2021 2021, Cham: Springer International Publishing, in Advances in

Intelligent Systems and Computing, pp. 324-336, doi: 10.1007/978-3-030-72651-5_32.

[22] Google."API Google Play Developer." https://developers.google.com/android-publisher?hl=fr (accessed March 28, 2023, 2023).

[23] A. Occhipinti, L. Rogers, and C. Angione, "A pipeline and comparative study of 12 machine learning models for text classification," Expert Systems with Applications, vol. 201, p. 117193, 2022/09/01/ 2022, doi: https://doi.org/10.1016/j.eswa.2022.117193.

[24] "NLTK::Natural Language Toolkit.". Available: https://www.nltk.org/. (accessed March 28, 2023).

[25] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey," International Journal on Digital Libraries, vol. 17, no. 4, pp. 305-338, 2016/11/01, doi: 10.1007/s00799-015-0156-0.

[26] "scikit-learn: machine learning in Python — scikit-learn 1.2.1 documentation." [Online]. Available: https://scikit-learn.org/stable/. accessed March 28, 2023).

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," jair, vol. 16, pp. 321-357, 2002/06/01/ 2002, doi: 10.1613/jair.953.

[28] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," 1992/07/01/ 1992, New York, NY, USA: Association for Computing Machinery, in COLT '92, pp. 144-152, doi: 10.1145/130385.130401.

[29] S. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," Neural Networks, vol. 12, no. 6, pp. 783-789,1999/07/01, doi: https://doi.org/10.1016/S0893-6080(99)00032-5.

[30] A. Agresti, Foundations of linear and generalized linear models (Wiley series in probability and statistics). Hoboken, New Jersey: John Wiley & Sons Inc, 2015, p. 1.

[31] M. Lintean and V. Rus, "Large scale experiments with naive bayes and decision trees for function tagging," Int. J. Artif. Intell. Tools, vol. 17, no. 03, pp. 483-499, 2008/06// 2008, doi: 10.1142/S0218213008004011.

[32] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001/10/01 2001, doi: 10.1023/A:1010933404324.

[33] A. Zakrani, A. Najm, and A. Marzak, "Support Vector Regression Based on Grid-Search Method for Agile Software Effort Prediction," in 2018 IEEE 5th International Congress on Information Science and Technology (CiSt), 21-27 Oct. 2018, pp. 1-6, doi: 10.1109/CIST.2018.8596370.

[34] A. Tharwat, "Classification assessment methods," Applied computing and informatics, vol. 17, no. 1, pp. 168-192, 2021.

[35] F. Zangenehnejad and Y. Gao, "GNSS smartphones positioning: advances, challenges, opportunities, and future perspectives," Satellite Navigation, vol. 2, no. 1, p. 24, 2021/11/16/ 2021, doi: 10.1186/s43020-021-00054-y.

[36] C. Bauer, "On the (In-)Accuracy of GPS Measures of Smartphones: A Study of Running Tracking Applications," in International Conference, 2013 2013, Vienna, Austria: ACM Press, pp. 335-341, doi: 10.1145/2536853.2536893.

[37] K. Merry and P. Bettinger, "Smartphone GPS accuracy study in an urban environment," (in en), PLoS ONE, vol. 14, no. 7, p. e0219890, 2019/07/18/ 2019, doi: 10.1371/journal.pone.0219890.

[38] P. A. Zandbergen, "Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning," (in en), Transactions in GIS, vol. 13, pp. 5-25, 2009/06// 2009, doi: 10.1111/j.1467-9671.2009.01152.x.

[39] Z. Peng, Y. Gao, C. Gao, R. Shang, and L. Gan, "Improving Smartphone GNSS Positioning Accuracy Using Inequality Constraints," (in en), Remote Sensing, vol. 15, no. 8, p. 2062, 2023/04/13/ 2023, doi: 10.3390/rs15082062.

[40] J. Hwang, H. Yun, Y. Suh, J. Cho, and D. Lee, "Development of an RTK-GPS Positioning Application with an Improved Position Error Model for Smartphones," (in en), Sensors, vol. 12, no. 10, pp. 12988-13001, 2012/09/25/ 2012, doi: 10.3390/s121012988.

Bitcoin Optimized Signal Allocation Strategies using Decomposition

Sherin M.Omran[1*], Wessam H. El-Behaidy[2], Aliaa A. A. Youssif[3]
Department of Computer Science, College of Computers and Artificial Intelligence, Helwan University, Cairo 11795, Egypt[1, 2]
College of Computing and Information Technology, Arab Academy for Science, Technology & Maritime Transport (AASTMT), Smart Village 12577, Egypt[3]

***Abstract*—Bitcoin is the first and most famous cryptocurrency. It is a virtual currency that is operated in a decentralized form using cryptographic strategies called blockchains. Although it has experienced significant market acceptance by traders and investors in recent years, it also suffers from volatility and riskiness. Technical analysis is one of the most powerful tools used for trading signals' allocation using some algorithmic strategies called technical indicators. In this research, a newly proposed multi-objectives decomposition-based particle swarm optimization algorithm is used to find the best parameter values for some technical indicators, which in turn generates the best trading signals for Bitcoin trading. In this context, three conflicting objectives have been used, i.e., the return on investment, the Sortino-ratio, and the number of trades. The proposed algorithm is compared to the original MOEA/D algorithm as well as the indicators using their original parameters. Results showed the superiority of the proposed algorithm during the training and testing periods over the other benchmarks.**

***Keywords*—Bitcoin; technical analysis; decomposition; particle swarm optimization; MOEA/D**

I. Introduction

Unlike the well-known physical currencies such as dollars or euros, bitcoin is a sort of digital currency backed by cryptographic protocols called blockchains [1]. These cryptographic protocols facilitate secure online payment with no need for intermediaries. There exist significant fluctuations in the bitcoin prices every day which increases its volatility and raises the level of risk [2]. This volatility enormously affects the traders' outcomes. So, powerful algorithms are always needed to help the traders and investors get the best returns with the minimum level of risk.

The algorithmic trading strategies for cryptocurrencies can be classified into three different models, i.e., models based on machine learning, Portfolio Optimization (PO) models, and trading-strategies optimization models.

Machine learning and deep learning are used in the literature as predictive models to forecast the future price patterns based on the analysis of historical market data [3], [4], [5], and [6].

PO in the context of cryptocurrencies entails selecting a combination of cryptocurrencies and determining the appropriate weights for each asset in order to achieve the desired portfolio characteristics. Depending on the preferences of the investor, the objective may be to maximize returns, minimize risk, or achieve a specific risk-return trade-off [7]. Portfolio optimization researches can be categorized into statistical models such as Modern Portfolio Theory (MPT) as found in [8], [7], and [9] and Multi-Objective Evolutionary Algorithms (MOEA) models such as [10], and [11].

The trading strategies optimization models mainly aim to allocate the optimal trading signals that enhance the trading outcomes. In study [12] a signal herding model is proposed in order to enhance the decision-making process. In study [13] the trading signals are generated based on the market tweets sentiments whereas in study [14], both time series analysis and social signals are used to produce tractable trading strategies.

Trading signals generation or allocation can be achieved using Technical Analysis (TA) based indicators. Although the optimization of the TA based trading-strategies was found in the literature for other types of markets such as physical currencies and stock markets, it was not found for cryptocurrencies. So, this research tries to cover the lack of this research point by proposing new optimized algorithmic trading-strategies for some Technical Indicators (TI).

As can be seen previously, cryptocurrency trading can be considered as a Multi-objective Optimization (MO) problem. As it involves handling a set of conflicting objectives simultaneously such as maximizing the returns and the percentage of the profitable trades to the non-profitable ones, minimizing risks, the transaction costs, and the number of trades, etc. [15].

The MO problem is described as in Eq. (1) [16]:

$$\text{Maximize } F(x) = (f_1(x), \dots \dots f_m(x))^T \quad (1)$$

subject to $x \in \Omega$

Such that: Ω is the variable or decision space, F: $\Omega \rightarrow R^m$ is the objective space, where m is the total number of objectives. As there are multiple contradictions between the different objectives, there can never be a unique solution that satisfies all the objectives simultaneously but rather a set containing the whole non-dominated solutions referred to as Pareto Set (PS). Let $x_1, x_2 \in R^m$, it can be said that x_1 dominates x_2 only if $x_1^i \geq x_2^i$ for each $i \in \{1, 2, \dots, m\}$ and $x_1^j > x_2^j$ at least once for $j \in \{1,2, \dots, m\}$. The set of the objective vectors corresponding to the points in the PS is called the Pareto Front (PF).

MOEA based on Decomposition (MOEA/D) is one of the most simple and efficient techniques to solve MO problems. It

can efficiently find the best set of non-dominated solutions regardless of the increasing number of objectives [17], [18]. On the contrary, MOEA based on Pareto dominance can efficiently handle problems with lower number of objectives, however by increasing the number of objectives it can hardly cover the entire (PF) [18].

MOEA/D transforms the MO problem into a simple set of scalar sub-problems then, it solves each of them simultaneously and independently. This transformation is achieved with the help of two basic factors, i.e., aggregation or also called Scalarization Functions (SF) and a set of well-selected weight vectors. These are the basic factors that control the performance of MOEA/D algorithms.

New decomposition strategies were found in the literature either by proposing new SFs as in [19], and [20] or by using a combination of different SFs [21], and [22].

Some changes to the original weight generation mechanism that was originally proposed by Zhang [17] were also found in the literature such as MOEA/D-URAW [23], AWD-MOEA/D [24] and MOEA/D-AWG [25].

The MOEA/D algorithms were successfully applied to different application areas such as PO [11], [26], image segmentation [27], and network routing [28].

In this paper a new MO Particle Swarm Optimization Algorithm using Decomposition (MOPSO/D) is proposed for BTC trading signals' allocation. The algorithm optimizes the original parameters of three of the most used TIs. A new weight generation strategy was also presented and used in order to further improve the performance of the algorithm. The proposed algorithm is compared to the original MOEA/D and the TIs with their original parameters based on three objectives, i.e., the return, the risk, and the number of trades.

The rest of this paper will be organized as follows: In Section II, the trading signals allocation mechanisms are described. Section III shows the proposed algorithm, whereas Section IV presents the results and the conclusion in Section V.

II. Trading Signals Allocation Mechanisms

Cryptocurrencies are well-known for their extreme volatility, which refers to the huge and sudden price changes they encounter over short periods of time. This nature raises the riskiness level of such markets, making them hardly predictable. So, for the investors to get the benefits from crypto-market trading, there should be powerful algorithms that ensure the profitability and safety of their trading. As seen in Fig. 1, the Bitcoin value started in 2017 at about 1,000 USD and has risen to 20,000 USD. Then, it dropped back again, up to 4,000 USD by 2019.

TA tools (specifically the Technical Indicators TA) are frequently used by investors for their simplicity. They are used either as the basic trading strategies or as confirmation tools to generate and allocate the trading signals, i.e., buy and sell. For example, most PO tools basically depend on them in order to get their final trading decisions [29] and [30].

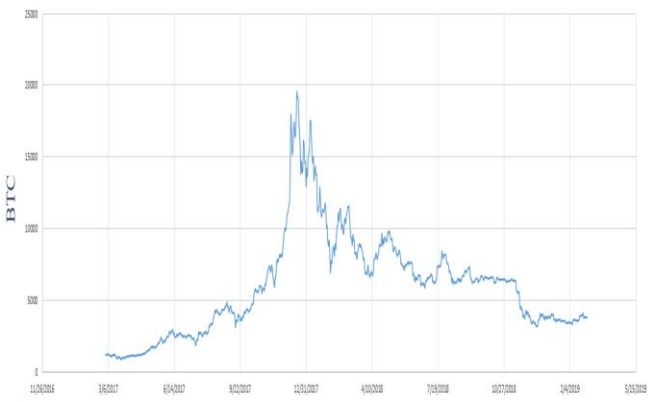


Fig. 1. The value of Bitcoin (BTC) in USD from 2017 to 2019.

A. Technical Indicators (TIs)

The TIs are mathematical formulations (equations) that are operated in an algorithmic manner. Each of these indicators has its own parameter combinations that controls its final performance [31]. The original values of these parameters are generated by their creators, however with more challenging markets such as the crypto market it could not always provide the required performance . The main objective of this study is to find the optimal set of parameters that helps allocating the best trading signals (buy and sell).

Three TIs are considered in this study, i.e., Double Weighted Moving-Average (DWMA), Exponentially Smoothed Rate of Change (ES-RoC) and Stochastic Relative Strength Index (S-RSI).

The DWMA indicator in [32] is one of the versions of Moving Average indicators (MAs). This type of indicators is used to generate an updated price average based on the selected time frame. The term double here means that the signals are generated based on the crossovers between two WMAs with different time frames.

A buy signal is provided when the shorter WMA crosses from below to above the longer WMA and vice versa for the sell signal. The WMA is calculated as seen in Eq. (2), such that n is the number of days used for calculations and P refers to the daily prices. For example, P_1 is the closing price for the first day of calculations and P_n.is the most recent price value at day n. The DWMA has two parameters n_1 and n_2 for the short and long WMAs consequently, such that the original parameters values are {20,50}.

$$WMA = \frac{(P_1 * n) + (P_2 * n-1) + \cdots + (P_{n-1} * 2) + \square(P_n)}{\frac{[n*(n+1)]}{2}} \quad (2)$$

The ES-RoC [31] indicator calculates the rate of change for the values of the Exponential MA (EMA) values over the past n days. The EMA is a type of MAs where the weighting for the days of calculations grows exponentially rather than the linear weighting as in WMA. The range of the outputs of this indicator is always in the range ±100.

The signals are generated by crossing above or below the center line (which is zero in this case) for buy and sell signals in sequence without considering a definite overbought or

oversold level. This indicator is mainly used to evaluate the strength of the upcoming trends. The ES-RoC is calculated as shown in Eq. (3), where $EMA_{current}$ is the value of the EMA on the last day of calculation and EMA_n is the value of the EMA *n*-days ago.

In this study, two extra parameters, i.e., overbought and oversold levels are produced for trading signals generation. Rather than crossing above or below zero line, the buy and sell signals are generated by crossing through the oversold and overbought levels for buy and sell signals consequently. The original parameter values for ES-RoC are 14 and 20 days for the RoC and the EMA sequentially.

$$ES_{RoC} = \frac{EMA_{current} - EMA_n}{EMA_n} \times 100 \quad (3)$$

S-RSI [31] is a mixed indicator that applies the Stochastic-Oscillator (SO) indicator to the RSI value. RSI is another indicator that analyses current price levels against those of the recent past. The output values lie in the range from 0 and 100 that is used to identify the oversold and overbought levels.

The signals are generated by the crossovers with these levels the same way as the ES-RoC. The value of S-RSI [31] is calculated as shown in Eq. (4). Such that *n* and *t* are the timespans for RSI and S-RSI in sequence. The original parameters of the S-RSI are 14 days for both *n* and *t* with an overbought level of 70 and oversold level of 30.

$$S\_RSI = \frac{RSI_{current} - min(RSI_t)}{max(RSI_t) - min(RSI_t)} \quad (4)$$

$$RSI_{current} = 100 - \frac{100}{1 + \frac{avg\_gain(n)}{avg\_loss(n)}} \quad (5)$$

B. Objective functions

For the optimization process at hand, three conflicting objectives are selected.

- The percentage Return on Investments (RoI): it evaluates the profitability of the investment strategy by comparing the investment net gain to the total investment costs as shown in Eq. (6). This is a maximization objective as the investors always aim to maximize their profits.

$$RoI = \frac{Net\_gains}{Investment\_costs} * 100 \quad (6)$$

- Sortino Ratio (SR): is a measure of the risk of investment. It is calculated as in Eq. (7), where σ_{down} is the Standard Deviation (SD) of the returns that are below the average (the downward SD). The target here is to maximize the SR in order to minimize the risk.

$$SR = \frac{aggregated\_return}{\sigma_d} \quad (7)$$

- The number of trades: The target here is to get the minimum number of trades.

III. The Proposed Algorithm

As mentioned before, the idea behind the decomposition-based algorithms is to simplify the MO problem and convert it into a group of single objective SPs. This simplification is accomplished using the SF or aggregation function. Different scalarization approaches were found in the literature. Among the most recommended SFs due to its simplicity and adaptability to different types of problems is the Chebyshev approach [33]. As a result, a number of variants were proposed in the literature to enhance its performances.

The Augmented Chebyshev (ACh) is one of the variants with an additional augmentation term that was proposed in order to improve the quality of the PF solutions and to discard the weak optimal solutions [34]. This is the one which is adopted in this research.

The approximated PF of the MO problem represented in Eq. (1) can be obtained by simplifying it into simpler scalar SP. The objective function of the j^{th} SP is given as in Eq. (8).

minimize

$$g^{ACh}(x \mid w^j, z^*) = max_{1 \le i \le N}\{ w_i^j \left| \frac{f_i(x) - z_i^{nad}}{z^* - z_i^{nad}} \right| \} + \rho \sum_{i=1}^{N} |f_i(x) - z_i^*| \quad (8)$$

Such that: $z^* = (z_1^*, z_2^*, \dots \dots z_m^*)^T$ is the reference point, i.e., $z_i^* = \max\{f_i(x) | x \in \Omega\}$ for $i = 1 \rightarrow m$ objectives and z_i^{nad} is the nadir point, where $z_i^{nad} = \min\{f_i(x) | x \in \Omega\}$.

The second part of the previous equation is the augmentation term, where ρ is the augmentation parameter which is a very small value such that $\rho \in [0.001, 0.1]$ [35].

Each SP is assigned a separate weight vector, such that each vector has m elements one for each objective $w = \{w_1, w_2, \dots . w_m\}$, where each element value ≥ 0 and the summation of all elements equals 1.

A neighborhood technique is utilized to optimize each SP based on its neighboring SPs. The j^{th} SP neighborhood comprises the set of SPs that are with T distance from j, such that T represents the neighborhood size. The steps of the original MOEA/D can be found in [18].

The proposed algorithm (algorithm 1) aims to optimize the parameters of three algorithmic trading approaches over three challenging objective functions. In this research, a new MOPSO/D algorithm is implemented, where the algorithm starts by randomly generating the initial particles' positions, velocities, such that, $P^j(t)$ and $v^j(t)$ are the current position and velocity of particle j at time t in sequence. In our application, the particles' positions represent the indicators' parameters.

The PSO considers the interactions and movements of particles as a swarm to locate the optimal points in the search space. Over different iterations, the particles cluster into an ideal position in the search space by employing both exploration and exploitation. The particles attempt to improve their position in the known beneficial regions by exploitation, while also exploring undiscovered regions of the feasible space.

Both exploitation and exploration processes can be managed during the velocity update step. Such that the accelerators c_1 and c_2 are used for the exploitation purposes, whereas. The inertia component ω is needed for exploration.

Algorithm 1: The proposed MOPSO/D algorithm for BTC trading

Inputs:

'C historical prices.

: The number of SPs.

The number of objectives.

: An initial set of weight vectors.

The size of the neighborhood.

ıpty External Archive (EA).

Steps:

1. Initialization:

- Generate a swarm of particles $x^1 \rightarrow x^N$ at random.
- For each particle j, initiate the current position P^j. The personal and global best positions, i.e., P^j_{Pbest}, and P^j_{Gbest} are initially set as P^j.
- For each particle j, initialize the velocity v^j=0.
- Initialize the reference point z^* and nadir point z^{nad}
- Calculate the Euclidean distances for each couple of weight vectors.
- Define the neighborhood of each particle j as, $B(j) = \{j_1, \cdots, j_T\}$, where $w^{j_1}, \cdots, w^{j_T}$ are the T closest weight vector to w^j.

2. Update

For (each iteration i = 1→ $iter$), do

For (each particle j = 1→ N), do

Calculate: g^{ACh}.

According to the g^{ACh} *value*, update P^i_{Pbest} and P^i_{Gbest}.

Update the reference and nadir points.

Update the velocity as:

$$v^j(t+1) = \omega\, v^j(t) + c_1 r_1 \left(P^j_{Pbest} - P^j(t)\right) + c_2 r_2 \left(P^j_{Gbest} - P^j(t)\right)$$

where, ω is the inertia component, c_1 and c_2 are two predefined constant accelerators, r_1, and r_2 are two random variables ∈ [0,1].

Update the current position as:

$$P^j(t+1) = P^j(t) + v^j(t+1)$$

Apply mutation operator.

Update the EA.

End

Update weight vectors W (algorithm 2).

Update neighborhood $B(j)$.

End

3. Return EA.

For further exploration of the search space, a uniform mutation is added to the algorithm to enhance the diversity of the current algorithm; however the resultant particles after mutation are subjected to a repair process that ensures that the new particles positions are never worse than the previous ones. Finally, the nondominant solutions are returned in a final External Archive (EA).

A. Weight Assignment

As mentioned before, the weight assignment method plays a crucial role in influencing the search process of MOEA/D algorithms. The diversity or the distribution of the weight vectors affects the quality of the final solutions to a high degree. Similar or identical vectors generate poor solutions that could not efficiently cover the whole PF.

The methods for generating weight vectors can be categorized as either uniform or random weight creation. Uniform weight creation involves constructing weights in repetitive patterns to guarantee the equitable distribution of vectors over the PF.

In [17], Zhang introduced a structured or uniform weight distribution architecture, such that each weight vector ∈ $\left\{0, \frac{1}{H}, \frac{2}{H}, \ldots, \frac{H}{H}\right\}$, with H being a positive integer parameter used for regulation. The number of SPs or weight vectors $N = C^{m-1}_{H+m-1}$ (C refers to the mathematical combinations).

The uniform distribution is effective in problems with continuous PFs. Nevertheless, it is not suitable for problems with complicated or scattered PFs [36]. A further problem with uniform distribution is that it occasionally generates vectors that are similar or extremely near to each other, resulting in duplicate solutions.

On the contrary, the random distribution of weights allows for a more comprehensive investigation of the search space, as it produces vectors that are not necessarily evenly distributed throughout the PF [18]. This, in turn, yields a range of distinctive and varied solutions since it generates vectors that are dissimilar from each other. The drawback with randomly distributed weights is that there is no assurance that the resulting vectors can accurately represent the whole PF [18].

Since the shape of the (PF) for real world problems is scattered and cannot be easily covered, finding the Pareto optimal points cannot be obtained by simple search strategies. To overcome this problem, an alternative scenario suggesting a Dual Weight assignment mechanism has been implemented denoted as (MOPSO/D-DW).

As seen, both uniform and random assignments have their benefits and drawbacks. In this study, a new weight assignment method is developed to combine the benefits of both strategies into a single algorithm (Algorithm 2).

The proposed strategy is a simple yet efficient. The main idea is to switch between the uniform and random generations over the different iterations.

This process ensures keeping the whole PF covered through the uniform distribution iterations while hitting other random areas during the random iterations. In this case, the particles neighbors are recalculated during each iteration which

improves the particles experience through the search process improving both the diversity and convergence.

Algorithm 2: Dual weight assignment algorithm

Inputs:

- *H:* A regulating integer parameter greater than zero.
- *m:* The number of objectives.
- *iter:* The number of iterations.

Steps:

1. Calculate the number of weight vectors $= C_{H+m-1}^{m-1}$.
2. Let U be a set of values in the range [0,1] with an increment of $1/\mathrm{H}$. U = $\{0, 1/\mathrm{H}, 2/\mathrm{H}, \ldots \ldots 1\}$.
3. Evaluate

For (*j*=1→*N),* do

  For (*i*=1→*m),* do

    - Generate a new non repeated weight vector w_i^j, such that $w_i \in U$ and $\sum_{i=1}^{m} w_i = 1$.
    - Append w_i^j to $\mathrm{W_U}$, such that $\mathrm{W_U}$ is the set of all uniformly generated vectors.

  End

End

For (*i*=1→ *iter*), do

If (*i* % 2== 0*)*, then

For (*j*=1→*N),* do

    - Generate a weight vector w_i^j, with w_i randomly selected from [0,1] and $\sum_{i=1}^{\mathcal{N}} w_i = 1$.
    - Append w_i^j to W.

  End

Else W← W_U

End

End

4. Return W

To summarize the optimization process in a more understandable way, the algorithmic trader first generates a set of random indicators' parameters (particles positions) along with a collection of a uniformly generated weight vectors one for each particle. The particles positions are then assessed by applying the indicators parameters to the training data set and calculating the resultant values for each of the objectives.

Moreover, a fitness aggregation technique in Eq. (8) is then employed to aggregate the three objectives. Based on the aggregated fitness value, the personal and global best positions are updated which in turn are used for updating both the velocities and positions of the particles. The algorithm continues till the maximum iterations and the final results (the best parameters) are obtained from the EA. These parameters are then tested on the testing data set.

IV. Experimental Results

The proposed algorithm is used to find the optimal trading signals for (BTC versus USD) trading during the interval from 3/1/2017 till 3/1/2019 which is considered as the training interval. The optimal set of parameters found during the search are tested upon the closing prices during the testing interval 3/1/2019 till 3/1/2021.

A. Evaluation Metrics

An evaluation metric serves as an indicator or assessment of the quality of the solutions that have been developed. Multiple measures exist for evaluating MO algorithms, each designed to assess distinct characteristics [37]. The convergence, diversity, and statistical measures are crucial assessment factors. Various metrics have been discovered that assess either a single criterion or multiple criteria concurrently. The study included three distinct indicators: the Generational-Distance (GD), the Hypervolume (HV), and the Average Fitness (AF).

- The GD serves as a metric to quantify the distance between the produced non-dominated solutions (the estimated PF) and the actual or true PF [38]. In real-world problems, it is possible to utilize a reference set derived from the collection of estimated PFs produced from all the search techniques under consideration. The lowest GD value indicates the proximity of the derived solution set to the true PF or reference set, and conversely, the higher the GD value, the more away the solution set is from the true Pareto Front [38].
- The HV indicator quantifies both the diversity and the convergence. It measures the *m*-dimensional volume of the objective space region, which is defined by the estimated PF and a reference point that is dominated by all solutions in the front [37].
- The AF is the average of the fitness values of the PF solutions found along a set of independent runs.

B. Parameter Settings

The parameter settings are as follows: the swarm size N = 351, the uniform weight regulating parameter H = 25. The total number of iterations is 150. The augmentation variable ρ = 0.05. The mutation rate=0.15. A neighborhood size (T) = 20. For PSO parameters, the inertia component ω =0.8, both constant accelerators c_1 and c_2 are set as 0.5. For the original MOEA/D, the crossover rate =0.8 and the other parameters are kept as before.

The indicators parameters are as follows: the time spans for all the indicators ∈ [3, 120] days. The ES-RoC over bought and sold levels range within ±10% from the center line. The S-RSI overbought level ∈ [60, 90], while the oversold level ∈ [10, 40].

C. Results

To evaluate convergence and diversity of the solutions generated by the proposed algorithm (MOPSO/D-DW) against the original MOEA/D, each algorithm is evaluated over 10 independent runs. The reported results in the following tables are the best and average values for each algorithm over these runs.

Table I shows a comparison between the values of the best and average GDs obtained by both algorithms. The best results among the two alternatives are displayed in bold. Again, the best GDs are the lowest values.

Table II shows a comparison between the best and average HVs obtained by the algorithms, where the best values (the higher values) are also shown in bold as before.

As seen from the tables, the proposed MOPSO/D-DW algorithm showed the best values in terms of both metrics, i.e., GD and HV. The proposed strategy always showed lower GDs and higher HVs which ensures closer solutions to the true PF with the best distribution of solutions.

To evaluate the efficiency of trading signals generated by the proposed algorithm, the AF of the obtained solutions over 10 independent runs are compared against both the MOEA/D and the TIs using their original parameters.

Due to the conflicts among the different objectives, the evaluation of the counterpart strategies is performed through a ranking methodology, such that each objective is ranked as compared to the same objective over the three counterpart strategies. The best rank is given a value of one, while the worst is given a rank of three. Finally, the total ranks obtained by each strategy are summed in order to evaluate its overall performance.

Table III shows a comparison of the three trading strategies, i.e., the TIs using their original parameters, the MOEA/D and the proposed MOPSO/D-DW during training.

TABLE I. A COMPARISON OF THE BEST AND AVERAGE GD OBTAINED BY EACH ALGORITHM

		MOEA/D	MOPSO/D-DW
DWMA	Best	1.40E-03	**0.00E+00**
	Average	2.53E-02	**0.00E+00**
ES-RoC	Best	5.91E-02	**3.20E-03**
	Average	7.83E-02	**4.34E-03**
S-RSI	Best	4.32E-01	**1.73E-01**
	Average	8.74E-01	**2.95E-01**

TABLE II. A COMPARISON OF THE BEST AND AVERAGE HV OBTAINED BY EACH ALGORITHM

		MOEA/D	MOPSO/D-DW
DWMA	Best	9.09E-02	**9.72E-02**
	Average	5.98E-02	**9.72E-02**
ES-RoC	Best	2.00E-03	**2.19E-02**
	Average	1.54E-03	**1.14E-02**
S-RSI	Best	5.06E-02	**2.91E-01**
	Average	2.97E-02	**2.64E-01**

TABLE III. THE AF OBTAINED BY THE THREE TRADING STRATRGIES FOR EACH INDICATOR OVER 10 INDEPENDENT RUNS DURING TRAINING

		Original parameters			MOEA/D			MOPSO/D-DW		
		RoI%	**SR**	**Trades**	**RoI%**	**SR**	**Trades**	**RoI%**	**SR**	**Trades**
DWMA	AF	518.49	0.87	10	559.91	0.88	5.92	704.18	1.08	7.41
	Rank	3	3	3	2	2	1	1	1	2
ES-RoC	AF	-29.85	-0.07	7	102.19	5.03	1.46	264.27	7.3	2.44
	Rank	3	3	3	2	2	1	1	1	2
S-RSI	AF	-75.6	-0.34	59	216.3	0.37	17.97	437.97	0.57	16.98
	Rank	3	3	3	2	2	2	1	1	1
Total ranking			27			16			11	

TABLE IV. THE AF OBTAINED BY THE THREE TRADING STRATRGIES FOR EACH INDICATOR OVER 10 INDEPENDENT RUNS DURING TESTING

		Original parameters			MOEA/D			MOPSO/D-DW		
		RoI%	**SR**	**Trades**	**RoI%**	**SR**	**Trades**	**RoI%**	**SR**	**Trades**
DWMA	AF	**983.77**	**1.03**	7	627.34	1.1	**6.26**	804.8	1.01	7.94
	Rank	1	1	2	3	3	1	2	2	3
ES-RoC	AF	-4.29	0.02	50	72.76	0.16	**20.06**	**76.63**	**0.28**	22.65
	Rank	3	3	3	2	2	1	1	1	2
S-RSI	AF	25.17	0.42	50	87.58	1.68	0.63	**203.11**	**2.05**	**1.01**
	Rank	3	3	3	2	2	2	1	1	1
Total ranking			22			18			14	

As seen from the table, the proposed MOPSO/D-DW algorithm could provide the best ranking followed by the MOEA/D with the TI's original parameters are given the worst ranking.

The same process is repeated in order to evaluate the benchmark methodologies during testing. As seen from Table IV, the proposed algorithm again provided the best ranking followed by MOEA/D.

Fig. 2 shows the summation of the ranks obtained by the last tables during both training and testing for each of the algorithms. Again, the algorithm assigned the lowest sum of ranks is the best. As can be seen, the proposed algorithm could achieve the best ranking during both the periods under study.

For further analysis, the best and median values obtained by the proposed algorithm are examined during both training and testing periods, such that the comparison is performed based on the average performance of each objective independently. In this case, the best obtainable RoI values for each indicator (i.e., DWMA, S-RoC, S-RSI) are averaged during both training and testing, and so on for the rest of the objectives. This process is repeated for the median value for each objective during both training and testing. The best and median values are compared to the original TIs in order to check the overall performance of the generated trading strategies in different ways, especially during the testing period, which is the main challenge.

Fig. 3 shows the best and median RoI values obtained by the proposed algorithm MOPSO/D-DW over 10 runs during both training and testing compared to the original TI's. As seen, during training, the best and median RoI values extremely exceed the indicators' original parameters RoI. During testing, it can be noted that both the median and the original indicators' parameters provide returns (RoI) that are close to each other, however, the median RoI of the proposed algorithm is still higher than the original parameters.

Fig. 4 shows the best and median SR obtained by the proposed algorithm versus the original indicators during both training and testing. Again, SR is a measure of the risk with higher values indicate lower risk. It can be seen that, the proposed algorithm provides better SR in all cases.

To figure out the effect of the selected parameters on the final trading signals in more detail, an example showing the difference between the effect of trading using the DWMA original parameters, i.e., 20-50 days and one of the generated solutions by the proposed algorithm, i.e., 18-38 days during the testing period for BTC, is clarified through Fig. 6 and Fig. 7 consequently.

Fig. 5 shows the best and median number of trades obtained during the two periods. In this case each buy-sell pair is considered as a single trade. As previously mentioned, lower values are always required as this in turn reduces the trading commissions. The proposed algorithm could also maintain the best number of trades in all the cases.

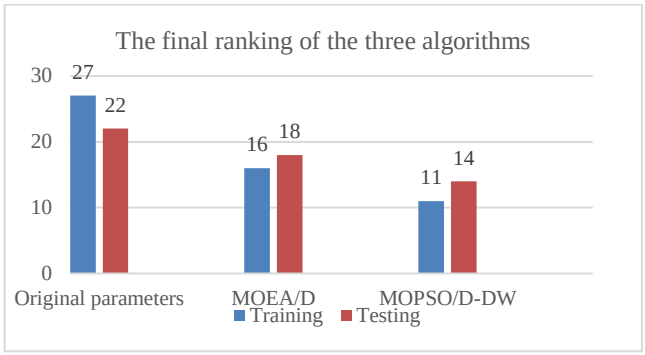


Fig. 2. The final summation of the ranks generated by the three trading strategies.

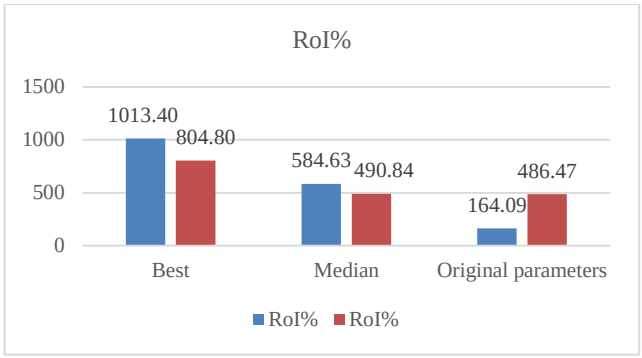


Fig. 3. The best and median RoI values obtained by the MOPSO/D-DW compared against the original TIs during both training and testing.

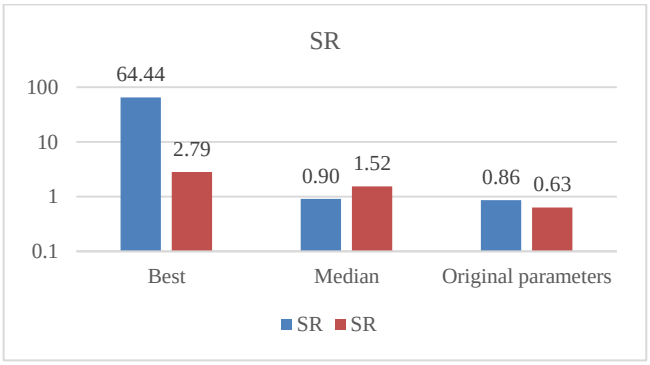


Fig. 4. The best and median SR values obtained by the MOPSO/D-DW compared against the original TIs during both training and testing.

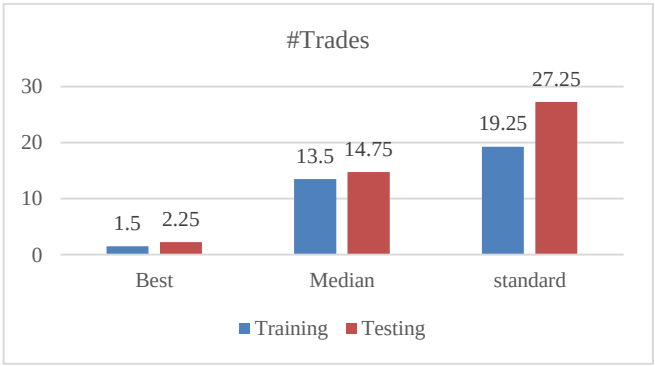


Fig. 5. The best and median number of trades obtained by the MOPSO/D-DW compared against the original TIs during both training and testing.

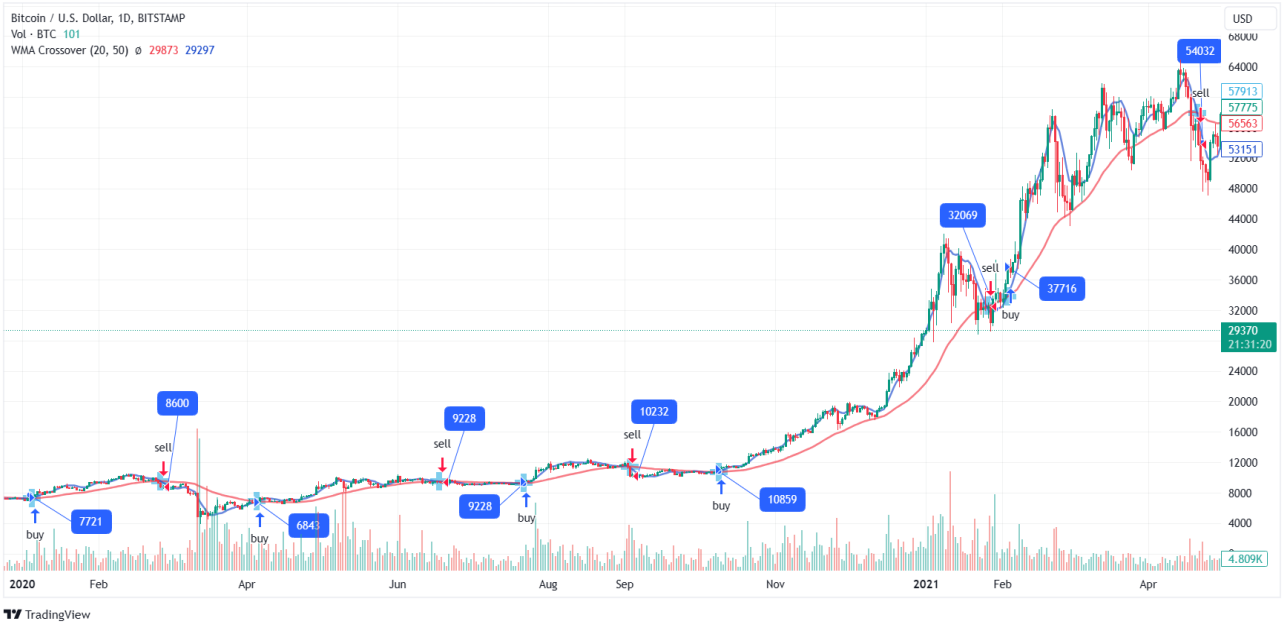


Fig. 6. The Crossovers between an 20-50-days DWMA indicator (the red line for the short WMA, whereas the blue line is for the long WMA) for BTC / USD trading (The buying and selling prices are highlighted in blue in each case).

Fig. 7. The Crossovers between an 18-38-days DWMA indicator (the red line for the short WMA, whereas the blue line is for the long WMA) for BTC / USD trading (The buying and selling prices are highlighted in blue in each case).

The figures show the set of generated signals, i.e., buy and sell, by each parameter combination, showing the buying and selling price in each case. As seen, both of the parameter sets provide a set of profitable trades during the period under study; however, the generated parameter combination in this case is more sensitive to market changes, providing more profitable trades.

V. Conclusion

Cryptocurrency, specifically Bitcoin (BTC), is a decentralized digital payment system that takes its name from the encryption mechanism used for verifying their digital transactions. As there is no need for traditional banking strategies, it attracted millions of traders all over the world. Crypto markets always suffer from instability or volatility which either yields to high returns or extremely harmful losses.

So, the study and analysis of the market is a major demand for investors to get the best returns with the least possible risks.

In this research, a new algorithm is proposed to optimize the trading signals' allocation strategies found in the literature named TIs. The algorithm proposes a new MOPSO/D that is based on a new dual weight assignment methodology MOPSO/D-DW. The algorithm is used to optimize three of the most famous and widely used TIs named DWMA, ES-RoC, and S-RSI.

The algorithm is compared to the MOEA/D and the TIs with their original parameters based on three objectives. The RoI is used for the evaluation of the investment returns, the SR is used for risk evaluation and the third objective is a calculation of the total number of trades. Results showed that the proposed algorithm showed promising results in terms of all the evaluation metrics during the training and testing intervals.

The problem with this optimization process is that the parameters that provide the best returns during the training period are not always the best during testing, but they still maintain good performance with high returns or at least no general losses. This is due to the extreme volatility found during both training and testing, such that the market shows a sideway performance during the testing period, with an extreme jump at the end of the period. For future research, a more complex normalization process could be tested instead of the linear normalization used in this research. The optimization process can be extended to more indicators and a large number of cryptocurrencies. The proposed dual weight-generation strategy could also be tested under different conditions.

REFERENCES

[1] M. V. Alstyne, "Why Bitcoin Has Value," *Communications of the ACM*, vol. 57, no. 5, pp. 30-32, 2014.

[2] N. Daskalakis and P. Georgitseas, An Introduction to Cryptocurrencies: The Crypto Market Ecosystem, London, 2020.

[3] J. Chen, "Analysis of Bitcoin Price Prediction Using Machine Learning," *J. Risk Financial Manag.*, vol. 16, no. 51, 2023.

[4] M. J. Hamayel and A. Y. Owda, "A Novel Cryptocurrency Price Prediction Model Using GRU, LSTM and bi-LSTM Machine Learning Algorithms," *AI*, vol. 2, no. 4, pp. 477-496, 2021.

[5] P. Jaquart, D. Dann and C. Weinhardt, "Short-term bitcoin market prediction via machine learning," *The Journal of Finance and Data Science*, vol. 7, pp. 45-66, 2021.

[6] G. Borrageiro, N. Firoozye and P. Barucca, "The Recurrent Reinforcement Learning Crypto Agent," *IEEE Access*, vol. 10, pp. 38590-38599, 2022.

[7] L. Lorenzo and J. Arroyo , "Online risk-based portfolio allocation on subsets of crypto assets applying a prototype-based clustering algorithm," *Financial Innovation*, vol. 9, no. 25, 2023.

[8] B. Joemon, M. A. Ghazanfar, M. A. Azam, N. Z. Jhanjhi and A. A. Khan, "Novel heuristics for Stock portfolio optimization using machine learning and Modern Portfolio Theory," in *International Conference on Business Analytics for Technology and Security (ICBATS)*, Dubai, 2023.

[9] P. Hrytsiuk, T. Babych and L. Bachyshyna, "Cryptocurrency Portfolio Optimization Using Value-At-Risk Measure," in *6th International Conference on Strategies, Models and Technologies of Economic Systems Management (SMTESM 2019)*, 2019.

[10] Y. He and . C. Aranha, "Solving Portfolio Optimization Problems Using MOEA/D and L´evy Flight," *ArXiv*, vol. abs/2003.06737, 2020.

[11] Q. Zhang, H. Li, D. Maringer and a. E. Tsang, "MOEA/D with NBI-style Tchebycheff approach for Portfolio Management," in *IEEE Congress on Evolutionary Computation*, Barcelona, Spain , 2010.

[12] D. Philippas, N. Philippas, P. Tziogkidis and H. Rjiba, "Signal-herding in cryptocurrencies," *Journal of International Financial Markets, Institutions and Money*, vol. 65, 2020.

[13] M.-F. Leung, L. Chan, W.-C. Hung, S.-F. Tsoi, C.-H. Lam and Y.-H. Cheng, "An Intelligent System for Trading Signal of Cryptocurrency Based on Market Tweets Sentiments," *FinTech.*, vol. 2, no. 1, pp. 153-169, 2023.

[14] D. Garcia and F. Schweitzer, "Social signals and algorithmic trading of Bitcoin," *R. Soc. open sci.*, vol. 2, no. 9, 2015.

[15] T. E. Koker and D. Koutmos, "Cryptocurrency Trading Using Machine Learning," *Risk and Financial Management*, vol. 13, p. 178, 2020.

[16] K. Deb, Multi-Objective Optimization Using Evolutionary Algorithms: An Introduction", In: Wang, L., Ng, A., Deb, K. (eds) Multi-objective Evolutionary Optimisation for Product Design and Manufacturing. Springer, 2011.

[17] Q. Zhang and a. H. Li, "MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 6, pp. 712 - 731, 2007.

[18] S. M. Omran, W. H. El-Behaidy and A. A. A. Youssif, "Decomposition Based Multi-objectives Evolutionary Algorithms Challenges and Circumvention," *Intelligent Computing*, vol. 1229, pp. 82-93, Advances in Intelligent Systems and Computing. Springer SAI 2020.

[19] S. Jiang, S. Yang, Y. Wang and a. X. Liu, "Scalarizing Functions in Decomposition-Based Multiobjective Evolutionary Algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 2, pp. 296 - 313, 2018.

[20] M. Grabisch, J.-L. Marichal, R. Mesiar and E. Pap, "Aggregation Functions. Part I: Means," *Information Sciences, Elsevier*, vol. 181, no. 1, pp. 1-22, 2011.

[21] M. Pescador-Rojas and a. C. A. C. Coello, "Collaborative and Adaptive Strategies of Different Scalarizing Functions in MOEA/D," in *IEEE Congress on Evolutionary Computation (CEC)*, 2018.

[22] X. Ma, Q. Zhang, G. Tian, J. Yang and a. Z. Zhu, "On Tchebycheff Decomposition Approaches for Multi-objective Evolutionary Optimization," *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 2, pp. 226 - 244, 2018.

[23] L. R. C. d. Farias, P. H. M. Braga, H. F. Bassani and a. A. F. R. Araújo, "MOEA/D with Uniformly Randomly Adaptive Weights," in *GECCO '18*, Kyoto, Japan, 2018.

[24] X. Guo, X. Wang and a. Z. Wei, "MOEA/D with Adaptive Weight Vector Design," in *11th International Conference on Computational Intelligence and Security*, Shenzhen, China, 2015.

[25] M. Wu, S. Kwong, Y. Jia, K. Li and a. Q. Zhang, "Adaptive weights generation for decomposition-based multi-objective optimization using Gaussian process regression," in *GECCO '17 Proceedings of the Genetic and Evolutionary Computation Conference*, Berlin, Germany, 2017.

[26] K. Erwin and A. Engelbrecht, "Meta-heuristics for portfolio optimization," *Soft Computing*, 2023.

[27] S. Sarkar, S. Das and a. S. S. Chaudhuri, "Multi-Level Thresholding with a Decomposition-based Multi-Objective Evolutionary Algorithm for Segmenting Natural and Medical Images," *Applied Soft Computing*, vol. 50, pp. 142-157, 2017.

[28] H. Xing, Z. Wang, T. Li and a. H. Li, "An Improved MOEA/D Algorithm for Multi-objective Multicast Routing with Network Coding," *Applied soft computing* , vol. 59, pp. 88-103, 2017.

[29] K. Metaxiotis and K. Liagkouras, "Multiobjective Evolutionary Algorithms for Portfolio Management: A comprehensive literature review," *Expert Systems with Applications*, vol. 39, no. 14, pp. 11685-11698, 2012.

[30] K. Frajtova-Michalikova, E. Spuchľakova and M. Misankova, "Portfolio Optimization," in *4th World Conference on Business, Economics and Management, WCBEM*, 2015.

[31] M. A. Lim, The Handbook of Technical Analysis, Singapore: John Wiley & Sons , 2016.

[32] Y. Su, C. Cui and H. Qu, "Self-Attentive Moving Average for Time Series Prediction," *MDPI in Applied sciences*, vol. 12, 2022.

[33] X. Ma, Q. Zhang, G. Tian, J. Yang and a. Z. Zhu, "On Tchebycheff Decomposition Approaches for Multi-objective Evolutionary Optimization," *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 2, pp. 226 - 244, 2018.

[34] T. Chugh, "Scalarizing Functions in Bayesian Multiobjective Optimization," in *2020 IEEE Congress on Evolutionary Computation (CEC)*, Glasgow, UK, 2020.

[35] K. Dachert, J. Gorski and K. Klamroth, "An augmented weighted Tchebycheff method with adaptively chosen parameters for discrete bicriteria optimization problems," *Computers & Operations Research*, vol. 39, pp. 2929-2943, 2012.

[36] X. Chen, J. Yin, D. Yu and X. Fan , "A decomposition-based many-objective evolutionary algorithm with adaptive weight vector strategy," *Applied Soft Computing*, vol. 128, 2022.

[37] C. Audet, J. Bigeon, D. Cartier, S. Le Digabel and L. Salomon, "Performance indicators in multiobjective optimization," *European Journal of Operational Research*, vol. 292, no. 2, pp. 397-422, 2020.

[38] Miqing Li and T. Chen, "How to Evaluate Solutions in Pareto-based Search-Based Software Engineering? A Critical Review and Methodological Guidance," *IEEE Transactions on Software Engineering*, vol. 48, no. 5, pp. 1771-1799, 2022.

A New Method for Revealing Traffic Patterns in Video Surveillance using a Topic Model

Yao Wang*
Henan Industry and Trade, Vocational College, Zhengzhou Henan, 450053, China

***Abstract*—Research on video surveillance systems, for instance, in intelligent transportation systems, has advanced due to the growing requirement for monitoring, control, and intelligent management. One of the next issues is extracting patterns and automatically classifying them, given the volume of data produced by these systems. In this study, a theme approach was utilized to translate visual patterns into visual words in order to reveal and extract traffic patterns at crossings. The supplied video is first cut up into segments. The optical flux technique is then used to determine the clips' optical flux characteristics, which are based on a lot of local motion vector data, and translate them into visual words. The thin-group thematic coding method is then used to teach traffic patterns to the proposed system using a non-probable thematic model. By responding to a behavioral query like "Where is a vehicle going?" these patterns convey observable motion that can be utilized to characterize a scene. The results of applying the suggested method to the QM_UL video database demonstrated that the suggested method can accurately identify and depict significant traffic patterns such as left turns, right turns, and intersection crossings.**

Keywords*—Group thin topic coding; QM_UL video; optical flux; traffic patterns

I. Introduction

The deployment of surveillance cameras is one of the most common strategies for boosting security and managing public spaces better. The majority of public spaces and intelligent transportation networks virtually always have these cameras in situ [1]. These cameras are crucial in video surveillance because they may give decision-makers access to all the details and activities occurring in the monitored area [2]. For instance, in smart transportation systems, these binoculars can be used to monitor and control the volume and kind of traffic, vehicle offenses, and traffic patterns at crossings. On the one hand, the use of these video surveillance cameras has become quite widespread [3]. On the other hand, as their use has increased, a vast amount of multimedia data has been produced, making it virtually impossible for people to monitor these cameras, necessitating the need for a system. Intelligent monitoring, automatic pattern detection, and pattern extraction are all clearly audible [4]. There are two sorts of studies that are accessible for the examination of traffic scenes: a) studies based on the route. A training dataset for machine learning is created using paths that have been observed over a lengthy period of time [5] and [6]. In fact, such a study is still unreliable in challenging circumstances because of the lack of steady and dependable multipurpose tracking algorithms [7]. Rapid response to unexpected shifts in traffic [8] also commonly causes analysis problems. b) Analysis that does not involve tracking and is based solely on the low-level motion vector Optical flux [9] is the most commonly used technique and contains a lot of data about local movement. Using this simple motion feature, more complex models, like thematic ones, might be built for evaluating complex traffic scenarios [10]. Some examples of probabilistic topic models are the Probabilistic Latent Structure Analysis [11], the Probabilistic Conceptual Latent Analysis [12], the Dirichlet Latent Allocation LDA [13], and the Hierarchical Dirichlet Process HDP [14]. were originally developed to unearth concealed headers in a massive corpus of text documents before they were adopted by academics for video analysis. In [15], a fully sparse topic model (TM) (FSTM) is proposed as a simple variant of LDA and PLSA. According to research [16], thin topic coding (STC) using a non-probabilistic topic model (NPTM) can be used to learn hierarchical hidden representations of enormous data sets. Group Thin Topic Coding (GS-TC), a brand-new non-probabilistic topic model, is suggested in [17] for learning thin hidden representations of big text document sets. An effort has been made in [18] to highlight odd and uncommon actions, such as the fire engine stopping the regular flow of traffic. Thin group thematic coding has been enhanced for this purpose, and the thin light flux method has also been applied to extract movement patterns. When compared to previous similar works, the key distinction between our work and others is the utilization of dense light flux in this article to extract movement pattern features and the thematic coding of the primary thin group. Even though using dense light flux requires a larger computation area than using thin light flux, more movement patterns are still recovered. The approaches used to convert the prevalent visual traffic movement patterns into descriptive phrases allowed us to zero down on eight key traffic patterns within the dataset. These eight traffic patterns are: a) Crossing the intersection from the eastern side to the western side; b) Crossing the intersection from the southern side to the northern side; c) Turning right from the eastern side to the northern side; d) Turning left from the western side to the northern side; d) Turning left from the west to the northern side; e) Crossing the intersection from west to east; f) Turning right from south to east; g) Crossing the intersection from north to south; and h) Finding traffic patterns directly results in a useful scene model and streamlines scene analysis. Despite the fact that there have been numerous studies in this area, machine vision systems still struggle with this problem. The automatic detection of traffic patterns in this paper is accomplished using the group thin thematic coding framework (GS-TC). First, a series of separate, non-overlapping clips are created from the surveillance camera video. Then, the discrete

visual stream words are translated into discrete light flux characteristics for each pair of subsequent frames in the clips. The words in the visual stream are then translated into words contained in each video clip, which is then treated as a document. To put it another way, visual elements become visual words. In order to find hidden patterns that reveal the distribution of common motion in the scene, the GS-TC approach is then used. The results demonstrate that useful traffic patterns, such as left turn, right turn, and crossing the intersection, can be retrieved from intersections using the proposed method when applied to real footage. The reasons for choosing the "proposed method" that we consider suitable for dealing with such problems are the following:

1) Growth in video surveillance data: This study confirms the increasing use of surveillance cameras in various applications, leading to significant growth in video data. This increasing amount of data makes manual monitoring impractical and necessitates the development of automatic pattern recognition methods. The proposed method offers a solution to this problem by automatically identifying and classifying traffic patterns.

2) Communication with intelligent transportation systems: surveillance cameras are important in intelligent transportation systems. These systems need to monitor and control traffic patterns at intersections, which makes the proposed method particularly relevant to this field.

3) Using optical flux features: This method uses optical flux features extracted from video frames. Optical flux data contains information about local movement, which is very important for analyzing traffic scenarios. This research discusses the use of optical flux and its potential to build more complex models, such as thematic models, to assess traffic patterns.

4) Thematic coding approach: In this work, a thematic coding approach is chosen, which is usually used in natural language and text processing. Applying this approach to video data, they aim to transform visual traffic movement patterns into descriptive expressions that enable the identification and classification of traffic patterns.

5) Identifying common traffic behaviors: This article also emphasizes the importance of recognizing common traffic behaviors such as turning left, turning right, and moving straight through intersections. These behaviors are considered significant traffic patterns and the proposed method is designed to identify them. The authors' overall contributions to this paper can be summarized as follows:

- Providing a non-surveillance way for video surveillance to identify traffic trends.
- Using the thematic paradigm, drivers at crossings are instructed on traffic patterns.

The overview of related studies follows in Section II. Section III provides background information on the theory behind flow detection and traffic patterns. The traffic dataset, the execution strategy, and the useful outcomes are all covered in Section IV. Section V concludes with recommendations for additional research.

II. BACKGROUND

In Europe, road traffic noise is a big problem, exposing almost 20% of the population to dangerously high noise levels. In order to facilitate decision-making techniques aimed at controlling or reducing noise exposure, effective monitoring and measurement of sound levels in sensitive regions is essential [31]. However, expensive equipment and maintenance duties are needed for spatiotemporal measurements with continuous range and lengthy duration. Thus, the goal of the research [32] is to create an intelligent mobility approach that is affordable and can be used to estimate traffic noise levels using roadside video images. The created method consists of an algorithm that uses dynamic microscopic models to evaluate noise and extract traffic volume, identify vehicle classes, and estimate each vehicle's speed from video records. These later models are based on the already available noise emission models (NEMs) for estimating source sound power levels and a sound emission model that can estimate the corresponding A-weighted sound pressure levels given any road vehicle speed as input slow. The created method is distinguished by its modular structure, which makes it simple to add new variables to the sound propagation model and/or replace the NEM. The process is put to the test under various service levels on a medium-sized city's rural roads. The findings reveal that the noise estimation errors are less than 1 dBA, indicating great accuracy. The issue of traffic congestion in large cities is getting worse due to the swift growth in both the population living in metropolitan areas and the quantity of motor vehicles. The paper aims to perform cluster analysis for daily traffic congestion index curves in order to discover patterns of traffic congestion and examine their spatial-temporal changes. Initially, the coefficient of variation is used to apply weights in order to improve the K-means clustering method because the significance of sample points varies slightly over different time segments. A better K-means clustering algorithm is suggested to find patterns of traffic congestion. Second, changes in traffic congestion patterns over time and space are analyzed using the paired t-test method.

A. Exposing Traffic Patterns

Fig. 1 displays an example of traffic monitoring scenarios at an intersection using a surveillance camera positioned above the intersection. Traffic movement patterns are regular and circular traffic situations, such as "straight passage" and "right turn". The technologies and algorithms for detection, identification, tracking, and categorization of items suffer substantially as a result of the complexity of many of these settings. In these situations, topic models can be utilized to locate and highlight patterns in order to map words, documents, and subjects to certain pattern recognition ideas.

B. Coding of Thin Group Topic

Consider a collection of documents $D = \{w_1, \dots, w_D\}$ that includes N terms from the vocabulary collection V. A document is only a vector | I| after which $D = \{w_1, \dots, w_{|I|}\}$, appears, where I is a set of m word indices. The nth entry (n $\in$I) in w_n indicates the number of occurrences of the desired

word in a particular document. We consider the parameter $\beta \in R^{K*N}$ as a unitary 2 distribution in V is used to describe a dictionary with K bases, each of which is taken to be the foundation of the subject [19].

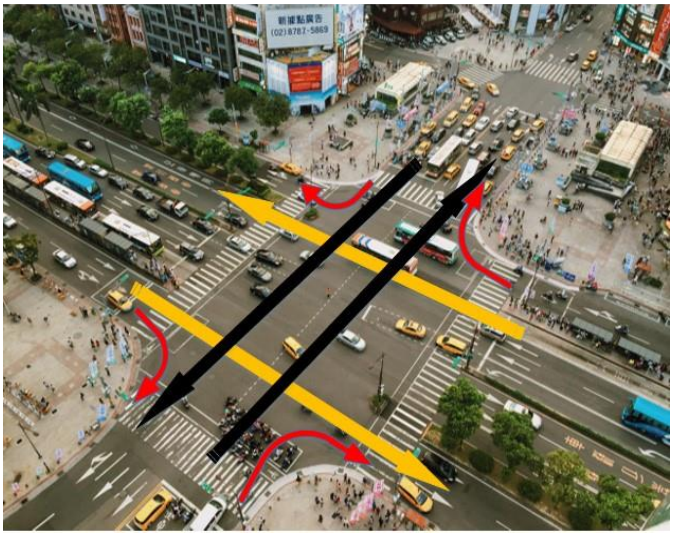


Fig. 1. An intersection traffic scene includes typical traffic movements including left turns, right turns, and crossing the intersection.

For the dth document (w_d), the GS-TC method maps it to a semantic space assigned by a set of auto-learned β addresses and directly encodes the non-normal word $s_{d,.m} \in R^K$ determines for each specific word in the document w_d.

Then the mixture ratio of the whole document w can be derived from the code word $s = \{s_{.1}, \ldots, s_{|I|}\}$ and the headings β. The optimization issue is resolved by the GS-TC approach in accordance with (1). The first part (1) is the non-normal difference KL between the observed words $w_{d,.n}$ and their reconstruction $\beta_{.n}\ s_{d,.m.}^T$.

$$\min \sum_{d=1}^{D} \sum_{m=1}^{|I_d|} \left(\sum_{k=1}^{K} S_{d,kn}\ \beta_{kn} - w_{d,m} \ln \left(\sum_{k=1}^{K} S_{d,km}\ \beta_{km} \right) \right) + \lambda \sum_{d=1}^{D} \sum_{k=1}^{K} \left\| S_{d,k.} \right\|_2 + C \quad (1)$$

$$\text{s.t}\ s_{d,.n} \geq 0\ \forall d, m$$

$$\sum_{k=1}^{K} \beta_{km} = 1, \forall k$$

$$\theta_k = \frac{\sum_{m=1}^{|I|} S_{km}\ \beta_{km}}{\sum_{m=1}^{|I|} \sum_{l=1}^{K} S_{\mathrm{Im}}\ \beta_{\mathrm{Im}}} \quad (2)$$

The second step involves using the LASSO approach to apply a variety of norms for the reconstruction coefficients matrix, which results in proportionate thin coding at the document level. The symbols used in (1) are defined in Table I. It should be noted that the word codes can be used to determine the mixture ratio of the document surface. The variable θ_k is used in (2) [20] to indicate the contribution vector of the k_{th} title in the document w.

TABLE I. THE DEFINITION OF THE SYMPTOMS AND VARIABLES UTILIZED IN (1) AND OPTIMIZATION OF THE GS-TC TECHNIQUE

Signs	Description
$\boldsymbol{m = 1, \ldots, N}$	Index_of words
$\boldsymbol{\theta_d}$	Percentage_of titles in the dth _document
$\boldsymbol{d = 1, \ldots, D}$	Document_index
β	Titles_Dictionary
$\boldsymbol{S_{d,m}}$	mth-word representation in dth-document
$\boldsymbol{w_{d,m}}$	How many times the word n appears_in the given document
$\boldsymbol{k = 1, \ldots, K}$	Index of _titles
$\boldsymbol{I_d}$	Set of index words _that appear in document dth

C. Related Works

Most methods for analyzing traffic can be classified into two broad categories. In the first category, the target object, such as a car or a person, is first recognized, followed by tracking, and the paths gleaned are then used for additional analysis. Undoubtedly, the grouping of traces is a straightforward technique that makes it possible to identify anomalies [21]. Regardless, the accuracy and dependability of these approaches heavily rely on the detection and tracking techniques, which are susceptible to errors brought on by obstruction, noise, shifting lighting conditions, and changing weather. Additionally, the grouping of routes necessitates a comparison of all samples' similarity, which can be computationally taxing [22]. In the second type, information about motion and appearance is gleaned from video frames without the aid of detection and tracking methods. The strategy utilized in these articles [7] and [23] is to directly develop a model of movements and activities using these extracted attributes. A sizable number of research projects have recently concentrated on the use of topic models (TM). For instance, utilizing probabilistic conceptual latent analysis (PLSA), anomaly detection and motion-based scene segmentation have been carried out in [24]. Using the two-level Dirichlet process (dp), normal and aberrant activity were distinguished in [25]. In [26], a two-level LDA topic model was applied to extracting rules and recognizing anomalies after it was utilized to identify anomalies and learn behavior for the efficient display of clips with a dispersed set of motion patterns. The author in [19] presents an unsupervised method for anomaly detection that uses the thin thematic group coding (GS-TC) framework to learn movement patterns. In [27], it is planned to use license plate number data obtained from video surveillance cameras to determine urban road vehicle density, city-wide regional vehicle density, and hot routes. To improve the precision of the visualization's impact, this article used a method for detecting outliers based on Dixon's detection approach and Internet crawling technologies during data analysis and processing. This study developed an urban road vehicle traffic index for the visualization map design in order to visually and

numerically depict the region's traffic operating state. An experiment was carried out in Guiyang using the information from the road video surveillance camera system to confirm the method's viability. From three visualization maps, a number of geographical and temporal characteristics of urban traffic are clearly and effectively identified. The outcomes demonstrate the suggested framework's satisfactory performance in terms of visual analysis, which helps with traffic management and operations. The author in [28] introduces DeWiCam, a smartphone-based detection system that is small and powerful. Utilizing the natural traffic patterns of wireless camera streams is the core concept behind DeWiCam. DeWiCam is more difficult to utilize than conventional traffic pattern analysis because it cannot access data that has been packed in data packets. DeWiCam, on the other hand, circumvents this issue and can accurately find neighboring Wi-Fi cameras. A human-assisted recognition approach is suggested to further determine whether a camera is interested in a certain room. Image resolution and audio channel inference are two additional DeWiCam add-on functions that are available to enhance further security. DeWiCam is put to use on the Android operating system and assessed using in-depth tests on 20 cameras. DeWiCam can identify cameras with 99% accuracy in 7:2 seconds, according to test data.

III. Proposed Method for Detecting Traffic Patterns

A. Process of the Proposed Method

Fig. 2 and 3 depict the general and specific steps of the suggested strategy that makes use of the topic model. When considering an input video, the video is first momentarily split into *D* non-overlapping clips, with each clip being treated as a separate w_d document. The scene is initially divided into $C_x * C_y$ square cells, each of which covers $p*p$ pixels, in order to construct stream words. After that, using the optical flux technique, motion vectors are extracted for each pair of subsequent frames. In order to eliminate noise and preserve dependable flows, a motion vectors are subjected to a threshold value, t. The other motion vectors $s_i = (x, y, u, v)$, whose positions (x and y) are fixed in a grid with a distance of p pixels, are sampled to create flow words. The examples of motion vectors include then split into an O-number of directions based on their displacement (u, v). Finally, a group of fixed words is generated, $N = C_X \times C_y \times O$ and $V = \{1, \dots N\}$, each of which has two content aspects: "position information" and "direction information" of movement. The frames include a collection of the video clips' stream words. Then, a video clip is shown as a vector with the form $w = (\omega_1, \dots, \omega_N)$, where ω_n denotes how many times the n_{th} word appears in the clip. The symptoms are depicted in Table II along with the video's counterparts.

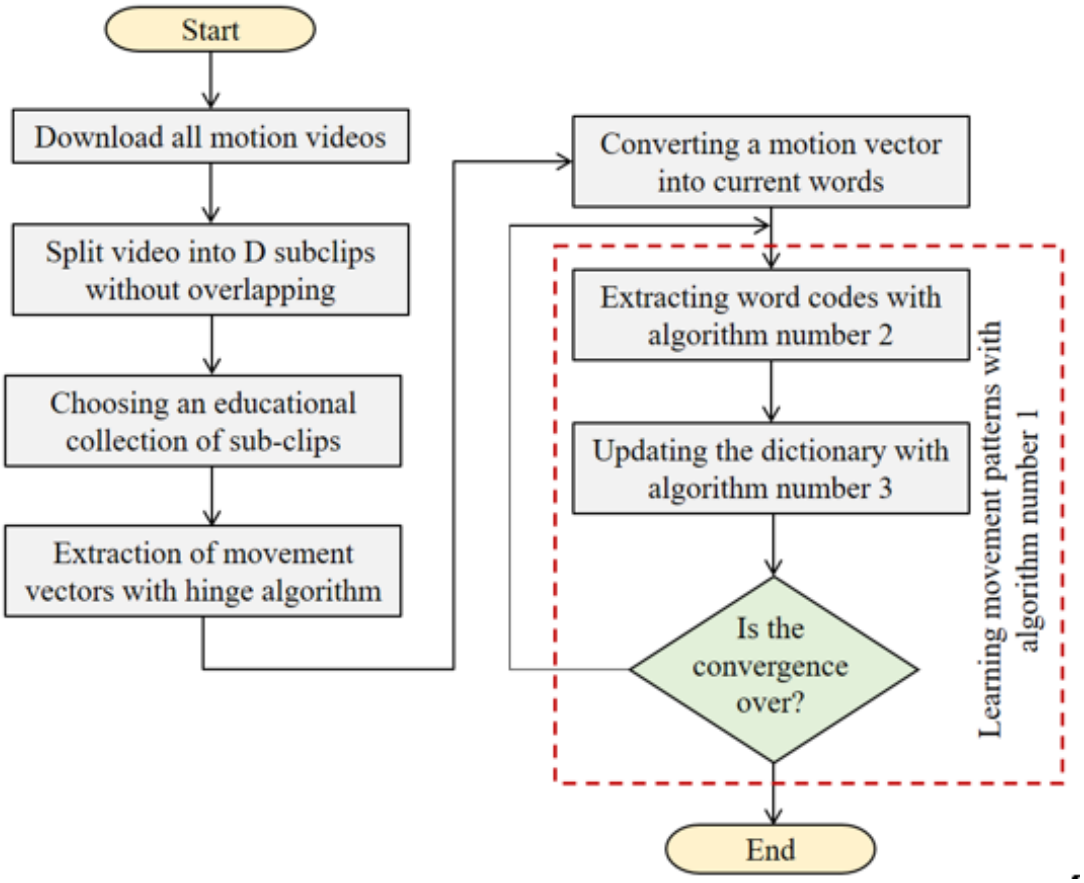


Fig. 2. The general process of the proposed method of detecting normal traffic patterns of vehicles.

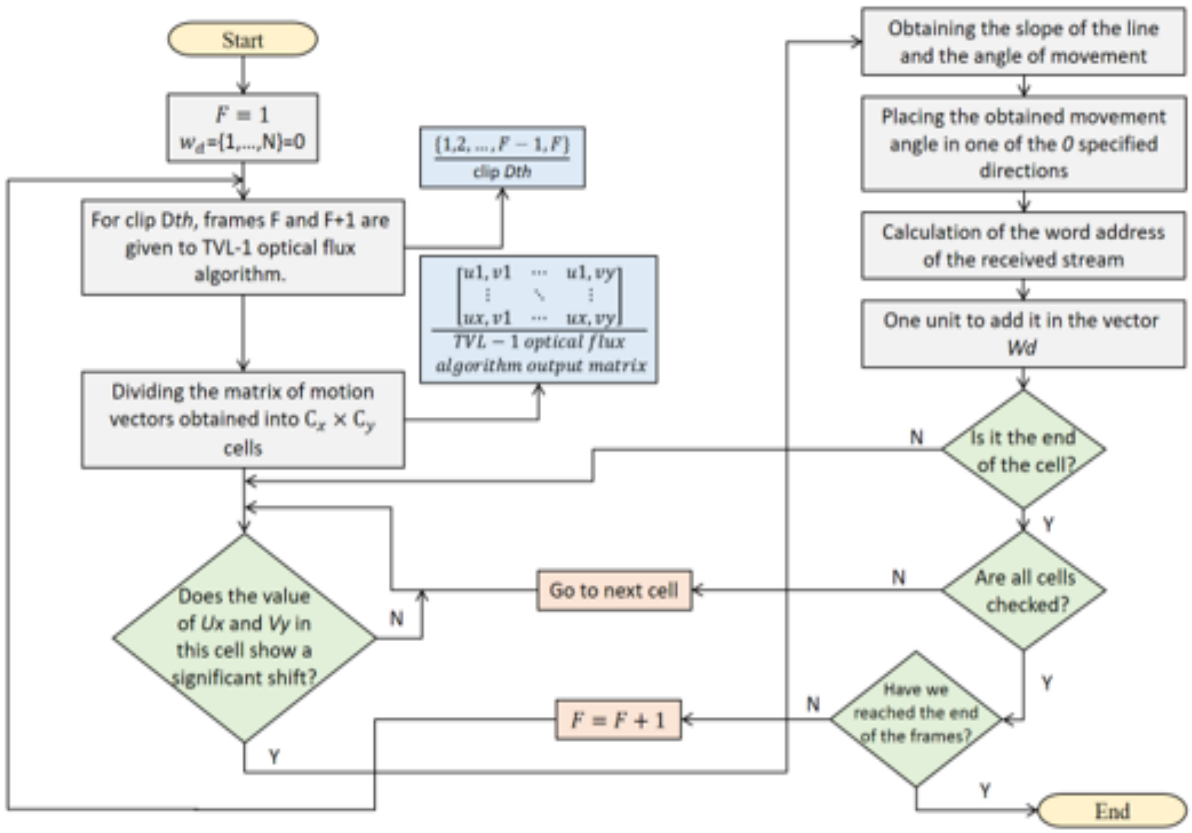


Fig. 3. Partial flowchart of the proposed method, video display with the subject model in detecting normal traffic patterns of vehicles at intersections. Identifying normal traffic patterns of vehicles.

TABLE II. SYMBOLS AND VARIABLES USED IN THE VIDEO, ALONG WITH THEIR EQUIVALENTS

Signs	Description
$\boldsymbol{d = 1, \dots, D}$	Index of clips
$\boldsymbol{k = 1, \dots, K}$	Index of movement patterns
$\boldsymbol{n = 1, \dots, N}$	A Word-Flow Index
$\boldsymbol{I_d}$	Index set of flow words that happened in clip dth
$\boldsymbol{w_{d,n}}$	number of occurrences _ word nth clip dth
β	A Movement Pattern Dictionary
$\boldsymbol{\theta_d}$	Clip share dth of templates
$\boldsymbol{S_{d,n}}$	How significant the nth stream word was in the dth clip

B. Learning Traffic Patterns using GS-TC

Eq. (3) provides a summary of the general GS-TC formula. The GS-TC formula in [18], which is based on reducing the variance of KL, does not apply to this equation because the 2 soft reduces reconstruction errors to a minimum. The optimization problem can be solved more easily by using soft minimization 2□ [21]. The objective function in (3) is convex, which means that while one of the two is constant, it is convex around $\{S_d\}_{d=1}^{D}$ and β. The CDA algorithm [29], which alternately optimizes on $\{S_d\}_{d=1}^{D}$ and β and is depicted in Algorithm 1, is a common solution.

Algorithm 1: Intelligent transportation systems' learning algorithms for understanding typical vehicle traffic patterns

Input: training video clips$\{w_d\}_{d=1}^{D}$, hyper parameter λ, and topic count
Output: dictionary β, word codes s
$\beta \in R^{K \times N}$ a positive-definite random matrix.
Initialize $\{S_d\}_{d=1}^{D} \in R^{D \times K \times N}$ to random matrices with positive elements
l^{old} = Using Eq. (3), figure out the cost function.
repeat
for $d = 1$: D
calculate S_d with Algorithm 2
end for
Apply Algorithm 3 to Dictionary β
l = A cost function can be calculated using formula (3).
If $(\boldsymbol{l - l^{old} < \varepsilon})$**then**
Break
Else
$l = l^{old}$ until the conditions for termination are met, usually convergence.

$$\min_{\{S_d\}_{d=1}^{D},\beta} \sum_{d=1}^{D} \left(\left| \left| w_d - diag(S_d^T \beta) \right| \right|_2^2 + \lambda \left| \left| S_d \right| \right|_{2,1} \right) \quad (3)$$

$$\text{s.t. } s_d \geq .0 \; \forall d, \beta_k \geq ., \left| \left| \beta_{k.} \right| \right|_1 \geq 1, \forall k$$

where $S_d \in R^{K\times N}$ and parameter λ is a non-negative parameter for thin control.

C. Extracting Word Codes using Thin Coding Algorithm

In the following, the word codes $\{S_d\}_{d=1}^{D}$ will be obtained by optimization according to Eq. (4) with the assumption of constant β dictionary.

Due to conditional independence, this step can be carried out separately for each document by addressing the optimization problem. Since the word codes are organized into titles and each title is unique, this optimization makes use of the BCD technique by solving the issue for each S_{kn}.

The difficulty of locating the roots of a complex quadratic equation is the end result of this approach. By expressing the optimization problem in accordance with Eq. (5) and (6), it is possible to acquire the entire matrix s without having to calculate each S_{kn} individually.

Algorithm 2: Thin coding algorithm

$P = diag(w_{d,1}, \dots, w_{d,N})$
l^{old} = Cost function calculation per formula (5)
Repeat
$R = diag(\frac{1}{\left\|S_{d,1.}\right\|_2 + \varepsilon}, \dots, \frac{1}{\left\|S_{d,k.}\right\|_2 + \varepsilon})$
$S_d = (\beta\beta^T + \lambda R)^{-1}\beta P$
l = Cost function calculation per formula (5)
If $(l - l^{old} < \varepsilon)$then
Break
Else
$l = l^{old}$
Until the convergence occurs or the end point is reached.

$$\min_{\{S_d\}_{d=1}^{D},\beta} \sum_{d=1}^{D}\left(\left||w_d - diag(S_d^T\beta)\right||_2^2 + \lambda\left||S_d|\right|_{2,1}\right) = \sum_{d=1}^{D}\left(\sum_{n=1}^{N}(W_{d,n} - S_{d,n}^T\beta_{.n})^2 + \lambda\sum_{k=1}^{K}\left\|S_{d,k.}\right\|_2\right) \quad (4)$$

$$\text{s.t. } s_d \geq .0, \forall d$$

$$\min_s \left||w_d - diag(s_d^T\beta)\right||_2^2 + \lambda\left||S_d|\right|_{2,1} \quad (5)$$

$$\text{s.t. } s > 0$$

$$\min_s trace(PP + s^T\beta\beta^T s - 2Ps^T\beta) + \lambda\, trace(s^T Rs) \quad (6)$$

$$\text{s.t. } s > 0$$

that in (6), $R = diag(\frac{1}{\left||s_{10}|\right|_2} + \varepsilon, \dots, 1/||S_{k.}||_2)$ and $P = diag(w_1, \dots, w_N)$ is. We consider the value of ε much smaller than the non-zero values of s and add it to the denominator to prevent division by zero. Since the norms of l_p for $p \geq 1$ are convex functions, according to (5) and its equivalent, Eq. (6), the problem is the optimization of convex functions with respect to the matrix s. Therefore, assuming zero slope can be a closed-form solution for calculating the ss matrix according to Eq. (7). Algorithm 2 shows the thin coding algorithm.

$$\beta\beta^T s - \beta P + \lambda Rs = 0$$
$$s = (\beta\beta^T + \lambda R)^{-1}\beta P \quad (7)$$

D. Dictionary Update Using Dictionary Learning Algorithm

After discovering each of the set's hidden word codes, the optimization problem (8) is solved to update the β dictionary. By locating the slope's root, Eq. (8)'s convex optimization problem can be successfully solved. Instead of calculating each β_{kn} separately to solve the dictionary learning problem, a general solution can be suggested to acquire the full β matrix by altering (8) to (9).

$$\min_\beta \sum_{d=1}^{D}\left(\left||w_d - diag(s_d^T\beta)\right||_2^2\right) = \sum_{d=1}^{D}\sum_{n=1}^{N}(w_{d,n} - s_{d,n.}^T\beta_{.n})^2 \quad (8)$$

$$\text{s.t. } \beta \geq 0, \sum_{n=1}^{N}\beta_{kn} = 1, \forall k$$

$$\min_\beta \sum_{d=1}^{D}\left(\left||w_d - diag(s_d^T\beta)\right||_2^2\right) = \sum_{d=1}^{D} trace(P^T P + s_d^T\beta\beta^T s_d - 2P^T s_d^T\beta) \quad (9)$$

which is $P = diag(w_{d,1}, \dots, w_{d,N})$. Setting the slope to zero yields the value of β per Eq. (10).

Algorithm 3: Dictionary learning algorithm

$P = diag(w_{d,1}, \dots, w_{d,N})$
$\beta = (\sum_{d=1}^{D} s_d\, s_d^T)^{-1}(\sum_{d=1}^{D} s_d\, P)$
for $k = 1 : K$
for $n = 1 : N$
$\beta_{kn} = \max(\beta_{kn}, 0)$
end
end
for $k = 1 : K$
$\beta_{k.} = \frac{\beta_{k.}}{\left||\beta_{k.}|\right|_1}$
End

$$\sum_{d=1}^{D}(S_d\, S_d^T\beta - s_d\, P) = 0$$
$$\beta = (\sum_{d=1}^{D} S_d\, S_d^T)^{-1}(\sum_{d=1}^{D} s_d\, P) \quad (10)$$

The proposed method is depicted in detail in Algorithm 3 of the dictionary learning process in Fig. 2.

E. Discussion

In this article, a new method for revealing traffic patterns in video surveillance using the topic model is presented. This method focuses on extracting and classifying traffic patterns from video footage, with specific application in intelligent transportation systems. The proposed method involves segmenting video clips, extracting optical flux features, and using a topic coding approach to train the system about traffic patterns. The results of applying this method in the video data set are promising to identify and depict different traffic patterns. The importance of the article begins by emphasizing the increasing use of surveillance cameras in various applications, including intelligent transportation systems. The growth of video data is a challenge for manual monitoring and necessitates the development of automatic pattern recognition methods. This research deals with a practical problem and the identification of traffic patterns is very important for traffic management and analysis. It briefly discusses the available approaches for analyzing traffic scenes. It mentions the challenges associated with tracking-based methods and the potential of optical flux data to create more complex models. The introduction provides valuable background to the proposed method and provides context for its importance. This paper also discusses the identification of eight key traffic patterns in the dataset and provides a list of these patterns. However, it would be useful to include more information on how these patterns are determined and how they relate to real-world traffic scenarios. In addition, the paper notes that two traffic patterns were indistinguishable due to camera positioning, which raises questions about the method's limitations and potential improvements. In summary, the paper introduces an attractive approach to traffic pattern detection in video surveillance, but could benefit from a more detailed description of the method, a deeper analysis of the results, and a discussion of its broader significance and potential limitations.

IV. Implementation of the Proposed Method and Results

A. Implementation Environment

To acquire the optical flux of the clips, the proposed method was implemented in the C++ programming environment using open_CV functions. The methods were implemented using the C++ linear algebra package Armadillo [27], and the software was run on a computer with an Intel Core i4790 7 processor and 8 GB of memory.

B. Total Data and their Characteristics

The video images used in this article are from the QMUL1 dataset, which is shown in Fig. 4 as an example.

The activity analysis and behavior comprehension applications of this traffic data collection are particularly beneficial [28]. The tough video images in this bank have a resolution of 288 x 360, a frame rate of 25 frames per second, and a total of one hour (90,000 frames). It should be noted that the topic model has swiftly adopted this image bank as a go-to resource. Fig. 5 displays the eleven typical traffic patterns from the QMUL dataset, and Table III lists each one's description. They are as follows: Turning left from the south side to the west is traffic pattern number one. Crossing the intersection from the east to the west is traffic pattern number two. Turning right from the north side to the west is traffic pattern number four. Turning left from the west to the north is traffic pattern number five. Turning right from the east side to the north is traffic pattern number six. Turning left from the north side to the east is traffic pattern number seven. It should be noted that the camera's angle and height off the ground have a significant impact on how accurately the optical flux algorithm detects motion vectors. Also, Fig. 6 shows examples of traffic patterns in this collection.

Fig. 4. QMUL dataset containing one hour (90,000 frames) of high-volume traffic video collected at the intersection.

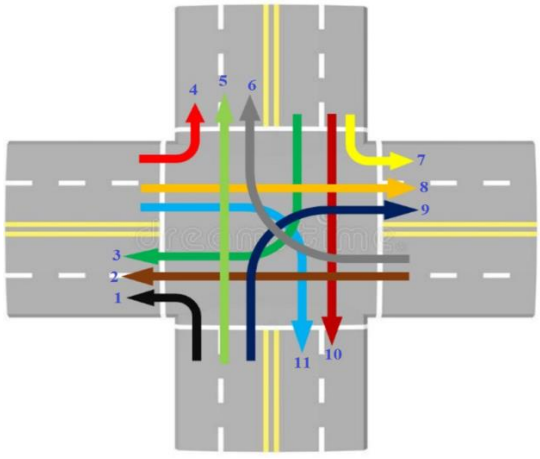


Fig. 5. Common traffic patterns in the QMUL data set (the description and numbering of the patterns are given in Table III).

TABLE III. TRAFFIC PATTERNS THAT ARE TYPICAL AND FREQUENT IN THE QUML DATASET

Traffic pattern number	name of the movement pattern	Recognizability
1	From the south side, turn left and head west.	□
2	moving from the east side of the intersection to the west side	□
3	From the north side, turn right to reach the west side.	□
4	From the west side, make a left turn to the north side.	□
5	moving from the south side of the crossroads to the north side	□
6	From the east side, make a right turn to the north side.	□
7	From the north side, make a left turn to the east side.	□
8	moving from the west to the east across the intersection	□
9	On the south side, turn right toward the east.	□
10	from the north side to the south side of the intersection	□
11	From the west side, turn right to go to the south side.	□

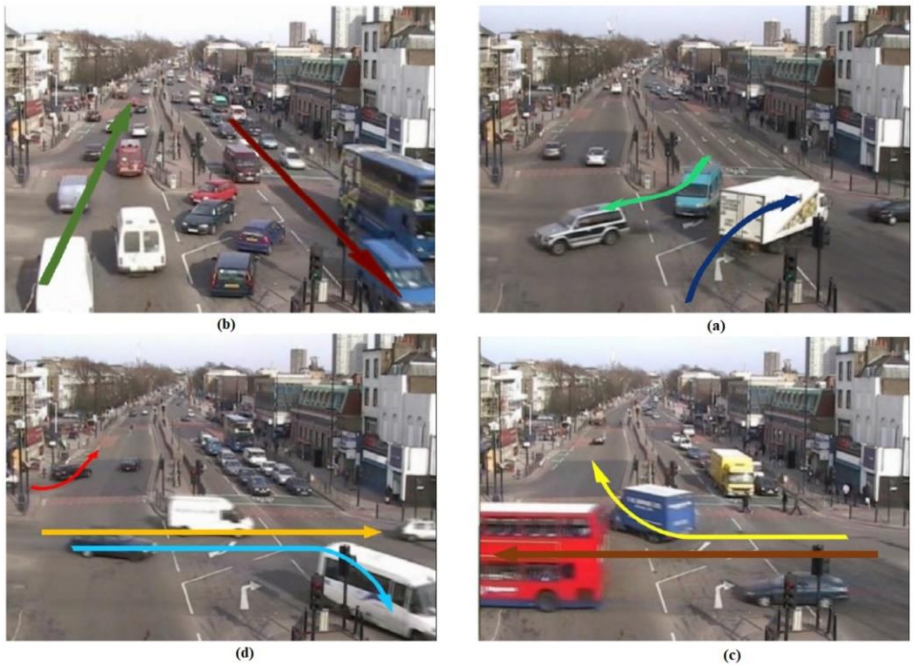


Fig. 6. Examples of traffic patterns in the QMUL image bank, (a) pattern 3 and 9, (b) pattern 5 and 10, (c) pattern 2 and 6 and (d) pattern 4, 8 and 11.

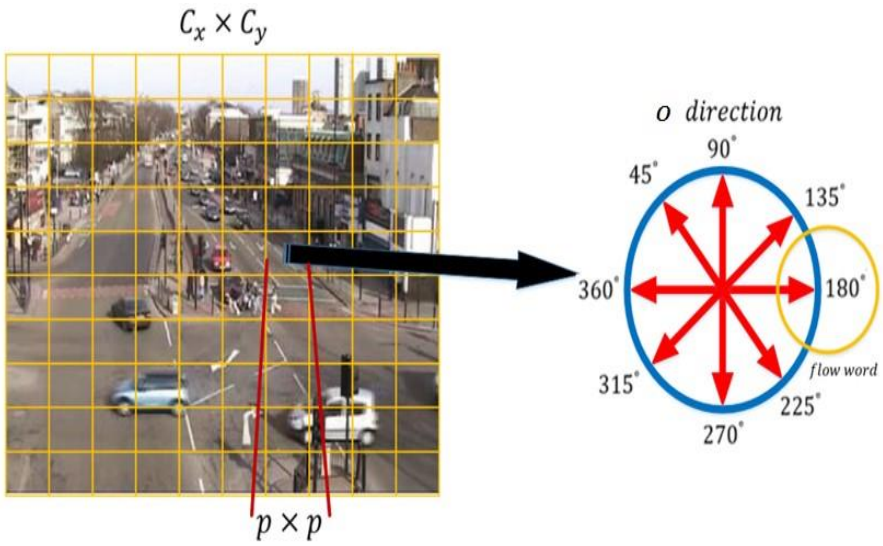


Fig. 7. How to make stream words.

C. Implementation Processes of the Proposed Method

In implementing the proposed method, the input video was first divided into three-second sub-clips. Then the optical flux characteristics were extracted for each pair of consecutive frames in each clip using the TVL-1 optical flux algorithm [29]. After that, as shown in Fig. 7, the output of the optical flux algorithm was divided into 10 x 10 cells. According to the resolution of each frame, there are 29 x 36 cells, and each of the motion vector values in each cell is one of 8 directions: 45º, 90º, 135º, 180º, 225º, 270º, 315º, and 360º. Then this process is repeated for each clip, and finally each clip is displayed with a vector length of $N = 8 \times 36 \times 29 = 8352$. Each member of the vector represents the number of occurrences of the current word. We experimentally set the λ parameter equal to 0.1 and the K value equal to 25. We considered the maximum number of repetitions of the program to be 20, and the value of ε for the convergence of the algorithms was 0.01. In this experiment, 200 clips were used for training.

D. Implementation Results

After running the proposed algorithm on the QMUL bank, traffic patterns were extracted. As can be seen in Fig. 8, three common movement behaviors—turning left [Fig. 8(a)], turning right [Fig. 8(b)], and crossing the intersection [Fig. 8(c)]—have been extracted. Also, in Fig. 8(d) to 8(i), two traffic patterns have been extracted simultaneously. As mentioned before, due to the importance of the location of the camera in the detection of traffic patterns, traffic patterns one and seven are undetectable in the QMUL bank and are not observed in the detected traffic patterns. Finally, according to the number and timing of red lights and traffic routines of cars in this data set, the proposed method has been able to correctly extract eight meaningful flows and patterns from the nine existing traffic patterns (accuracy 88.8%). The use of local movements, which is the movement of pixels between two frames, as features of "direction of flow detection" and "motion patterns", increases the possibility of misdiagnosis. For example, the presence of pedestrians or other moving objects increases the possibility of detecting meaningless patterns.

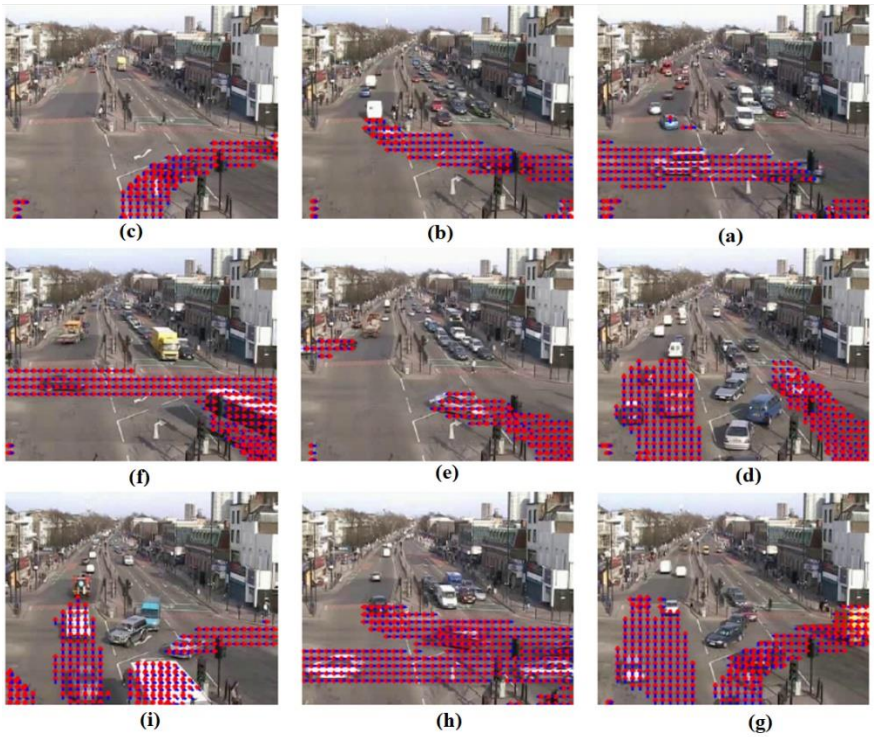


Fig. 8. Traffic flows and patterns obtained by the proposed method, (a) going left (b) turning north (c) a right turn from south to east, (d) crossing the intersection (e)) a right turn from west to south (f) going right from the west side to the south side (g) crossing the intersection from the south side to the north side (h) turning right from the east side to the north (i) turning to Right from south to east.

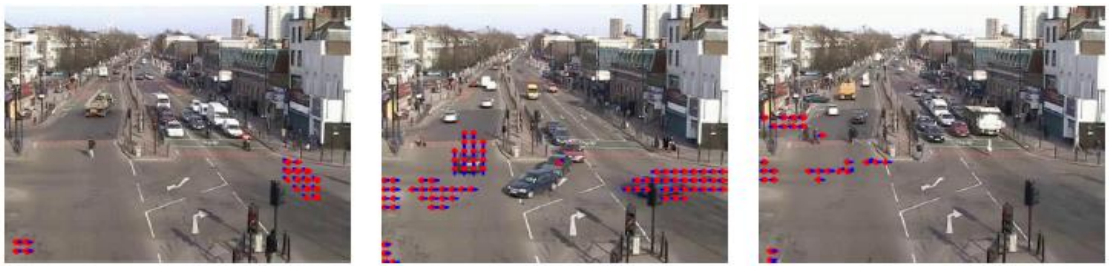


Fig. 9. Traffic patterns extracted from QMUL bank that do not have a specific meaning.

Naturally, scientists have employed these patterns to identify anomalies and peculiar movements [1]. There are other traffic patterns in the bank, as seen in Fig. 9 taken from QMUL, that have no clear significance. Therefore, it will be challenging to analyze the scenario, uncover the rules, and apply them using the patterns gleaned from the thematic model. Some traffic flows and patterns recovered have no particular significance because of the topic model's nature, which was primarily created and utilized for processing natural language and text and does not account for difficulties such as abrupt changes in illumination, image size, or camera viewing angle.

This paper has several significant contributions in the field of traffic pattern detection in video surveillance, which can be summarized as follows:

1) Unsupervised Traffic Pattern Recognition: The main contribution of this work is the development of an unsupervised method to identify and classify traffic patterns in video surveillance data. Traditional approaches often rely on predefined training datasets or tracking techniques, which may be limited by the need for extensive manual labeling or tracking accuracy issues. The proposed method takes a different approach by using a topic coding model to automatically extract traffic patterns without the need for a previous training dataset. This non-supervised nature of the method makes it adaptable to a wide range of traffic scenarios, including scenarios with diverse and evolving patterns.

2) Topic coding with optical flux features: This paper introduces the use of topic coding, a technique commonly used in natural language processing, in the context of video analysis. This method transforms visual traffic movement patterns into descriptive expressions that enable the identification and categorization of significant traffic patterns. To achieve this goal, the authors use optical flux features extracted from video frames that contain detailed information about local motion. This combination of thematic coding and optical flux features provides a new and promising approach for traffic pattern recognition. The contributions of this work extend to the development of an automated system to recognize common traffic behaviors at intersections, such as left turns, right turns, and crossing intersections. These contributions have significant relevance for applications in intelligent transportation systems, traffic management, and public safety, where accurate and automatic traffic pattern detection is essential for efficient monitoring and decision-making.

V. Conclusion

In urban traffic, crossroads play a significant role, and surveillance cameras are frequently utilized to regulate and control public spaces. The traditional method of video surveillance does not function well due to the vast volume of video data and the unreliability of humans; therefore, a system that automatically collects traffic flows and patterns is necessary. A non-supervisory strategy to identify traffic patterns in video surveillance was put forth in this paper. First, the optical flux algorithm is used to determine the traffic flow in each frame. The Car was taught using the theme model of these traffic patterns because behaviors like turning left, turning right, and traveling straight through an intersection are considered to be common significant traffic patterns in the images observed by the camera. According to the position of the camera in the QMUL bank, the implementation results revealed that the suggested method was able to accurately calculate eight significant motion patterns out of a total of nine conceivable patterns. Turning left, turning right, and passing straight are frequent and permitted in areas where traffic is controlled, such as junctions, according to traffic regulations and driving laws. Now, if there are movements that deviate from these authorized and typical patterns, they can first be identified as unexpected events before suitable choices, like recording a violation, can be taken into account. In other words, future studies may focus on the identification of violations in terms of atypical movement. A future study might examine how well the suggested algorithm holds up to environmental elements in the image, such as changes in light intensity, camera position, and difficulties similar to those seen in image processing. Some limitations and potential challenges in this study are as follows:

Reliability of tracking-based methods, one class of which involves first tracking objects (e.g. cars or people) and then using the tracked trajectories for further analysis: This indicates that these methods may suffer from challenges related to the accuracy and reliability of object detection and tracking.

Inadequate anomaly detection, where some existing methods may not effectively detect anomalies or unusual traffic behavior: This shows that the proposed approach can potentially detect unexpected events or violations by identifying patterns that deviate from the permitted and normal traffic behaviors. This indicates that existing methods may have limitations in detecting and classifying anomalies in traffic patterns.

Environmental variability, where the location and angle of surveillance cameras can significantly affect the accuracy of optical flux-based motion vector detection: This suggests that existing methods may be limited by environmental factors, such as changes in lighting conditions, changes in camera positions, and other challenges commonly encountered in video surveillance.

Acknowledgment

This research was supported by Key Technology Research and Development Program of Henan Province China (Grant No.232102210194).

References

[1] H. Zou, K. Cao, and C. Jiang, “Spatio-temporal visual analysis for urban traffic characters based on video surveillance camera data,” ISPRS Int J Geoinf, vol. 10, no. 3, p. 177, 2021.

[2] L. Song, F. Jiang, Z. Shi, R. Molina, and A. K. Katsaggelos, “Toward dynamic scene understanding by hierarchical motion pattern mining,” IEEE Transactions on Intelligent Transportation Systems, vol. 15, no. 3, pp. 1273–1285, 2014.

[3] I. Tvoroshenko and M. Dziubenko, “Modern methods of analysis of the movement scheme using video detection of vehicles,” 2020.

[4] Y. Cheng, X. Ji, T. Lu, and W. Xu, "On detecting hidden wireless cameras: A traffic pattern-based approach," IEEE Trans Mob Comput, vol. 19, no. 4, pp. 907–921, 2019.

[5] J. Varadarajan, R. Emonet, and J.-M. Odobez, "A sequential topic model for mining recurrent activities from long term video logs," Int J Comput Vis, vol. 103, no. 1, pp. 100–126, 2013.

[6] M. C. R. Murça, R. J. Hansman, L. Li, and P. Ren, "Flight trajectory data analytics for characterization of air traffic flows: A comparative analysis of terminal area operations between New York, Hong Kong and Sao Paulo," Transp Res Part C Emerg Technol, vol. 97, pp. 324–347, 2018.

[7] N. Apthorpe, D. Reisman, and N. Feamster, "A smart home is no castle: Privacy vulnerabilities of encrypted iot traffic," arXiv preprint arXiv:1705.06805, 2017.

[8] X. Ji, Y. Cheng, W. Xu, and X. Zhou, "User presence inference via encrypted traffic of wireless camera in smart homes," Security and Communication Networks, vol. 2018, 2018.

[9] F. Duarte and C. Ratti, "What urban cameras reveal about the city: The work of the Senseable City Lab," Urban Informatics, pp. 491–502, 2021.

[10] N. Apthorpe, D. Reisman, S. Sundaresan, A. Narayanan, and N. Feamster, "Spying on the smart home: Privacy attacks and defenses on encrypted iot traffic," arXiv preprint arXiv:1708.05044, 2017.

[11] Y. Cheng, X. Ji, T. Lu, and W. Xu, "Dewicam: Detecting hidden wireless cameras via smartphones," in Proceedings of the 2018 on Asia Conference on Computer and Communications Security, 2018, pp. 1–13.

[12] J. Barthélemy, N. Verstaevel, H. Forehead, and P. Perez, "Edge-computing video analytics for real-time traffic monitoring in a smart city," Sensors, vol. 19, no. 9, p. 2048, 2019.

[13] M. C. R. Murca and R. J. Hansman, "Identification, characterization, and prediction of traffic flow patterns in multi-airport systems," IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 5, pp. 1683–1696, 2018.

[14] Khezri, E., Zeinali, E., & Sargolzaey, H. (2023). SGHRP: Secure Greedy Highway Routing Protocol with authentication and increased privacy in vehicular ad hoc networks. Plos one, 18(4), e0282031.

[15] H. Luo, J. Liu, W. Fang, P. E. D. Love, Q. Yu, and Z. Lu, "Real-time smart video surveillance to manage safety: A case study of a transport mega-project," Advanced Engineering Informatics, vol. 45, p. 101100, 2020.

[16] Khezri, E., & Zeinali, E. (2021). A review on highway routing protocols in vehicular ad hoc networks. SN Computer Science, 2, 1-22.

[17] Zhang, H., Zou, Q., Ju, Y., Song, C., & Chen, D. (2022). Distance-based support vector machine to predict DNA N6-methyladenine modification. Current Bioinformatics, 17(5), 473-482.

[18] Cao, C., Wang, J., Kwok, D., Cui, F., Zhang, Z., Zhao, D., ... & Zou, Q. (2022). webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. Nucleic acids research, 50(D1), D1123-D1130.

[19] Zhao, H., Wang, H., Xu, N., Zhao, X., & Sharaf, S. (2023). Fuzzy approximation-based optimal consensus control for nonlinear multiagent systems via adaptive dynamic programming. Neurocomputing, 553, 126529.

[20] M. Samiei, A. Hassani, S. Sarspy, I. E. Komari, M. Trik, and F. Hassanpour, "Classification of skin cancer stages using a AHP fuzzy technique within the context of big data healthcare," J Cancer Res Clin Oncol, pp. 1–15, 2023.

[21] J. Sun, Y. Zhang, and M. Trik, "PBPHS: a profile-based predictive handover strategy for 5G networks," Cybern Syst, pp. 1–22, 2022.

[22] M. Trik, H. Akhavan, A. M. Bidgoli, A. M. N. G. Molk, H. Vashani, and S. P. Mozaffari, "A new adaptive selection strategy for reducing latency in networks on chip," Integration, vol. 89, pp. 9–24, 2023.

[23] Trik, M., Pour Mozafari, S., & Bidgoli, A. M. (2021). An adaptive routing strategy to reduce energy consumption in network on chip. Journal of Advances in Computer Research, 12(3), 13-26.

[24] Yue, S., Niu, B., Wang, H., Zhang, L., & Ahmad, A. M. (2023). Hierarchical sliding mode-based adaptive fuzzy control for uncertain switched under-actuated nonlinear systems with input saturation and dead-zone. Robotic Intelligence and Automation, 43(5), 523-536.

[25] Wang, T., Zhang, L., Xu, N., & Alharbi, K. H. (2023). Adaptive critic learning for approximate optimal event-triggered tracking control of nonlinear systems with prescribed performances. International Journal of Control, 1-15.

[26] M. Trik, A. M. N. G. Molk, F. Ghasemi, and P. Pouryeganeh, "A hybrid selection strategy based on traffic analysis for improving performance in networks on chip," J Sens, vol. 2022, 2022.

[27] Khezri, E., Zeinali, E., & Sargolzaey, H. (2022). A novel highway routing protocol in vehicular ad hoc networks using VMaSC-LTE and DBA-MAC protocols. Wireless Communications and Mobile Computing, 2022.

[28] Cao, Y., Xu, N., Wang, H., Zhao, X., & Ahmad, A. M. (2023). Neural networks-based adaptive tracking control for full-state constrained switched nonlinear systems with periodic disturbances and actuator saturation. International Journal of Systems Science, 54(14), 2689-2704.

[29] C. Yang, K. Clarke, S. Shekhar, and C. V. Tao, "Big Spatiotemporal Data Analytics: A research and innovation frontier," International Journal of Geographical Information Science, vol. 34, no. 6. Taylor & Francis, pp. 1075–1088, 2020.

[30] Wang, Z., Jin, Z., Yang, Z., Zhao, W., & Trik, M. (2023). Increasing efficiency for routing in internet of things using Binary Gray Wolf Optimization and fuzzy logic. Journal of King Saud University-Computer and Information Sciences, 35(9), 101732.

[31] Li, X., Gui, J., & Liu, J. (2023). Data-driven traffic congestion patterns analysis: A case of Beijing. Journal of Ambient Intelligence and Humanized Computing, 14(7), 9035-9048.

[32] Pascale, A., Macedo, E., Guarnaccia, C., & Coelho, M. C. (2023). Smart mobility procedure for road traffic noise dynamic estimation by video analysis. Applied Acoustics, 208, 109381.

Improving Deep Reinforcement Learning Training Convergence using Fuzzy Logic for Autonomous Mobile Robot Navigation

Abdurrahman bin Kamarulariffin, Azhar bin Mohd Ibrahim*, Alala Bahamid
Department of Mechatronics Engineering, International Islamic University Malaysia, Kuala Lumpur, Malaysia

***Abstract*—Autonomous robotic navigation has become hotspot research, particularly in complex environments, where inefficient exploration can lead to inefficient navigation. Previous approaches often had a wide range of assumptions and prior knowledge. Adaptations of machine learning (ML) approaches, especially deep learning, play a vital role in the applications of navigation, detection, and prediction about robotic analysis. Further development is needed due to the fast growth of urban megacities. The main problem of training convergence time in deep reinforcement learning (DRL) for mobile robot navigation refers to the amount of time it takes for the agent to learn an optimal policy through trial and error and is caused by the need to collect a large amount of data and computational demands of training deep neural networks. Meanwhile, the assumption of reward in DRL for navigation is problematic as it can be difficult or impossible to define a clear reward function in real-world scenarios, making it challenging to train the agent to navigate effectively. This paper proposes a neuro-symbolic approach that combine the strengths of deep reinforcement learning and fuzzy logic to address the challenges of deep reinforcement learning for mobile robot navigation in terms of training time and the assumption of reward by incorporating symbolic representations to guide the learning process, and inferring the underlying objectives of the task which is expected to reduce the training convergence time.**

Keywords*—Autonomous navigation; deep reinforcement learning; mobile robots; neuro-symbolic; Fuzzy Logic

I. INTRODUCTION

Advancements in robot navigation have spurred the development of algorithms that leverage basic rules and environmental mapping to optimize path planning. Rule-based methods, such as Fuzzy logic and Neuro-fuzzy techniques, have been extensively explored to enhance navigation decisions and tracking performance under uncertain conditions [1], [2]. While these methods offer valuable insights, they often require extensive justification and may not fully meet the demands for efficient and accurate path planning.

To address this challenge, researchers have turned to bio-inspired approaches, such as genetic algorithms and swarm optimization, which draw inspiration from biological behavior and incorporate prior knowledge to simulate human cognitive processes [3], [4]. One particularly promising area in navigation research is reinforcement learning (RL), which enables autonomous agents to learn and make sequential decisions in complex environments. Machine learning models, including supervised, unsupervised, and reinforcement learning, have played a pivotal role in robotics research, enabling learning, adaptation, and effective detection and classification. Deep reinforcement learning (DRL), a fusion of RL and deep neural networks, has emerged as a powerful approach for decision-making tasks involving high-dimensional inputs [5], [6].This article aims to delve into the application of RL techniques, specifically Q-learning and deep Q-networks, for mobile robot path planning. By seamlessly integrating these techniques with widely used frameworks such as ROS, Gazebo, and OpenAI, a robust and autonomous navigation system can be developed, leading to improved performance, optimized routes, and efficient obstacle avoidance in complex environments. The evaluation of this system will undoubtedly contribute to the advancement of autonomous robotics. The trial-and-error learning process inherent in RL offers immense potential for building human-level agents and has been extensively explored in various domains [7] [8]. Deep learning (DL), characterized by its ability to extract meaningful patterns and classifications from raw sensory data through deep neural networks, has revolutionized the field of machine learning. When combined with RL, in the form of DRL, this integration has shown remarkable success in tackling challenges associated with sequential decision-making [9], [10]. Notably, DRL excels in scenarios involving a vast number of states, making it an ideal candidate for addressing navigation complexities. Nevertheless, achieving optimal navigation remains an ongoing challenge, necessitating further optimization and effective handling of high-dimensional data. Reinforcement learning methods offer valuable approaches for learning and planning navigation, empowering agents to interact with their environment and make autonomous decisions. Various studies have proposed agent-based DRL approaches for navigation, successfully simulating diverse scenarios without the need for intricate rule-based systems or laborious parameter tuning. However, there is still room for improvement in terms of achieving the shortest and fastest routes. To enhance navigation performance and optimize evacuation paths, researchers have explored techniques such as look-ahead crowded estimation and Q-learning, which have demonstrated superior results compared to other RL algorithms [6]. Additionally, CNN-based robot-assisted evacuation systems have been developed to maximize pedestrian outflow by extracting specific features from high-dimensional images. Furthermore, iterative, and incremental learning strategies,

like vector quantization with Q learning (VQQL), have been proposed to expedite the learning process and optimize navigation by gradually improving interactions among agents [11], [12]. These advancements in DRL continue to show great promise in addressing the speed of agent learning and optimizing navigation processes. In the realm of task planning, the ability to find a series of steps that transform initial conditions into desired states is crucial. Task planning becomes especially important when atomic actions alone cannot accomplish a task. Neuro-symbolic task planning has emerged as an effective approach, allowing for the incorporation of restrictions, guidelines, and requirements in each activity. However, traditional task planners often rely on detailed hand-coded explanations, limiting their scalability. To overcome this limitation, a combination of deep learning and symbolic planning, known as a neuro-symbolic approach, has shown potential by leveraging visual information instead of hand-coded explanations [3], [13], [14]. However, collecting image data for neuro-symbolic models in robotic applications is a labor-intensive process that involves steps such as creating problem instances, defining initial and goal states, operating robots, and capturing scene images. The challenges associated with data collection have hindered the widespread adoption of neuro-symbolic models in robot task planning. Neuro-symbolic models excel in reasoning, providing explanations and manipulating complex data structures. Conversely, numerical models, such as neuronal models, are preferred for pattern recognition due to their generalization and learning abilities. A unified strategy proposes that the characteristic properties of symbolic artificial intelligence can emerge from distributed local computations performed by neuronal models, spanning cognitive functions from the neuron level to the structural level of the nervous system. By integrating neuro-symbolic and numerical models, a comprehensive framework can be established to leverage the strengths of both approaches in robotics. This integrated approach holds the potential to enable efficient task planning, grounding symbols in perceptual information, and enhancing pattern recognition capabilities. Ultimately, this integration could advance cognitive functions and pave the way for the creation of more sophisticated robotic systems.

This paper is organized as follows. Section II presents the proposed method which integrates the reinforcement learning (RL) and fuzzy logic for mobile robot path planning, aiming to create a robust autonomous navigation system that optimizes routes and efficiently avoids obstacles in complex environments. Section III illustrates the simulation set-up, while Section IV provides an evaluation of the training process of the policy optimization. Finally, Section V presents the evaluation and verification of the developed policy based on the proposed method, followed by the conclusion.

II. METHODS

The methodology for this project involves the utilization of simulation tools, namely Gazebo, ROS (Robot Operating System), and OpenAI Gym. Gazebo provides a realistic environment for simulating the mobile robot path planning system, while ROS serves as a comprehensive framework for controlling the robot and interfacing with its sensors and actuators. OpenAI Gym is used to train and evaluate the reinforcement learning algorithms. The main focus of this project is to apply reinforcement learning techniques to mobile robot path planning. Unlike traditional approaches that rely on SLAM or mapping techniques, the project aims to enable the robot to learn the optimal path through a reward and punishment system. By using reinforcement learning algorithms such as Q-learning, SARSA, and DQN, the robot can learn to navigate its environment efficiently and safely. To facilitate communication between the simulation and the robot, ROS integration is implemented. This integration allows the robot to receive sensor data, send control commands, and interact with the simulation environment seamlessly. By leveraging the capabilities of ROS, the reinforcement learning algorithms can effectively interface with the robot's actions and observations [15]–[17].The reinforcement learning algorithms receive feedback through a reward and punishment system based on the robot's performance in reaching the goal while avoiding collisions and obstacles. The training aims to optimize the robot's decision-making and path planning abilities. Performance analysis is conducted to assess the effectiveness of the trained reinforcement learning models. Metrics such as the time taken to reach the goal, collision occurrences with static and dynamic obstacles, and the number of pathing alterations are measured and analyzed. These metrics provide insights into the path planning efficiency, collision avoidance capabilities, and adaptability of the reinforcement learning approach. In conclusion, the methodology of this project involves using simulation tools (Gazebo, ROS, and OpenAI Gym) to evaluate the application of reinforcement learning algorithms (Q-learning, SARSA, and DQN) in mobile robot path planning. The integration of ROS ensures seamless communication between the simulation environment and the robot, while the OpenAI Gym environment provides a standardized framework for training and evaluating the algorithms. The methodology enables rigorous testing and analysis of the robot's performance in terms of path planning, collision avoidance, and adaptability to dynamic environments. This following subsection discusses the mathematical model of Q-learning with fuzzy logic approach theory towards navigation problems, and experimentation setup that is used in this work.

In the context of agents utilizing visual SLAM, traditional algorithms are still employed for final path planning on the map. However, RL offers numerous applications, and in mobile robot navigation, it can replace the path planning part. The RL model, after training, can effectively make decisions, enabling the agent to select its path from one location to another based on interactions with the environment [18], [19]. The environment is abstracted into a grid map representation, with each position on the map corresponding to an agent's state. Transitioning from one state to another reflects the actual movement of the entity, while the agent's behavioral decision-making is represented by its state choice at each step in the RL model. The reward value plays a pivotal role in guiding path selection. Early Q-learning recorded reward values between position states in a table, guiding the next state selection. As depth-enhanced learning emerges, the DL model is integrated, replacing the table with a neural network, which provides corresponding decision results by inputting the state [20], [21]. The weighting parameters in the neural network

influence the choice of the next state. On the other hand, when incorporating fuzzy logic into the RL model, the decision-making process becomes more nuanced and interpretable. Fuzzy logic allows for handling uncertainties and imprecise information, enabling the agent to reason with vague input and output values. By combining RL and fuzzy logic, the agent can make more human-like decisions, considering both the environment's precise measurements and the agent's subjective understanding of the surroundings. This fusion can enhance path planning in complex and dynamic environments by considering various factors and optimizing the decision-making process.

A. Q-Learning Algorithm

RL defines any decision maker as an agent and everything outside the agent as the environment. The agent aims to maximize the accumulated reward and obtains a reward value as a feedback signal for training through interaction with the environment. Beyond the agent (who perform actions) and the environment (which made of states), there are three major elements of a reinforcement learning system:

- Policy $\boldsymbol{\pi}$: It is to formalize an agent's decision and determine the agent's behaviour at a given time. A policy π is a function that maps between the perceived state and the action is taken from that state.
- Reward $\boldsymbol{r}$: The agent receives feedback known as rewards, r_{t+1} for each action at time step t, indicating the inherent desirability of that state. The main goal of the agent is to maximize the cumulative reward over time. The total sum of the rewards (return) is:

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \ldots r_T,\ T\text{: final time step}$$

The agent-environment interaction breaks into episodes where each episode ends in a state called the terminal state, followed by a reset to a standard starting state. In some cases, the episodes continue where final time step would be $T = \infty$, and the return become infinite. So, a discount factor γ is introduced. The discounted return is defined as:

$$R_t = r_{t+1} + \gamma\, r_{t+2} + \gamma^2\, r_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

$$0 < \gamma < 1$$

Rewards can be sparse (after a long sequence of actions), every time step, or at the end of the episodes.

- Value function: Most of the RL algorithms are based on estimating value functions (states or state action). Value function is used to estimate how good a certain state is for the agent to be in (state value function), or how good a certain action is to perform in a specific state (state-action value function). The state value functions under the policy π, denoted $V\pi(s)$, is the expected return,

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+1+k} | s_t = s\right\}$$

The state-action value function under policy π, denoted $Q^{\pi}(s, a)$, as the expected accumulated return from state s and action a. Q^{π} is also known as action value function or Q-Learning algorithm.

$$Q^{\pi}(s,a) = E_{\pi}\{ R_t | s_t = s, a_t = a\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+1+k} | s_t = s, a_t = a\}$$

$$Q^{\pi}(s,a) = E\,[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots | s, a\,]$$

Reinforcement learning is about finding an optimal policy that achieves a lot of reward over the long-term. A policy π is defined to be better than or equal to a policy π' if its expected return is greater than or equal to that of π' for all states.

$$\pi \geq \pi' \text{ if and only if } v^{\pi}(s) \geq V^{\pi'}(s), \text{for all states}$$

Optimal Value Functions must satisfy the below conditions:

$$V^*(s) = max\ V.(s), \text{ for all states}$$

$$Q^*(s, a) = maxQ.(s, a), \text{ for all states and actions}$$

We get the optimal policy by solving $Q*(s, a)$ to find the action that gives the most optimal state-action value function,

$$\pi^*(s) = arg \max_a Q^*(s,a)$$

Q-Learning algorithm is an off-policy value-based RL algorithm and very effective under unknown environment [6], [21], [22]. The value of a state-action can be decomposed into immediate reward plus the value of successor state-action $Q^{\pi}(s', a')$ with a discount factor (γ).

$$Q^{\pi}(s,a) = E_{s'a'}[r + \gamma Q^{\pi}(s', a') | s, a]$$

And according to the Bellman optimality, the optimal value function can be expressed as:

$$Q^*(s,a) = E_{s'a'}[r + \gamma \max_a Q^*(s', a') | s, a]$$

Update the value function iteratively to obtain optimal value function,

$$Q(s,a) \leftarrow Q(s,a) + \alpha.\,[r + \gamma \max a'\ Q(s', a') - Q(s,a)],$$

α: learning rate

$Q(s, a)$ converges to $Q^*(s, a)$ as $t \rightarrow \infty$.

Algorithm 1 illustrates the overall framework of the proposed Q-learning to generate the shortest route for navigation mapping.

Algorithm 1. Overall framework of the Q-Learning

```
Initialize Q (s, a) arbitrarily
repeat for each episode.
Initialize s.
  for each step of the episode do
    Choose a from s using € greedy policy.
    Do action a and observer r and s'
    Q(s, a) ← Q(s, a) + α. [r + γ max a'  Q (s', a') - Q (s, a)]
    s ← s'
  until s is terminal
```

B. Fuzzified Reward Function

The fuzzy logic-based approach has been adopted to enhance the decision-making [23], [24] process of the autonomous agent navigating through a maze. This work incorporated symbolic representation by adopting fuzzy logic into the reward function to guide the learning process and address the challenge of the computational demands of training. The proposed fuzzy reward function has three input variables and one output. The input variables are distance to obstacle (near, medium, far), distance to target (near, medium, far), and visual range (off target, medium, on target). The output variable is the reward points (see Fig. 4).

By employing three membership levels for each input variable (See Fig. 1, 2 and 3), a comprehensive set of 27 fuzzy rules has been devised (see Table I), effectively covering all possible combinations of the environment states. These rules dictate the agent's rewards, which are categorized as punishment (least), medium, and reward (most). By leveraging the flexibility and adaptability of fuzzy logic, the agent is guided through its learning process with a more nuanced and context-aware reward system, allowing it to make more informed decisions in a variety of maze scenarios and significantly improving its learning efficiency.

Table I presents a comprehensive and systematic overview of the fuzzy logic rules governing the agent's decision-making process in the maze navigation task. The table showcases the various combinations of input possibilities, encompassing distance with obstacles, distance with target location, and visual range, each categorized into appropriate linguistic variables (e.g., near, medium, far; off target, medium, on target). For every unique combination, the corresponding fuzzy logic "If/Then" rules are defined, determining the agent's rewards as punish, medium, or reward. Table I highlights the agent's adaptability and versatility through the vast array of rules, capturing the intricacies of different maze scenarios. With 27 distinct rules, the fuzzy logic system can precisely respond to the agent's real-time observations, guiding it towards optimal actions that lead to successful navigation. This rich and nuanced reward system empowers the agent to effectively learn from its experiences, enabling it to avoid obstacles, approach the target, and dynamically adjust its behavior based on varying visual cues. Consequently, the "If/Then Analysis Fuzzy Logic Rules Possibilities" table serves as a powerful tool in understanding and implementing the complex decision-making process of the agent, fostering efficient learning and successful maze navigation.

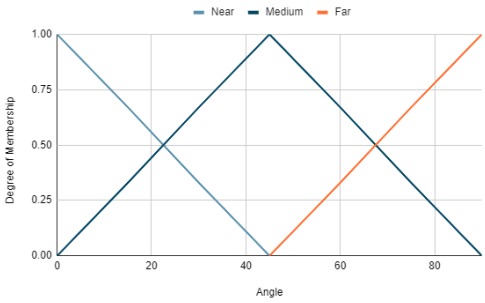


Fig. 1. Visual range.

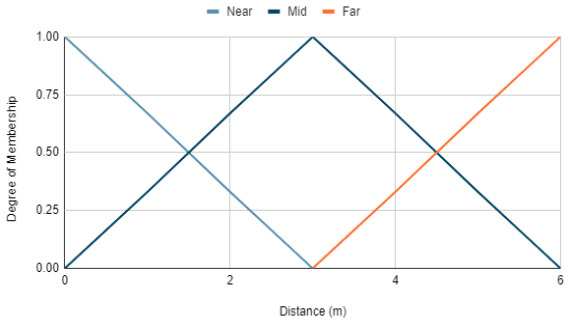


Fig. 2. Distance to target.

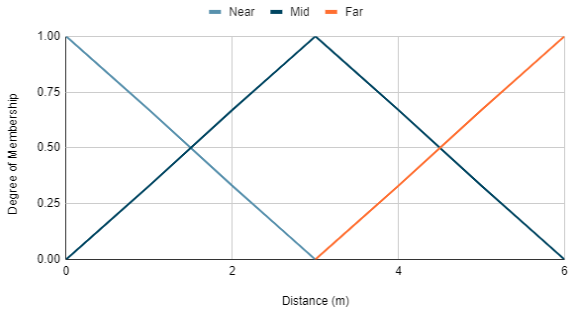


Fig. 3. Distance to obstacle.

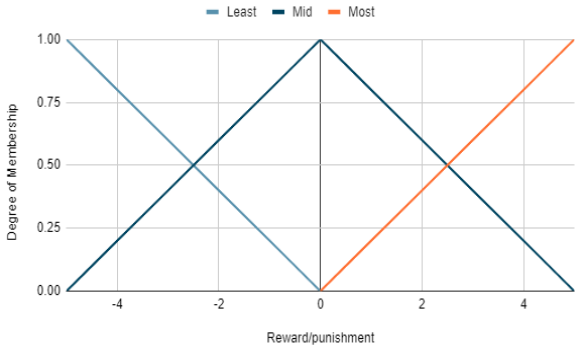


Fig. 4. Output fuzzy: reward points.

TABLE I. IF/THEN FUZZY LOGIC RULES FOR REWARDS

Distance	Visual Range	Obstacle (Near)	Obstacle (Medium)	Obstacle (Far)
Target (Near)	Near	High Positive	High Positive	High Positive
	Medium	Mid Positive	Mid Positive	Mid Positive
	Far	Low Positive	Low Positive	Low Positive
Target (Medium)	Near	High Middle	High Middle	High Middle
	Medium	Low Middle	Middle	Low Middle
	Far	Low Middle	Low Middle	Low Middle
Target (Far)	Near	Low Negative	Low Negative	Low Negative
	Medium	Mid Negative	Mid Negative	Mid Negative
	Far	High Negative	High Negative	High Negative

III. SIMULATION SETUP

To thoroughly evaluate the enhanced performance of the proposed method, this study embarked on constructing a detailed 3-D raster map model and crafting two distinct simulation maps, meticulously illustrated in Fig. 5(a) and Fig. 5(b). In this simulated environment, dynamic obstacles, symbolized by white cylinders, were strategically placed, posing challenges to a TurtleBot simulation machine car, visually presented in black. The machine car's laser range, portrayed in blue, scanned the surroundings as it navigated through the intricate maze. The red square pinpointed a target training point, emphasizing the complexity of the assigned tasks.

Fig. 5(a) specifically delves into a scenario where the TurtleBot is tasked with locating a singular target location represented by the red square amid a set of four cylindrical obstacles. This intricate setting simulates real-world challenges where the robot must efficiently identify and navigate towards a specific point among various hindrances.

Fig. 5(b) presents a more intricate scenario where the TurtleBot is assigned the mission of identifying two specific cylinders as target locations within a maze of block obstacles. This heightened complexity mirrors scenarios where the robot must discern and navigate through a maze-like environment to pinpoint multiple objectives. This detailed simulation environment allows for a comprehensive assessment of the proposed method's effectiveness in handling diverse and intricate navigation tasks.

The computer used for the simulations was equipped with a 4-core Intel i5 7400 CPU running at 3.00 GHz, 8 GB of RAM, running on the Ubuntu 16.04 operating system, and utilizing the ROS kinetic system. The article leveraged certain parameters for the 3-D environment model, which were sourced from the ROS open-source community. The corresponding parameter settings are as follows:

$$r_{goal} = 100,\ r_{obstacle} = -100,\ \varepsilon = -100,\ \sigma_x = \sigma_y = 1$$

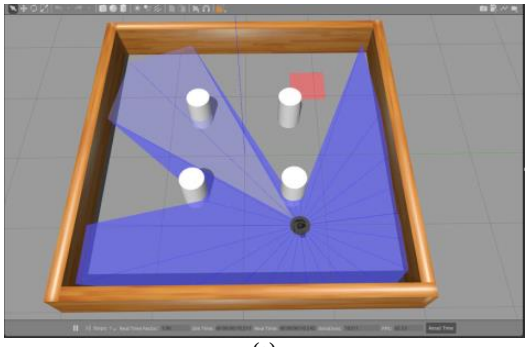
$$\gamma_{goal} = 0.9,\ r_{obstacle} = 0.9,\ r_{critical} = 0.8,\ r_{otherwise} = 0.75$$

In the context of this work:

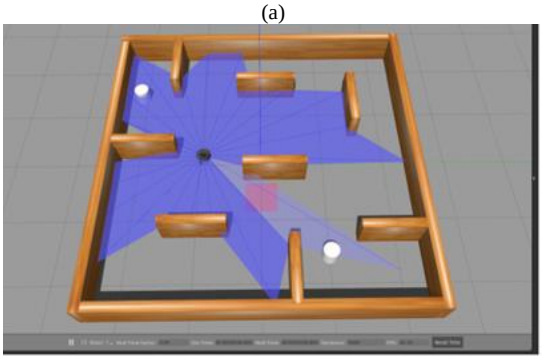
- " *r* " signifies a single reward.
- σ_x and σ_y represent the obstacle center coordinates.
- γ (gamma) serves as the discount factor, influencing the importance of future rewards.

In this context, *r* represents a reward, with r_{goal} and $r_{obstacle}$ being specific awards assigned to reaching the goal and encountering obstacles, respectively. The term γ serves as the discount factor, influencing the importance of future rewards in the context of reinforcement learning. The parameters ε, σ_x, and *and* σ_y represent the grid center coordinates, contributing to the spatial representation of the environment and the localization of obstacles.

Fig. 6 depicts The Turtlebot2 which is a popular mobile robot platform widely used in robotics research and applications.



(a)



(b)

Fig. 5. (a) and (b) Showcase maze circuits in the Gazebo simulation environment to test and evaluate the robot's path planning capabilities.

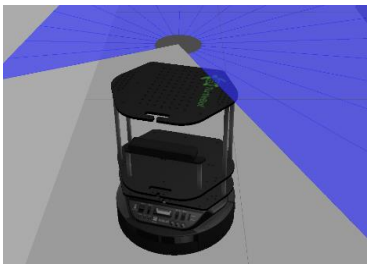


Fig. 6. The Turtlebot2 robot equipped with a lidar sensor.

IV. TRAINING PERFORMANCE OF THE PROPOSED METHOD

The visual representations in Fig. 5(a) and Fig. 5(b) aimed to illustrate the improved method's performance across different simulation maps. While these figures may not directly demonstrate capabilities, they serve as visual aids to showcase the distinct scenarios and complexities encountered by the agent in each environment. It's essential to acknowledge that the term "validation" in the context of comparing results from Reinforcement Learning (RL) and Fuzzy Logic (FL) approaches refers to a qualitative assessment rather than a formal validation process.

In Fig. 5(a), the experiment showcased the performance of the improved method on the first simulation map, providing insights into how the agent navigates a specific environment. On the other hand, Fig. 5(b) illustrated the capabilities of the improved method on the second simulation map, highlighting its adaptability to different scenarios. By comparing the results from RL and FL approaches, the article qualitatively validated the effectiveness of the enhanced technique in the complex 3-D environment. These visual representations offered a valuable qualitative assessment, helping to understand the

nuanced behaviors of the agent, its path planning strategies, and obstacle avoidance mechanisms in diverse settings. The improved method consistently demonstrated superior performance, efficiently finding optimal routes to reach the target point while navigating around obstacles effectively. The simulations offered valuable insights into the agent's behavior, path planning, and obstacle avoidance, elucidating fundamental aspects of autonomous robot navigation. The superior performance of the enhanced method was evident in its ability to navigate efficiently, choosing optimal routes while circumventing obstacles effectively.

Furthermore, the deliberate choice of Fig. 5(b) as a test run was made to rigorously assess the proposed method's robustness in scenarios with increased complexity and multiple target points. This strategic selection adds an additional layer of validation, demonstrating the algorithm's efficacy in handling intricate navigation tasks.

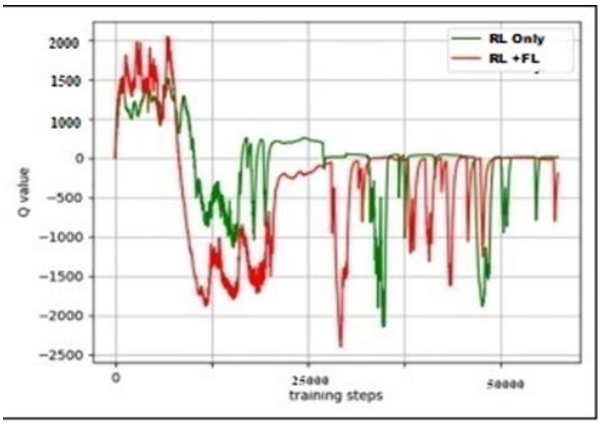


Fig. 7. Q-value comparison.

As shown in Fig. 7, in the Q-value map, the Fuzzy Logic (FL) example has a faster convergence speed, especially in 50 K training sessions, and after approximating 25 K trainings, the Q value of the FL algorithm is still richly transformed, showing the FL is less likely to fall into local optimum. This segment of our analysis offers a glimpse into the noteworthy performance attributes of the FL local search approach. As depicted in Fig. 8, which illustrates the bonus map, the FL example stands out due to its utilization of a multiple reward mechanism and a loop memory network. This distinction is most evident in the greater reward values attributed to the FL path, which correspondingly signify a reduced occurrence of repeated errors. In essence, a higher reward value in this context indicates a superior capacity to identify and follow an optimal path with fewer deviations. Turning our attention to Fig. 9, we delve into the loss diagram. Here, we observe a compelling trend: the loss associated with the FL example is consistently lower compared to that of the RL example. This finding is particularly significant, as it underscores the model's proficiency in minimizing error during the learning process. A lower loss value reflects a more accurate prediction and action selection by the model, emphasizing the effectiveness of the FL approach in optimizing path planning. To provide a closer examination of this phenomenon, Fig. 10 offers a magnified view of the loss diagram from Fig. 9. This detailed perspective reaffirms the rationality and effectiveness of the loss function employed in our model. The stability in parameter learning, particularly evident in the FL example, facilitates faster convergence to the optimal values. This not only enhances the efficiency of path planning but also showcases the model's robustness in navigating complex environments.

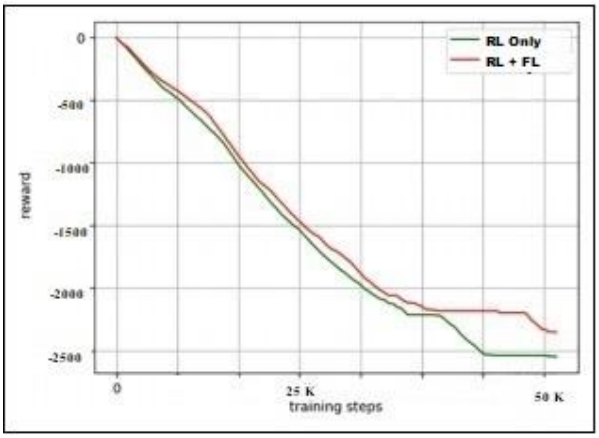


Fig. 8. Cumulative Rewards based on the proposed method vs. pure RL algorithm.

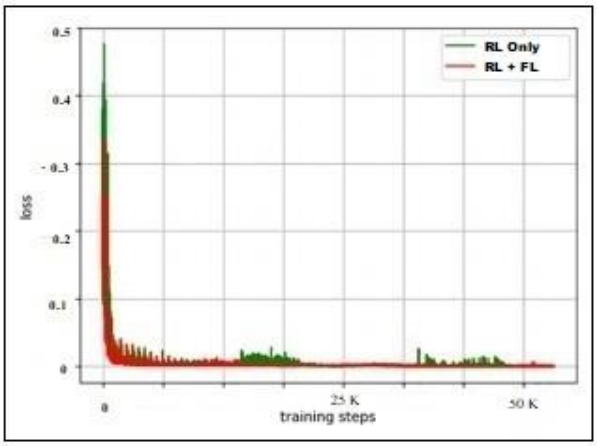


Fig. 9. Loss comparison.

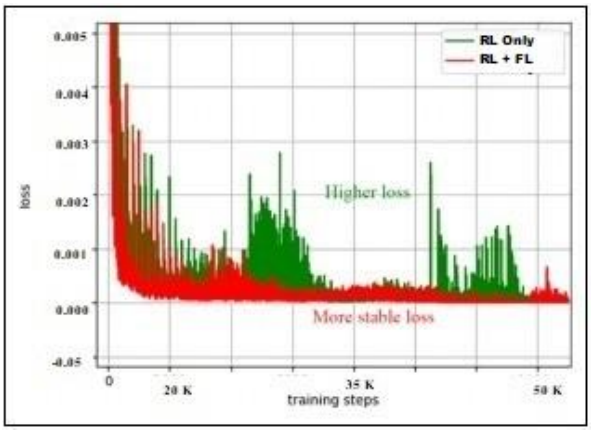


Fig. 10. Loss comparative enlarged view.

V. Evaluation and Verification of the Developed Policy

This work conducted three comprehensive tests to rigorously evaluate the performance of the proposed method. Each simulation aimed to assess the effectiveness of the respective algorithm in enabling the mobile robot to learn and navigate its environment autonomously.

To verify the practical performance of the model, physical tests were conducted on the robotic machine based on the robot operating system (ROS). The TurtleBot machine car was employed for these experiments to ensure consistency and reliability. The test environment comprised an obstacle zone constructed in the laboratory terrain, with the ideal distance from the starting point to the target point set at 8.3 meters. Fig. 5(a) and Fig. 5(b) depict the laser environment after its construction. It's important to note that the use of TurtleBot in these experiments is not meant to directly reduce errors in the algorithm. Instead, TurtleBot provides a standardized platform for testing, ensuring consistency and reliability across multiple trials. The choice of TurtleBot contributes to the creation of a controlled and reproducible testing environment, minimizing potential errors arising from variations in hardware and environmental conditions. This emphasis on error reduction pertains to the establishment of a robust and reliable basis for evaluating the proposed method's performance in real-world scenarios rather than directly mitigating errors in the algorithm or system. Following the integration of the trained model into the navigation function package, a meticulous series of verification tests was carried out to assess its performance. Each testing round consisted of five restarts, with three experiments conducted within each round to ensure the robustness of the evaluation process. For instance, in the first round of experiments, the robot's performance was tested through three individual trials: the first trial covered a distance of 8.8 meters in 77 seconds; the second trial covered 9.0 meters in 78 seconds, and the third trial spanned 8.6 meters in 73 seconds. By calculating the mean of these results, we obtained an average performance of 8.8 meters covered in 76 seconds.

Table III presents the detailed results of the first round, where the robot covered distances of 8.8 meters in 65 seconds, 8.6 meters in 53 seconds, and 8.7 meters in 56 seconds during the three tests. The calculated mean for Table II was 8.7 meters covered in 58 seconds. Notably, Table II exhibited a higher learning rate compared to Table II, indicating improved efficiency in path planning and execution.

TABLE II. The Example of RL Algorithm

Examples	**length/time**			
	Test 1	Test 2	Test 3	Mean
First Round	8.8 m/77 s	9.0 m/78 s	8.6 m/73 s	8.8 m/76 s
Second Round	9.3 m/86 s	9.1 m/83 s	8.9 m/74 s	9.1 m/81 s
Third Round	8.9 m/68 s	8.6 m/63 s	9.2 m/70 s	8.9 m/67 s
Fourth Round	9.1 m/78 s	8.9 m/73 s	8.7 m/71 s	8.9 m/74 s
Fifth Round	9.2 m/80 s	9.2 m/77 s	8.6 m/77 s	9.0 m/78 s

TABLE III. The Example of Reinforcement Learning with Fuzzy Logic Algorithm

Examples	**Length/Time**			
	Test 1	**Test 2**	**Test 3**	**Mean**
First Round	8.8 m/65 s	8.6 m/53 s	8.7 m/56 s	8.7 m/58 s
Second Round	8.8 m/63 s	8.9 m/69 s	8.7 m/60 s	8.8 m/64 s
Third Round	8.7 m/66 s	8.7 m/70 s	8.5 m/68 s	8.6 m/68 s
Fourth Round	8.7 m/73 s	8.8 m/65 s	8.6 m/66 s	8.7 m/68 s
Fifth Round	8.4 m/71 s	8.7 m/73 s	8.4 m/69 s	8.5 m/69 s

The overarching analysis of these comprehensive tests reveals that the Fuzzy Logic approach consistently outperforms other methods in terms of both time consumption and path length, particularly in the scenario represented in Fig. 5(b). It consistently finds shorter paths in less time, highlighting its superior efficiency. Additionally, the Fuzzy Logic method demonstrates remarkable stability in locating multiple paths, underscoring its prowess in complex environment path-finding.

However, it's important to acknowledge certain limitations associated with the Fuzzy Logic-based approach. While it excels in various aspects of path planning, it may face challenges when confronted with highly dynamic and rapidly changing environments. Fuzzy Logic, being rule-based and reliant on predetermined membership functions, might struggle to adapt swiftly to unpredictable obstacles or situations. Additionally, its performance could be impacted by the complexity and size of the environment, as processing a vast amount of data can introduce computational overhead.

Therefore, while the Fuzzy Logic approach proves highly effective in many scenarios, it may not be the optimal choice for applications demanding real-time adaptability in extremely dynamic settings. Exploring its boundaries and considering alternative approaches for such specific scenarios remains a valuable avenue for future research and development.

VI. Conclusion

This research introduced a novel navigation method based on Q- learning and fuzzy logic for efficient path planning of agents in diverse environments. The proposed approach combines the strengths of deep learning with symbolic reasoning, specifically Fuzzy Logic, to overcome the challenges faced by traditional DRL methods in mobile robot navigation, reducing the global path search time by 6-9% and shortening the average path search length by 4-10% compared to pure Q-learning. The incorporation of symbolic representations in the learning process leads to reduced training convergence time and more practical path planning results. The experimental results demonstrate its efficiency and effectiveness in complex environments, making it a promising solution for autonomous robotic navigation in urban megacities. As future work, the effectiveness of new RL algorithms will be explored in even more challenging environments, further advancing the field of autonomous robotic navigation.

REFERENCES

[1] U. Rakhman, J. Ahn, and C. Nam, "Fully automatic data collection for neuro-symbolic task planning for mobile robot navigation," in Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 450–455. doi: 10.1109/SMC52423.2021.9658822.

[2] A. Zhu and S. X. Yang, "Neurofuzzy-based approach to mobile robot navigation in unknown environments," IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, vol. 37, no. 4, pp. 610–621, Jul. 2007, doi: 10.1109/TSMCC.2007.897499.

[3] P. Coraggio and M. De Gregorio, "A Neurosymbolic Hybrid Approach for Landmark Recognition and Robot Localization."

[4] O. Castillo, R. Martínez-Marroquín, P. Melin, F. Valdez, and J. Soria, "Comparative study of bio-inspired algorithms applied to the optimization of type-1 and type-2 fuzzy controllers for an autonomous mobile robot," Inf Sci (N Y), vol. 192, pp. 19–38, Jun. 2012, doi: 10.1016/j.ins.2010.02.022.

[5] Y. Li, "Deep Reinforcement Learning: An Overview," Jan. 2017, [Online]. Available: http://arxiv.org/abs/1701.07274.

[6] H. Van Hasselt, A. Guez, and D. Silver, "Deep Reinforcement Learning with Double Q-Learning." [Online]. Available: www.aaai.org.

[7] K. Zhang, F. Niroui, M. Ficocelli, and G. Nejat, "Robot Navigation of Environments with Unknown Rough Terrain Using deep Reinforcement Learning," in 2018 IEEE International Symposium on Safety, Security, and Rescue Robotics, SSRR 2018, Institute of Electrical and Electronics Engineers Inc., Sep. 2018. doi: 10.1109/SSRR.2018.8468643.

[8] N. Altuntas, E. Imal, N. Emanet, and C. N. Öztürk, "Reinforcement learning-based mobile robot navigation," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 24, no. 3, pp. 1747–1767, 2016, doi: 10.3906/elk-1311-129.

[9] C. Pérez-D'Arpino, C. Liu, P. Goebel, R. Martín-Martín, and S. Savarese, "Robot Navigation in Constrained Pedestrian Environments using Reinforcement Learning," in Proceedings - IEEE International Conference on Robotics and Automation, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 1140–1146. doi: 10.1109/ICRA48506.2021.9560893.

[10] V. Zambaldi et al., "Relational Deep Reinforcement Learning," Jun. 2018, [Online]. Available: http://arxiv.org/abs/1806.01830.

[11] D. Dong, C. Chen, J. Chu, and T. J. Tarn, "Robust quantum-inspired reinforcement learning for robot navigation," IEEE/ASME Transactions on Mechatronics, vol. 17, no. 1, pp. 86–97, Feb. 2012, doi: 10.1109/TMECH.2010.2090896.

[12] Y. Zhu, Z. Wang, C. Chen, and D. Dong, "Rule-Based Reinforcement Learning for Efficient Robot Navigation With Space Reduction," IEEE/ASME Transactions on Mechatronics, vol. 27, no. 2, pp. 846–857, Apr. 2022, doi: 10.1109/TMECH.2021.3072675.

[13] J. Priya Inala et al., "Neurosymbolic Transformers for Multi-Agent Communication." [Online]. Available: https://github.com/jinala/.

[14] P. Coraggio, M. De Gregorio, and M. Forastiere, "ROBOT NAVIGATION BASED ON NEUROSYMBOLIC REASONING OVER LANDMARKS," 2008. [Online]. Available: www.worldscientific.com.

[15] M. Sokolov, R. Lavrenov, A. Gabdullin, I. Afanasyev, and E. Magid, "3D modelling and simulation of a crawler robot in ROS/Gazebo," in ACM International Conference Proceeding Series, Association for Computing Machinery, Dec. 2016, pp. 61–65. doi: 10.1145/3029610.3029641.

[16] K. Takaya, T. Asai, V. Kroumov, and F. Smarandache, Simulation Environment for Mobile Robots Testing Using ROS and Gazebo. 2016. doi: 10.0/Linux-x86_64.

[17] K. Sukvichai, K. Wongsuwan, N. Kaewnark, and P. Wisanuvej, "Implementation of Visual Odometry Estimation for Underwater Robot on ROS by using RaspberryPi 2."

[18] N. Botteghi, B. Sirmacek, K. A. A. Mustafa, M. Poel, and S. Stramigioli, "On Reward Shaping for Mobile Robot Navigation: A Reinforcement Learning and SLAM Based Approach," Feb. 2020, [Online]. Available: http://arxiv.org/abs/2002.04109.

[19] A. V. Bernstein, E. V. Burnaev, and O. N. Kachan, "Reinforcement learning for computer vision and robot navigation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2018, pp. 258–272. doi: 10.1007/978-3-319-96133-0_20.

[20] A. Newman, G. Yang, B. Wang, D. Arnold, and J. Saniie, "Embedded Mobile ROS Platform for SLAM Application with RGB-D Cameras," in IEEE International Conference on Electro Information Technology, IEEE Computer Society, Jul. 2020, pp. 449–453. doi: 10.1109/EIT48999.2020.9208310.

[21] Q. Jiang, "Path Planning Method of Mobile Robot Based on Q-learning," in Journal of Physics: Conference Series, IOP Publishing Ltd, Feb. 2022. doi: 10.1088/1742-6596/2181/1/012030.

[22] K.-H. Park, Y.-J. Kim, and J.-H. Kim, "Modular Q-learning based multi-agent cooperation for robot soccer," 2001. [Online]. Available: www.fira.net.

[23] G. Antonelli, S. Chiaverini, and G. Fusco, "A fuzzy-logic-based approach for mobile robot path tracking," IEEE Transactions on Fuzzy Systems, vol. 15, no. 2, pp. 211–221, Apr. 2007, doi: 10.1109/TFUZZ.2006.879998.

[24] E. Ayari, S. Hadouaj, and K. Ghedira, "A fuzzy logic method for autonomous robot navigation in dynamic and uncertain environment composed with complex traps," in Proceedings - 5th International Multi-Conference on Computing in the Global Information Technology, ICCGI 2010, 2010, pp. 18–23. doi: 10.1109/ICCGI.2010.47.

Brain Tumor Segmentation Algorithm Based on Asymmetric Encoder and Multimodal Cross-Collaboration

Pengyue Zhang[1], Qiaomei Ma[2]*
Software School, North University of China, Taiyuan 030051, China[1, 2]
Shanxi Medical Imaging Artificial Intelligence Engineering Technology Research Center
(Central North University), Taiyuan 030051, China[2]

***Abstract*—To address the challenges of insufficient multimodal information fusion and insufficient long-range dependencies features extraction for brain tumor segmentation, this paper propose a novel network based on asymmetric encoder and multimodal cross-collaboration. The network employs an asymmetric encoder-decoder architecture. Firstly, the invert ConvNext split convolution (ICSC) block is used in the local refinement encoder and improved SwinTransformer with DscMLP enhancements (DscSwinTransformer) module is used in global associative encoder. The local and long-range dependencies of each stage of two parallel encoders can be well extracted by hybrid fusion. Moreover, this paper adds a multimodal cross-collaboration (MCC) module at the beginning of the two encoders to fully exploit the complementary information between modalities and reduce the reliance on a single modality during model training. Coordinate Attention (CA) is used in the bridge part of the encoder and decoder to capture important spatial location information. Then, the depthwise separable convolution (DscConv) module is used in the decoder branch to reduce the computation while maintaining good feature extraction ability. Finally, this paper uses a hybrid loss function of BCE, Dice and L2 loss to mitigate the problem of class datas imbalance. Experimental results show that our model achieves Dice coefficients of 0.897, 0.905 and 0.824 in the whole, core and enhanced tumor regions, respectively. These results show that the performance of our proposed method outperforms in comparison with several existing methods in core and enhanced tumor regions.**

***Keywords*—Brain tumor; multimodal cross-collaboration; asymmetric encoder; coordinate attention**

I. Introduction

A brain tumor is a mass or cluster of aberrant cells in the brain that impairs brain tissue function. Brain tumors are primarily classified as malignant or benign [1]. Malignant brain tumors are one of the most severe cancers at present, posing an increasing threat to human health. Brain tumors are classified as I-IV by the World Health Organization. Because the higher the grade of the brain tumor, the shorter the patient's survival time [2-3], early detection and treatment of brain tumors is critical. However, because the shape, size, location, and border of MRI images from various brain tumor patients vary, it is difficult to properly segment the brain tumor area. Manual segmentation of brain tumors by doctors is very time-consuming and inconsistent among different doctors for the same patient, while automatic segmentation techniques based on brain tumor MRI images can automatically locate and segment the shape, position and boundary of the brain tumor area, thus assisting doctors in diagnosing patients' conditions and alleviating their workload. Therefore, the research of brain tumor segmentation algorithm has significant scientific value and clinical relevance for efficient diagnosis of brain tumors.

At present, most segmentation networks do not properly use multi-modal complementary information. This study proposes a multi-modal cross-coordination feature fusion module, which reduces the feature dependence on a single mode and obtains rich context information of different modal complementary information. In order to obtain long-range dependencies information while extracting local feature information, this paper uses dual encoders to obtain spatial and coordinate attention information at different stages. In this paper, the mixed loss function is further designed to alleviate the problem of class imbalance in brain tumor data sets, so that the model can effectively segment different types of brain tumor regions.

II. Related Work

With the advancement of deep learning technologies, convoluted neural networks have emerged as the primary way for diagnosing brain tumor locations. U-net [4] is a typical segmentation network based on the encoder-decoder structure. Later, the Unet-type structure was further developed, such as Unet++ [5] with nested and densely connected structures, DenseUnet [7] that combines DenseNet [6] network and U-net, and Vnet [8] structure for volumetric segmentation. Convolutional neural networks can capture local features, but they have difficulty in modeling explicit long-range dependencies from the global feature space.

However, Local and global features are essential for dense prediction tasks. Vision Transformer [9] leverages self-attention mechanism to model long-range information, enabling CNN hybrid Transformer to fuse and extract local and distant features effectively. In this regard, TransBTS [10] network is proposed, which incorporates Transformer into the 3D CNN encoder-decoder architecture for the first time, enhancing global feature extraction. TransBTSV2 [11] further improves TransBTS by redesigning the Transformer module

and introducing deformable bottleneck module to capture shape-sensitive local features. SwinBTS [12] structure employs 3D SwinTransformer as both encoder and decoder of the network to extract global information from feature maps efficiently, using convolution operation for upsampling and downsampling. Unetr [13] network connects the Transformer encoder to the decoder with different resolutions through skip connections, capturing global multi-scale information more effectively.

Moreover, different brain tumor regions have large scale differences, and the receptive field of ordinary convolution is not enough to extract rich contextual feature information. In this regard, Liu et al. [14] proposes a lightweight ADHDC-Net network that combines hierarchical convolution with different dilation rates and tumor region relation-guided attention; Chang et al. [15] proposes a dual-path and multi-scale attention fusion module that merges feature maps with different receptive fields for dense pixel prediction; Rehman et al. [16] designs SDS-MSA-Net, which extracts features from 3D and 2D inputs separately, and uses selective depth supervision to assist the output, accelerating the model convergence speed, but at the same time processing 3D and 2D resources increases the computational cost. The above improved structures enhance the extraction ability of global features and multi-scale attention features respectively, but most of the current networks are limited to simple concatenation fusion of multi-modal brain tumor data input level, which cannot fully utilize the complementary fusion information between different modalities. Therefore, Liu et al. [17] designs a two-stage network that performs pixel-level fusion and feature-level fusion of multi-modal images to achieve more fine-grained utilization of multi-modal information; Zhou et al. [18] proposes an attention feature fusion module that can fuse different modalities and selectively extract useful feature information, but the core of the above networks still needs to improve the segmentation accuracy of the enhanced tumor region.

To solve the aforementioned challenges, our study offers an asymmetric encoder and multimodal cross-collaboration brain tumor segmentation network (AEMCCNet). This paper's primary contributions are summarized as follows:

1) This paper proposes an asymmetric encoder-decoder structure, where parallel local refinement encoder and global associative encoder use redesigned invert ConvNext split convolution (ICSC) block and improved SwinTransformer [19] with DscMLP enhancements (DscSwinTransformer) module respectively, which can effectively capture the fusion information of local details and long-range dependencies features in three stages of the encoder.

2) To reduce the model's dependence on a single brain tumor modality during training, this paper proposes a multimodal cross-collaboration (MCC) module, which can fully utilize the complementary information between modalities.

3) To obtain more accurate segmentation results, this paper uses coordinate attention (CA) Module [20] in AEMCCNet, which encodes the channels along horizontal and vertical directions. This transformation can capture remote features along one spatial direction and preserve precise location information along another direction, which is very important for generating spatial detail selective information.

4) To tackle the class imbalance issue, this paper employs a hybrid loss function composed of binary cross entropy, Dice, and L_2, which enhances brain tumor segmentation accuracy even further.

III. Methodology

A. AEMCCNet Network

The overall architecture of the brain tumor segmentation network based on asymmetric encoder and multimodal cross-collaboration proposed within this study is seen in Fig. 1.

The network model is an asymmetric encoder-decoder structure, where T1 and T1ce modalities are the inputs of the local refinement encoder; T2 and Flair modalities are the inputs of the global associative encoder. Both the local refinement encoder and the global associative encoder first use MCC module designed in this paper to fully learn the cross-modal features and reduce the model's dependence on a single modality [21]. Then this paper uses the ICSC Block and DscSwinTransformer module designed in this paper respectively, and the parallel dual-stream encoders fuse with each other at each stage, increasing the link throughout low-level detail features and high-level semantic features. Moreover, CA module is applied to the fused feature maps along two dimensions of MRI images to aggregate features, model long-range dependencies and channel transformation, and enhances the extraction of useful information while suppressing the influence of invalid information on tumor segmentation performance. Finally, depthwise separable convolution (DsConv) [22] is used in the decoder module to acquire the semantic details of the fusion information obtained from asymmetric dual-stream encoder and low-level decoder modalities.

B. MCC Module

To reduce the dependence on a single brain tumor modality during the model training process, and to better utilize the complementarity between T1, T1ce modalities and T2, Flair modalities to cross-extract features, this paper designs a MCC module, as shown in Fig. 2.

First, modality A and modality B separately go through 7×7 channel-by-channel convolution (DwConv) to obtain rich context information of a single modality, and then cross-multiply with the features of another modality after 1×1 convolution to obtain y_{11} and y_{21}, respectively, as shown in Eq. (1) and Eq. (2), to extract recognizable features from one modality to assist in correcting another modality;

$$y_{11} = \mathrm{Dw7\times7}(B) \otimes \mathrm{Conv1\times1}(A) \tag{1}$$

$$y_{21} = \mathrm{Dw7\times7}(A) \otimes \mathrm{Conv1\times1}(B) \tag{2}$$

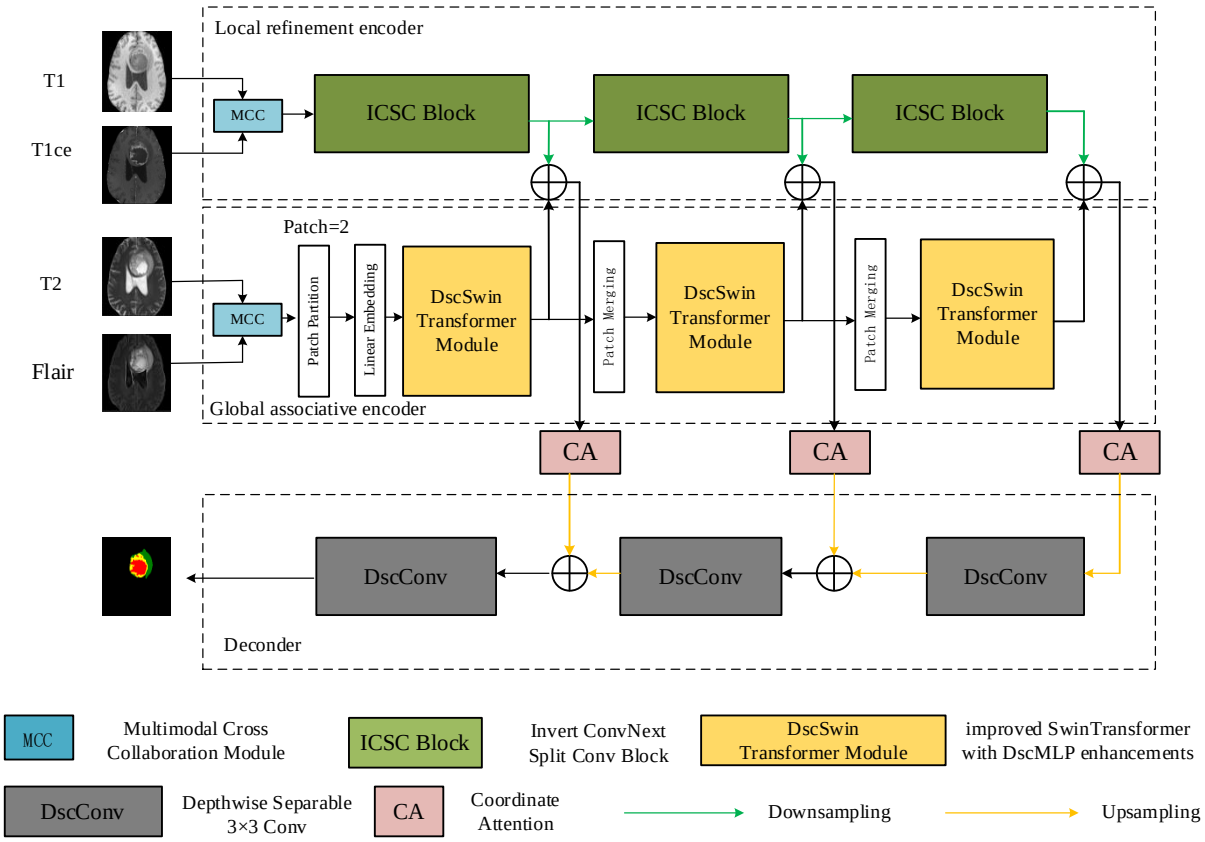


Fig. 1. Overall architecture of proposed AEMCCNet.

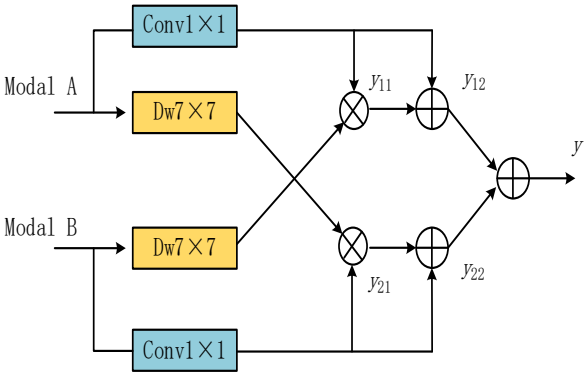


Fig. 2. Structure of a MCC module.

Then, the features of the same modality after 1×1 convolution are added and fused with the cross-fused features to generate the feature maps y_{12} and y_{22}, as stated in Eq. (3) and Eq. (4).

$$y_{12} = y_{11} + \text{Conv1}\times1(A) \tag{3}$$

$$y_{22} = y_{21} + \text{Conv1}\times1(B) \tag{4}$$

Finally, the branch modality features $y_{12}$and y_{22} are element-wise added and fused to generate the output feature map y of the module, as illustrated in Eq. (5).

$$y = y_{12} + y_{22} \tag{5}$$

C. ICSC Block

As shown in Fig. 3(a), MobileNetV2 [23] swaps the order of convolutional dimensionality increase and decrease in the Inverted Residuals structure, uses depthwise separable convolution to reduce the computational cost, and enhances the nonlinear expression ability of the network. As shown in Fig. 3(b), ConvNext Block [24] inherits the feature of wide convolutional dimension in the middle layer of Inverted Residul, and sets the depthwise separable convolution kernel size to 7×7, Padding=3. This paper draws on the advantages of these two modules and redesigns the ICSC Block of channel-split, as shown in Fig. 3 (c).

The input feature map of this module is $X \in R^{C\times H\times W}$,, where C is the number of channels. First, X is split into two $C/2$ branches X_{11} and X_{21} along the channel dimension. The left branch X_{11} uses 7×7 depthwise separable convolution to extract rich spatial context information, and then uses two 1×1 convolutions to increase and decrease the dimension respectively, obtaining the feature map X_1. The right branch X_{21} first uses 1×1 convolution to reduce the dimension, then uses 3×3 depthwise separable convolution to extract spatial rich information and increase the dimension, and then uses 1×1 convolution to reduce the dimension again, obtaining the feature map X_2. Then, the outputs of the two branches are concatenated and fused along the channel direction. Finally, the fused feature map is added with the original input feature map by identity connection to obtain the output feature map y of this module as shown in Eq. (6) to Eq. (8).

$$X_1 = Conv1(g(Conv1[L(Dw7(X_{11}))])) \quad (6)$$

$$X_2 = Conv1(Relu6(Dw3[B(Conv1(X_{21}))])) \quad (7)$$

$$Y = X + [X_1, X_2] \quad (8)$$

where, *DW7*() is a channel-wise convolution with a kernel size of 7×7, *Conv1*() is a convolution with a kernel size of 1×1, *L* is LN normalization operation, *B* is BN normalization operation, *g*() is gelu() activation function, *Relu6* is activation function, [·] is channel-wise concatenation and fusion.

D. DscSwinTransformer Module

Encoder-decoder structure based on CNN lacks the ability to capture long-range dependencies features, while lightweight SwinTransformer Block uses sliding window self-attention mechanism to capture global dependencies features information. In SwinTransformer Block [19], the multilayer perception (MLP) uses two fully connected layers for dimension transformation, but using fully connected layers in image segmentation causes partial segmentation information loss.

Inspired by the above content, this paper replaces the multilayer perception (MLP) in DscSwinTransformer Module with the designed depthwise separable perception (DscMLP), which can further refine the context information and improve the nonlinear transformation of features. As shown in Fig. 4, DscMLP takes the feature X after self-attention W-MSA, sliding window self-attention mechanism SW-MSA, and reshapes the shape of the feature map *X* first from [B, H×W, C] to X_1 in [B, C, H, W] dimensions, where B, C, H, and W are the batch size, the number of channels, the height, and the width, respectively, of the model training settings; Then it applies depthwise separable convolution and identity connection on X_1 in parallel respectively, and performs element-wise multiplication on the two-branch results; finally it reshapes the feature dimension to [B, H×W, C] dimension output feature map *Y*.

The DscSwinTransformer Module is used in the three stages of the global associative encoder, and the repetition number of SwinTransformer in each stage is 1; before the first stage the output feature map $y \in R^{C\times H\times W}$ of the multimodal cross-collaboration module is partitioned into *M* patches of size *P*×*P*,*P*=2 in the Patch Partition module, and each patch is reshaped into a one-dimensional vector $y_p \in R^{M\times(P\times P\times C)}$, then these patches are flattened along the channel direction and mapped to *D* dimensions by the Linear Projection module $E \in R^{(P\times P\times C)\times D}$, while adding a learnable position variable $E_{pos} \in R^{(P\times P\times C)\times D}$ to obtain the feature *z*, as shown in Eq. (9):

$$z = y_p E + E_{pos} \quad (9)$$

In the DscSwinTransformer Module, layer normalization (LN) is first used and residual connection is performed on W-MSA, SW-MSA, DscMLP, as shown in Fig. 4, the above process can be expressed as.

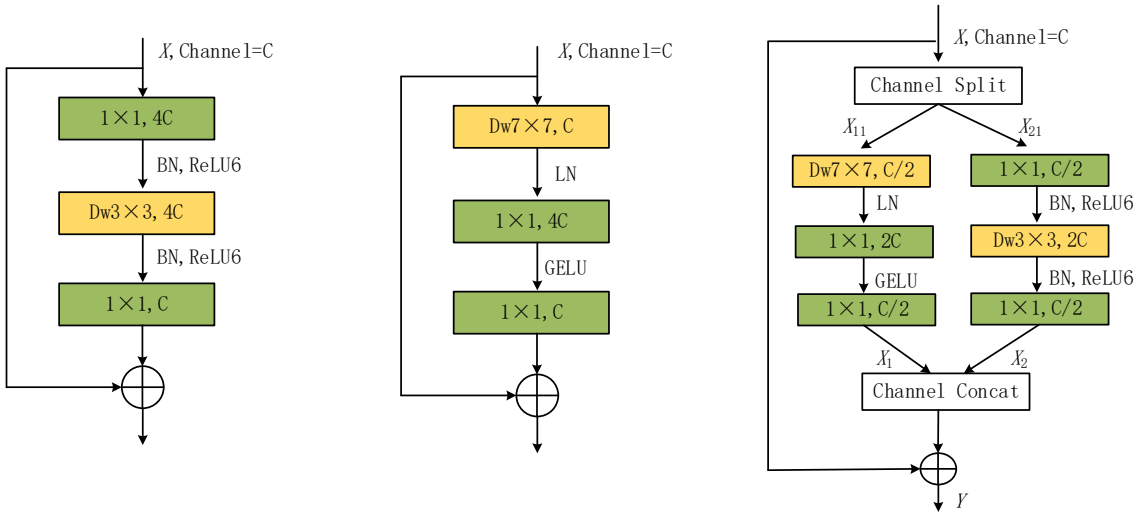
$$\hat{Z}^k = W - MSA(LN(Z^{k-1})) + Z^{k-1} \quad (10)$$

$$Z^k = DscMLP(LN(\hat{Z}^k)) + \hat{Z}^k \quad (11)$$

$$\hat{Z}^{k+1} = SW - MSA(LN(Z^k)) + Z^k \quad (12)$$

$$Z^{k+1} = DscMLP(LN(\hat{Z}^{k+1})) + \hat{Z}^{k+1} \quad (13)$$

To generate 2x downsampling, between the DscSwinTransformer modules use patch merging to increase dimensionality and decrease token numbers.



(a) InvertedResiduals Block (b) ConvNeXt Block (c) ICSC Block

Fig. 3. Comparison of Convolutional Blocks from left to right as (a) InvertedResidual Block, (b) ConvNeXt Block, (c) ICSC Block.

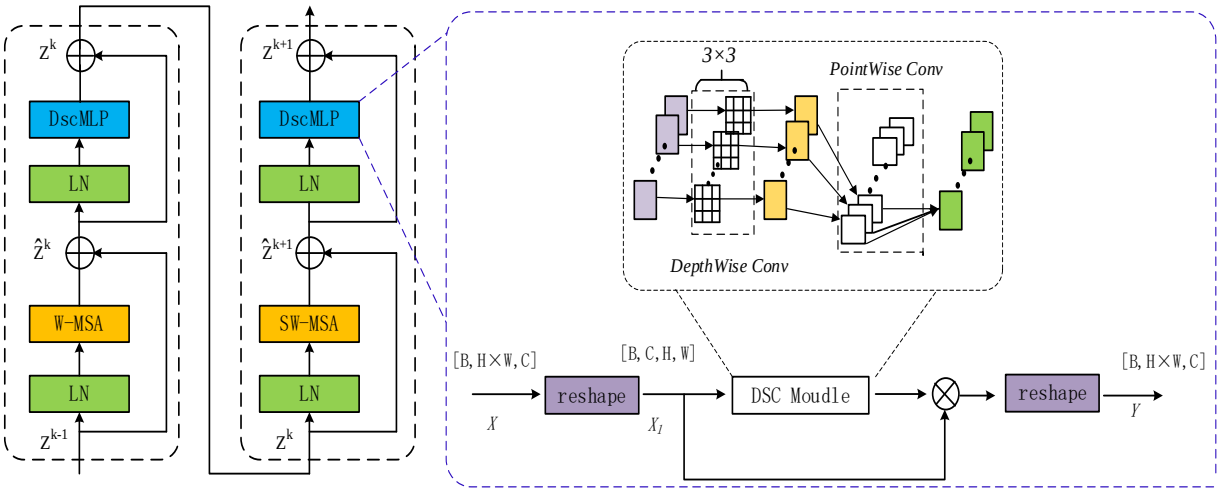


Fig. 4. Structure of DscSwinTransformer module.

E. CA Module

In deep network segmentation models, such as SE attention and CBAM attention, it has been proven that they can significantly enhance channel attention and spatial attention weights, and promote the model's segmentation performance, but they typically ignore positional details, which is vital for creating selective spatial features. Therefore, this paper introduces CA module [20], which embeds positional information into channel attention, as shown in Fig. 5.

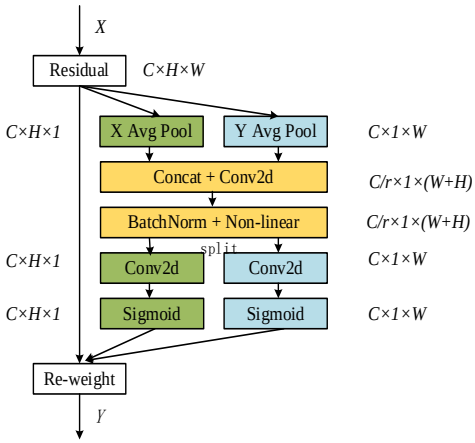


Fig. 5. Structure of CA module.

Coordinate information embedding alongside coordinated attention generation make up the two sections of the CA module. The coordinate details embedding implies encoding the input feature map *X* along both vertical and horizontal axes, respectively, using pooling kernels of shapes (H, 1) and (1, W) within the channels, to create two-dimensional feature maps that can capture distant features through one spatial direction and retain accurate positioning data along the other. Eq. (14) and Eq. (15) illustrate this:

$$Z_c^h(h)=\frac{1}{W}\sum_{0\le i\le W}X_c(h,i) \tag{14}$$

$$Z_c^w(w)=\frac{1}{H}\sum_{0\le j\le H}X_c(j,w) \tag{15}$$

where, $Z_c^h(h)$ is a result of the *C*th channel with height *h*, and $Z_c^w(w)$ is similarly.

The second transformation is the generation of CA module. First, $Z_c^h(h)$ and $Z_c^w(w)$ are concatenated, and then a 1×1 convolution function *F* and the feature mapping *f* of spatial data within multiple directions was extracted using an activation function that is nonlinear, as shown in Eq. (16).

$$f=\delta(F([Z^h,Z^w])) \tag{16}$$

Then *f* is decomposed into a pair of distinct tensors $f^h\in R^{C/r\times H}$ and $f^w\in R^{C/r\times W}$, and the convolution function F is used to transform them into tensors with the identical number of channels just like the input *X*, and the function of Sigmoid activating is utilizing to derive the attention weights g^h and g^w on two directions respectively. Ultimately, the module's initial feature map multiplies element by element with the two separate attention weights to yield the module's output *Y*, as indicated in Eq. (17) to Eq. (19).

$$g^h=\sigma(F(f^h)) \tag{17}$$

$$g^w=\sigma(F(f^w)) \tag{18}$$

$$Y(i,j)=X(i,j)\cdot g^h(i,j)\cdot g^w(i,j) \tag{19}$$

IV. Experimental Settings

A. Dataset

This paper uses the dataset from the Brain Tumor Segmentation (BraTS) competition in 2019, which contains 76 cases of low-grade glioma and 259 cases of high-grade glioma. This paper divides the training and testing data of the BraTS2019 dataset according to a ratio of 8:2. The

segmentation results are evaluated by the performance indicators of the whole tumor region (core tumor region and edema region), core tumor region (enhanced tumor region and necrosis region) and enhanced tumor region.

B. Data Preprocessing

Since training with 3D format images takes a long time and requires better GPU and more memory, this paper chooses to use 2D slices to train the proposed network. The size of the 3D data for each modality is 240×240×155. Since there is a lot of useless background information on the outer edge of the brain tumor data, which causes the problem of data class imbalance, this paper first crops the spatial size of the 3D data to 160×160×155 to eliminate the useless spatial background information, and then slices the data along the channel direction, transforming the 3D data into 155 slices of 160×160 2D slices to meet the needs of the model training in this paper.

Due to different imaging mechanisms, different modalities have different image contrast, so normalization is used to make the data intensity of different modalities balanced, which is conducive to the model using complementary information between different modalities. The normalization operation is shown in Eq. (20):

$$z = \frac{x - \bar{x}}{\lambda} \tag{20}$$

In the above equation, x is the cropped 2D sliced image, $\bar{x}$ is the mean value of the input image, λ is the standard deviation of the input image, and z is the normalized image.

As shown in Fig. 6, the images of four modalities and the ground truth label of a slice of a case after preprocessing in this paper are shown from left to right as (a) T1, (b) T1ce, (c)T2, (d) Flair and (e) Ground truth (GT) label. In the network model prediction image and the ground truth label, green represents edema tumor region, yellow represents enhanced tumor region and red represents necrosis and non-enhanced tumor region.

C. Experimental Environment Configuration

The software version is PyTorch 1.11.0 with Cuda 11.3, and the hardware environment consists of a 32 core CPU processor, 30GB RAM, and a GPU with NVIDIA RTX A5000, 24GB video memory. To update the model weights, this paper utilizes the Adaptive Moment Estimation (Adam) algorithm [25] as the optimizer; the detailed training experiment configuration is shown in Table I.

D. Evaluation Metrics

Four distinct assessment standards are employed to assess the segmentation accuracy of the model in this study to evaluate the how effective the suggested model algorithm. As described in Eq. (21), the Dice similarity coefficient is a value between 0 and 1. The closer to 1, the more similar the brain tumor segmentation result is to the manual label result, and the better the segmentation effect.

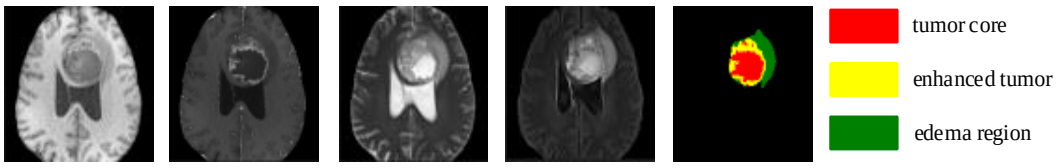
$$Dice = \frac{2TP}{2TP + FP + FN} \tag{21}$$

Sensitivity is a measure of the model's ability to predict positive pixels. Precision is used to measure the ability of the model to correctly predict pixels. As shown in Eq. (22) and Eq. (23).

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

Where TP stands for true positive pixels, FN represents for false negative pixels, FP means for false positive pixels, and TN indicates for true negative pixels.

$$Sensitivity = \frac{TP}{TP + FN} \tag{23}$$



(a) T1 (b) T1ce (c) T2 (d) Flair (e) GT

Fig. 6. Images after data preprocessing. from left to right as (a) T1, (b)T1ce, (c) T2, (d) Flair and (e) GT label.

TABLE I. EXPERIMENTAL CONFIGURATION

Configurations	Values
Software version	PyTorch11.0
GPU	NVIDIA RTX A5000
Optimizer	Adam
Initial learning rate	0.003
Momentum	0.09
Weight decay coefficient	0.00001
Batch szie	24
Tranning Epoches	300

To calculate the distance between the model segmentation border and the real label boundary, the Hausdorff distance (HD) has utilized, the higher the segmentation precision, the lower the Hausdorff distance. Eq. (24) shows the calculation formula.

$$Hausdorff = \max\left\{\max_{x\in X}\min_{y\in Y} d(x,y), \max_{y\in Y}\min_{x\in X} d(x,y)\right\} \quad (24)$$

Here y represents GT and x represents the predicted segmentation result, $d(x,y)$ is the Euclidean distance between x and y. In this paper, HD95 is used in the evaluation, which means taking the 95th percentile result.

E. Hybrid Loss Function

As indicated in Eq. (25), binary cross entropy (*BCE*) is often utilized as a loss function for performing segmentation operations for various medical data sets.

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \quad (25)$$

Where N denotes the overall amount of output pixels, y_i means the la-bel value for the ith pixel, and p_i defines the model prediction value for the ith pixel. Since the BCE loss function assigns equal weights to foreground and background pixels, but there is a large difference in the proportion of foreground and background pixels in multimodal brain tumors, and foreground pixels account for only a minority, there is a problem of data class imbalance. This study applies the *Dice* loss function to the *BCE* loss function to overcome the problem of data class imbalance, as seen in Eq. (26):

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i}^{N} p_i y_i + \varepsilon}{\sum_{i}^{N} p_i + \sum_{i}^{N} y_i + \varepsilon} \quad (26)$$

The value given by the parameter has been set to 10^{-6} to ensure data stability. In addition, overfitting is prone to occur during the training of the model, so based on this L_2 loss is introduced to alleviate the overfitting problem during the training of the model, which is advantageous to network convergence. The L_2 loss function is shown in Eq. (27):

$$L_2 = \frac{1}{N}\Sigma_i^N \left(y_i - p_i\right)^2 \quad (27)$$

In summary, the hybrid loss function in this paper is shown in Eq. (28):

$$L = L_{\text{Dice}} + \alpha L_{\text{BCE}} + \beta L_2 \quad (28)$$

The approach of controlling hyperparameters is applied in this paper to evaluate the most effective settings. The variation range of hyperparameters α and β values is shown in Table II

First, β is set to 0 to test the best hyperparameter α. As shown in Fig. 7, the fluctuation range of α is between 0 and 1. When α is 0.5, the model in this work hits the peak point in the whole tumor (WT), core tumor (TC), and enhanced tumor (ET) regions, implying that its predictive ability is best at this time.

On this basis, the experimentation of the optimal superparameter β was continued, as shown in Fig. 8, where the fluctuation of β ranges from 0 to 0.1 spacing. The Dice coefficient of this paper's model on WT, TC and ET regions reaches 0.897, 0.905 & 0.824 when the finalized parameter α is 0.5 and β is 0.05, and the prediction performance of this paper reaches the best.

The best hyperparameters α=0.5 and β=0.05 are substituted into the hybrid loss function. When the model training iteration number is 295 rounds, as illustrated in Fig. 9, the model's training and validation loss values tend to be optimum.

F. Ablation Experiment

To evaluate the efficacy of the design along with addition of modules in this study, under the same experimental conditions of using the same loss function and parameters, this paper replaces the original 3×3 convolution module of the encoder-decoder structure with depthwise separable convolution, which serves as the Baseline structure of our model. This paper adds different modules to the Baseline structure, and Table III displays the outcomes. Where MCC stands MCC module, ICSC represents ICSC block CA indicates CA module and DscSwinT represents DscSwinTransformer Module. By incorporating the MCC module to the first layer of the encoder before entering the Baseline, the Dice values of the whole tumor, core tumor and enhanced tumor regions in the model increase by 0.6%, 0.9% and 1.5%, respectively. Based on Baseline, using the ICSC block, the performance of the model in the WT, TC and ET regions is further improved. Similarly, based on Baseline, using the DscSwinTransformer to obtain the accuracy indicators of the ET and TC regions are significantly improved. Based on Baseline, adding CA module, the Dice indicators of the WT and TC are significantly improved.

TABLE II. VARIATION RANGE OF HYPERPARAMETERS

Hyperparameters	Variation Range
α	Between 0 and 1
β	Between 0 and 0.1

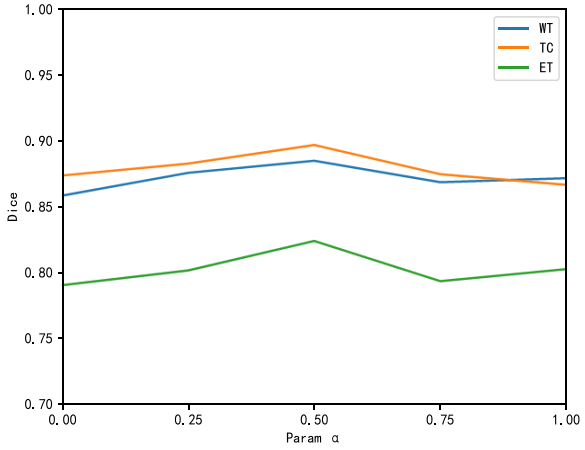


Fig. 7. Effect of hyperparameter α on model performance.

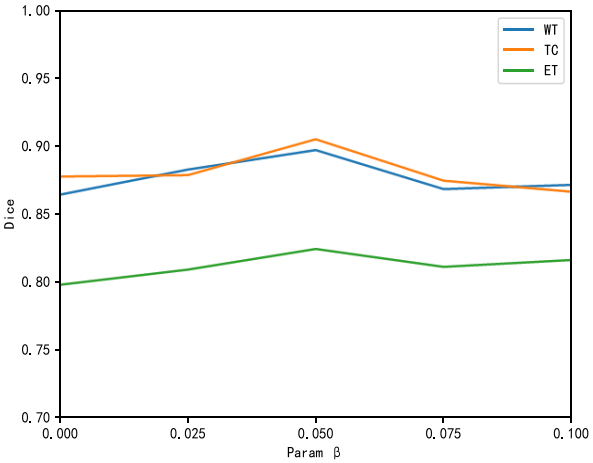


Fig. 8. Effect of hyperparameter β on model performance.

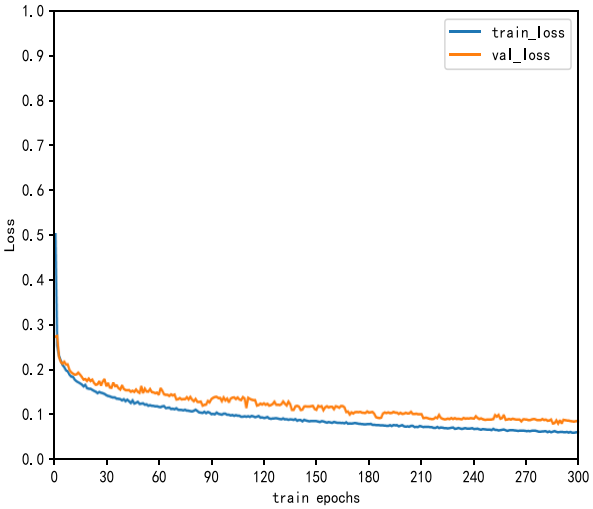


Fig. 9. Model training Loss variation.

Finally, this paper integrates the modules designed and used above into our structure. Compared with the Baseline structure, the Dice values of the WT, TC and ET increase by 5.1%, 5.3%, 6%, respectively. The hausdorff distance values of the three types of tumors are also the lowest values for ablation comparison experiments, proving the effectiveness of this method.

To test the effectiveness of adding different attentions to the bridging part between the asymmetric dual-stream encoder and decoder, this paper compares the CA module mechanism with SE [26] channel attention, CBAM [27] channel spatial attention, and ECA [28] efficient channel attention, respectively, as shown in Fig. 10 (a) and Fig. 10(b). After using the improved CA module in the decoder branch, the Dice similarity coefficient and HD95 of the model reach the best.

G. Experimental Results

To further verify the effectiveness of our method for multimodal brain tumor MR image segmentation, this paper uses part of the BraTS2019 dataset as the test set to compare with other advanced methods, and the results are shown in Table IV.

Compared with the classic 2D U-net and Unet++ networks, our model has a significant performance improvement in the whole, core and enhanced tumor regions. When compared with the advanced 3D models TransBTSV2, SwinBTS and MBANet, the 3D models have an advantage in segmenting the whole tumor due to their spatial continuity, but our 2D model uses DscSwinTransformer Module to strengthen the extraction of long-range dependencies features, making the Dice value of the whole tumor region close to the advanced 3D segmentation methods, and having the most optimal values in the overall evaluation indicators.

Finally, this paper segments some samples from the BraTS2019 test dataset and compare them with the input images and segmentation results as shown in Fig.11, where each row represents a patient case. U-net has over-segmentation phenomenon in the edema region of the second case, and also over-segments the core and enhanced tumor regions of the third case; The current advanced SwinBTS and TransBTSv2 networks have good segmentation effects on the edema region due to their spatial continuity, but they mis-segment part of the enhanced tumor region as necrotic tumor region in the second and third cases. Our proposed segmentation model improves by 5.7% and 2.2% respectively on the TC and ET tumor regions compared to the advanced TransBTSV2 network, showing that our method has better segmentation effects on the tumor core and enhanced regions.

TABLE III. MODULE ABLATION COMPARISON EXPERIMENTS

Baseline	MCC	ICSC	DscSwinT	CA	Dice			Hausdorff95(mm)		
					WT	TC	ET	WT	TC	ET
√					0.846	0.852	0.764	6.617	5.509	3.155
√	√				0.852	0.861	0.779	6.613	5.516	3.167
√		√			0.855	0.879	0.802	6.607	5.498	2.948
√			√		0.863	0.886	0.798	6.594	5.195	2.935
√				√	0.877	0.873	0.788	6.539	5.504	3.026
√	√	√	√	√	0.897	0.905	0.824	5.508	4.892	2.790

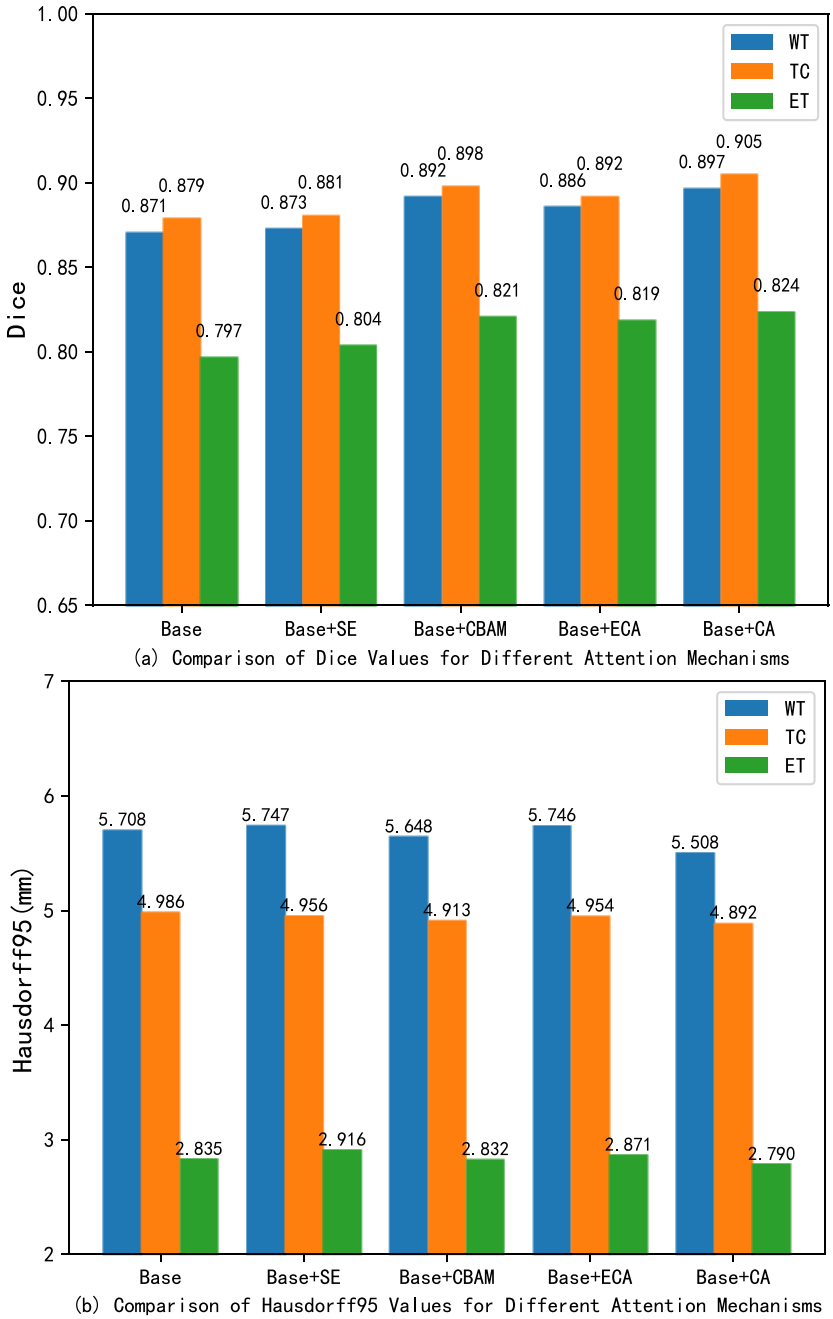
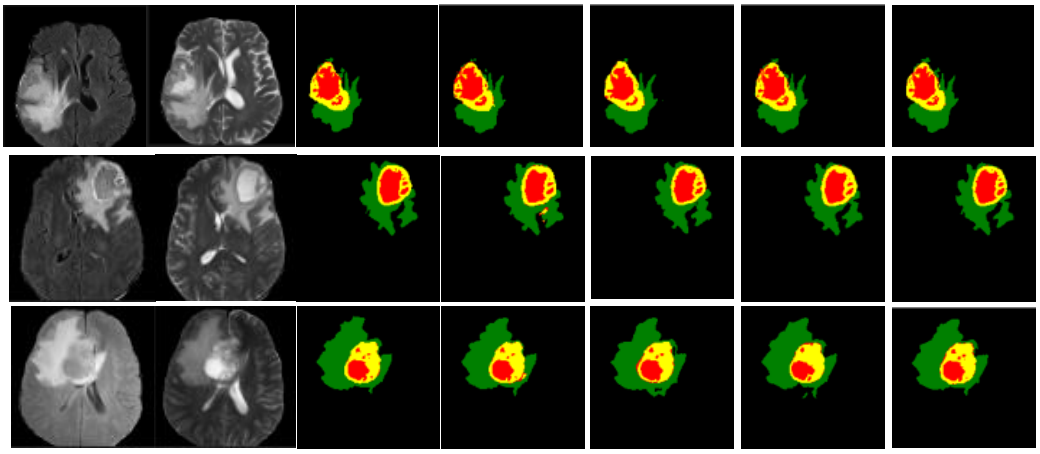


Fig. 10. Comparison of dice and hausdorff95 values with different attention modules added.

TABLE IV. COMPARATIVE EXPERIMENTS OF ADVANCED NETWORKS

Models	Dice↑			Precision↑			Sensitivity↑			Hausdorff95(mm) ↓		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
U-net [4]	0.834	0.823	0.769	0.871	0.898	0.800	0.856	0.908	0.813	6.648	6.596	4.062
Unet++ [5]	0.848	0.860	0.784	0.868	0.899	0.805	0.849	0.910	0.795	6.256	6.131	3.967
TransBTSv2 [11]	0.902	0.848	0.802	0.852	0.893	0.789	0.902	0.922	0.860	5.432	5.473	3.696
SwinBTS [12]	0.903	0.825	0.796	0.856	0.909	0.788	0.890	0.908	0.864	8.560	15.78	26.84
MBANet [29]	0.898	0.831	0.782	0.892	0.905	0.809	0.896	0.896	0.794	5.881	5.090	3.086
Ours	0.897	0.905	0.824	0.884	0.916	0.819	0.921	0.915	0.859	5.508	4.892	2.790



(a) Flair (b) T2 (c) GT (d) U-net (e) SwinBTS (f) TransBTSV2 (g) Ours

Fig. 11. Visual comparison of segmentation results: (a) Flair; (b) T2; (c) GT; (d) U-net; (e) SwinBTS; (f) TransBTSV2; (g) Ours.

H. Complexity Comparative Analysis

To better evaluate the performance of the model, we compare the number of parameters (Params) and the amount of computation (FLOPs) of the network model under the same input size, as shown in Table V. The quantity of parameters of our modelis reduced by about 75.45% compared to the CNN hybrid SwinTransformer structure SwinUnet, and the amount of computation is reduced by about 44.66%. Moreover, compared with the advanced 3D model SwinBTS, the number of parameters of our model is reduced by about 14.68%, and the amount of computation is reduced by about 70.09%. In summary, our model embeds depthwise separable convolution in each proposed module, and only uses one DscSwinTransformer Module in each stage of the global associative encoder. This substantially minimizes the number of parameters and the amount of computation required by the model, resulting in excellent segmentation performance for TC and ET regions.

TABLE V. COMPARATIVE ANALYSIS OF COMPLEXITY

Method	Params/M	FLOPs/G
U-net [4]	69.71	28.46
TransUnet [30]	96.07	48.34
TransBTS [10]	32.99	333.00
SwinBTS [12]	27.64	89.46
Unetr [13]	92.58	41.19
Ours	23.58	26.75

V. DISCUSSION

Due to insufficient hardware resources, this article only changed the number of rounds of model training on the basis of the same training parameters. The results showed that in the 295th round of model training, Dice, Precision, and HD95 in the model validation set were optimal. The sensitivity values in the TC and ET regions still had a gap compared to the optimal values of the network. This may be related to the removal of slices without lesion information during the process of slicing 3D images into 2D images in data preprocessing. Considering the segmentation performance presented by the four evaluation indicators, this article retains the 295th training model as the optimal model for the network to test the test set. If the GPU computing resources are sufficient, further increasing the training rounds can be considered to find the optimal value of the comprehensive evaluation indicators during the training process as the optimal model.

A lightweight AEMCCNet network model is proposed in this paper. It can be seen from subsections F and G of Section IV that compared with other network models, the proposed model has better segmentation effect in the core and enhanced tumor regions. In addition, the AEMCCNet network in Section IV, subsection H, reaches the optimal values of the number of parameters and the amount of computation. For the whole tumor area, 2D network cannot capture the information of adjacent sections of 3D brain tumor cases, but AEMCCNet is close to the segmentation effect of 3D network, and multiple adjacent sections can be combined for segmentation in the future.

VI. CONCLUSION

To address the problems of insufficient fusion of multimodal brain tumor information and inadequate extraction of long-range dependencies features, this paper adopts an asymmetric encoder-decoder structure, which incorporates the MCC module, ICSC block, DscSwinTransformer module, and CA module designed in this paper into the architecture, and offers an asymmetric encoder-based brain tumor segmentation algorithm with multimodal cross-collaboration. The MCC module can reduce the model's dependence on a single brain tumor modality during training and fully utilize the complementary information between modalities; the local refinement encoder branch uses the ICSC module to split channels and extract local detail features, enhancing the network's non-linear expression ability; the global associative encoder uses DscSwinTransformer module to strengthen the capture ability of long-range dependencies features; the bridge part between the asymmetric encoder and decoder uses the CA module to enhance the location weight of spatial detail

selective information; the decoder branch uses DscConv to maintain good feature extraction ability while reducing computation; During network training, a hybrid loss function is redesigned to handle the class imbalance and overfitting issues. The findings from the experiments reveal that the model's accurate segmentation of WT region is comparable to that of advanced 3D segmentation algorithms., while the Dice coefficient of TC and ET regions is better than other advanced models. At the same time, in the comparative experiment, it has the best values of other evaluation indicators, and uses DscConv throughout the model to minimize model parameters and calculations, but the extraction of edge detail information of enhanced tumor is still insufficient. Therefore, in future work, we will explore using an efficient encoder-decoder structure enhanced by edge operator attention or a low-parameter 2.5D network to further improve the segmentation accuracy of enhanced tumor region.

DATA AVAILABILITY

The datasets for this study can be found in the BraTS 2019 da taset available at: https://www.med.upenn. edu/cbica/brats2019/data.html.

ACKNOWLEDGMENT

We thank anonymous reviewers for valuable suggestions and comments. This work was supported by the Natural Science Foundation of Shanxi Province, China (No. 20210302123019).

REFERENCES

[1] Ostrom Q T, Cioffi G, Waite K, et al. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2014–2018. Neuro-oncology, 2021, 23(3): 1-105.

[2] Zhang R, Jia S, Adamu M J, et al. HMNet: Hierarchical Multi-Scale Brain Tumor Segmentation Network. Journal of Clinical Medicine, 2023, 12(2): 538.

[3] Wang J, Yu Z, Luan Z, et al. RDAU-Net: Based on a residual convolutional neural network with DFP and CBAM for brain tumor segmentation. Frontiers in Oncology, 2022,12:805263.

[4] Ronneberger O, Fischer P, Brox T.U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[5] Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation. 4th Deep Learning in Medical Image Analysis Work, 2018: 3-11.

[6] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[7] Kaku A, Hegde C V, Huang J, et al. DARTS: DenseUnet-based automatic rapid tool for brain segmentation. arXiv preprint arXiv:1911.05567, 2019.

[8] Milletari F, Navab N, Ahmadi S A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. 4th International Conference on 3DVision. IEEE, 2016 :565-571.

[9] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020. arXiv preprint arXiv:2010.11929, 2010.

[10] Wang W, Chen C, Ding M, et al. Transbts: Multimodal brain tumor segmentation using transformer. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021: 109-119.

[11] Li J, Wang W, Chen C, et al. TransBTSV2: Towards Better and More Efficient Volumetric Segmentation of Medical Images. arXiv preprint arXiv:2201.12785, 2022.

[12] Jiang Y, Zhang Y, Lin X, et al. SwinBTS: A method for 3D multimodal brain tumor segmentation using swin transformer. Brain sciences, 2022, 12(6): 797.

[13] Hatamizadeh A, Tang Y, Nath V, et al. Unetr: Transformers for 3d medical image segmentation[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022: 574-584.

[14] Liu H, Huo G, Li Q, et al. Multiscale lightweight 3D segmentation algorithm with attention mechanism: Brain tumor image segmentation. Expert Systems with Applications, 2023, 214: 119166.

[15] Chang Y, Zheng Z, Sun Y, et al. Dpafnet: A residual dual-path attention-fusion convolutional neural network for multimodal brain tumor segmentation. Biomedical Signal Processing and Control, 2023, 79: 104037.

[16] Rehman A, Usman M, Shahid A, et al. Selective Deeply Supervised Multi-Scale Attention Network for Brain Tumor Segmentation. Sensors, 2023, 23(4): 2346.

[17] Liu Y, Mu F, Shi Y, et al. Brain tumor segmentation in multimodal MRI via pixel-level and feature-level image fusion. Frontiers in Neuroscience, 2022, 16:1000587.

[18] Zhou T. Modality-level cross-connection and attentional feature fusion based deep neural network for multi-modal brain tumor segmentation. Biomedical Signal Processing and Control, 2023, 81: 104524.

[19] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[20] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.

[21] Zhou Y, Liang X, Gu Y, et al. Multi-classifier interactive learning for ambiguous speech emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 695-705.

[22] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[23] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.

[24] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11976-11986.

[25] Wu X, Huang F, Huang H. Fast stochastic recursive momentum methods for imbalanced data mining. 2022 IEEE International Conference on Data Mining (ICDM). IEEE, 2022: 578-587.

[26] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computervision and pattern recognition. 2018: 7132-7141.

[27] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module. Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[28] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[29] Cao Y, Zhou W, Zang M, et al. MBANet: A 3D convolutional neural network with multi-branch attention for brain tumor segmentation from MRI images. Biomedical Signal Processing and Control, 2023, 80: 104296.

[30] Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, 2021.

Blockchain Integrated Neural Networks: A New Frontier in MRI-based Brain Tumor Detection

Subrata Banik[1], Nani Gopal Barai[2], F M Javed Mehedi Shamrat[3]
Bangladesh Japan Information Technology Limited (BJIT Limited), Dhaka, Bangladesh[1, 2]
Department of Computer System and Technology, University of Malaya, Kuala Lumpur, Malaysia[3]

***Abstract*—Brain tumors originating from uncontrolled growth of abnormal cells in the brain, presents a significant challenge in healthcare due to their various symptoms and infrequency. While Magnetic Resonance Imaging (MRI) is essential for accurately identifying and diagnosing malignant tumors, manual interpretation is often complex and sensitive to mistakes. To address this, we introduce BrainTumorNet, a specialized convolutional neural network (CNN) created for MRI-based brain tumor diagnosis. We ensure improved image quality and a robust dataset for model training by including preprocessing approaches involving CLAHE and data augmentation. Additionally, we integrated a blockchain-based data retrieval technology to enhance the security, traceability, and collaboration in MRI data management across several medical institutions. This blockchain framework ensures that MRI data, once input from hospitals, stays immutable and can be safely retrieved based on unique hospital IDs, promoting a trustable environment for data exchange. Performance assessments conducted on multiple MRI datasets showcased BrainTumorNet's commendable proficiency, with accuracy rates of 98.66%, 97.17% and 94.24% on the dataset 1, dataset 2, and dataset 3, respectively. The model's performance was evaluated using a comprehensive set of metrics, including accuracy, specificity, recall, precision, f1-score, and confusion matrix. These measures are essential for evaluating a model's strengths and limits, emphasizing BrainTumorNet's ability to generate accurate and relevant predictions and its effectiveness in determining negative classification. BrainTumorNet's performance was compared with six renowned deep learning architectures: VGG16, ResNet50, AlexNet, MobileNetV2, InceptionV3, and DenseNet121. Our work highlights BrainTumorNet's potential capabilities in simplifying and boosting the accuracy of MRI-based brain tumor diagnosis while ensuring data integrity and collaboration through blockchain.**

Keywords*—Brain tumor; MRI imaging; BrainTumorNet; deep learning; image classification; augmentation

I. Introduction

Brain tumors develop from abnormal cell growth in the brain [1]. While cells normally follow a regular development and death cycle, sometimes they expand uncontrolled, resulting to harm in the brain. There are around 120 distinct forms of brain tumors and central nervous system (CNS) exist. In 2021, the American Cancer Society anticipated that brain and CNS cancers will cause 18,600 adult and 3,460 child deaths. The probability of surviving five years after being diagnosed is roughly 36%, and 10 years is 31% [2]. In 2019, the National Cancer Institute recorded 86,010 new cases of brain and CNS cancers in the U.S. It's believed that roughly 700,000 Americans live with a brain tumor, with 60,800 of them being non-malignant and 26,170 being cancerous [3]. Globally, the World Health Organization recorded roughly 9.6 million new cancer diagnoses in 2018 [4].

Early diagnosis of brain tumors is vital for patient survival. Analyzing brain tumor images effectively is vital to determining a patient's health state. Physicians and radiologists traditionally scan magnetic resonance (MR) images to discover abnormalities. However, this strategy depends greatly on the medical skill of the practitioner [5]. Differences in experience and the complexity of the imagery may make diagnosis with the human eye challenging. Doctors may struggle to rapidly analyze MR images because they often include several abnormalities or unnecessary data. The difficulty of accessing this large quantity of information increases as the number of data increases, making manual tumor identification time-consuming and costly. There's a rising demand for an autonomous computer-aided diagnostics (CAD) system to solve these issues. Such technologies may help doctors and radiologists by enabling fast and precise identification of cancers, thereby contributing to preserving human lives.

Artificial intelligence (AI) provides automation with capacities following human brain functions, such as learning and problem-solving. In the field of brain tumor identification and diagnosis, AI's accuracy provides crucial help, particularly considering the sensitive nature of the task. There are several initiatives to improve brain tumor categorization. However, the variation in tumor properties, such as their form, texture, and contrast variations across people, continues to be a problem. Machine learning (ML) and deep learning (DL), two branches of AI, have brought in a new era for the practice of neurosurgery. These cutting-edge methods include data preparation, feature extraction, selection, reduction, and classification as their final steps. Recent research [6] reveals that AI permits neurosurgeons to perform surgeries with unsurpassed confidence, enabling more precise brain tumor diagnosis.

Deep learning (DL) is an advanced version of machine learning that dives into data using multi-layered representations. By establishing a feature hierarchy, DL ensures fundamental features aid in developing advanced ones. This technique strengthens classic neural networks by incorporating several hidden layers between input and output, allowing them to capture complicated, non-linear relationships. Because of its excellent performance in recent years, DL has become the frontrunner in many medical image analysis difficulties, including tasks like image denoising, segmentation, authentication, and classification.

Brain tumors, due to their various appearances and sporadic frequency, have long been a main topic of worry in the medical community. Though useful, traditional diagnostic approaches often depend on human mistakes, particularly given the delicate nuances of MRI imaging. Recognizing these problems, we designed BrainTumorNet. Designed to overcome current diagnostic inadequacies, BrainTumorNet employs the ability of deep learning to negotiate the difficulties of MRI-based tumor identification. Our objectives for BrainTumorNet are not simply confined to enhanced diagnostic capabilities; we also emphasize a smooth, transparent, and most crucially, a secure data flow. This is where the integration of blockchain technology plays a key role, bringing in an approach where data integrity and trustworthiness become the standard rather than the exception. As our study develops a comprehensive strategy, blending cutting-edge AI with the resilience of blockchain, we intend to redefine the standards in brain tumor detection. The key findings of this research are:

- BrainTumorNet, a novel proposed model, demonstrated significant proficiency in identifying brain tumors using multiple MRI datasets.
- The addition of the CLAHE preprocessing approach significantly improved the image quality, leading to better model performance.
- Utilized a rigorous data augmentation method to increase the dataset's size and prevent the model from overfitting and improve its generalizability across various MRI images.
- We implemented a blockchain-based system for MRI data retrieval in recognition of the essential requirement for data security and traceability in medical diagnostics. This integration promises a clear, dependable, and immutable data handling procedure.
- Using six thorough performance indicators on three different datasets, we systematically evaluated BrainTumorNet's effectiveness. This comprehensive assessment confirmed the model's consistently good performance, proving its durability and dependability.

II. Literature Review

CNN has been extensively employed to address various issues, but its performance in health-related image processing applications is outstanding. Several techniques have been developed based on DL to identify brain tumors on MRI images in recent years. Most of them focused on binary segmentation to identify brain tumors.

Zhao et al. [7] developed an inventive method for brain tumor segmentation by integrating a complete Convolutional Neural Network (CNN) with Conditional Random Fields (CRFs). This unified framework assured visual excellence and spatial coherence in the segmentation outcomes. They employed three segmentation models trained on 2D image segments and slices from axial, coronal, and sagittal views. These models were combined using a voting-based fusion approach for precise tumor segmentation. On the other hand, Mohsen et al. [8] utilized a Deep Neural Network (DNN) classifier to differentiate an MRI dataset into four categories: normal tissue, glioblastoma, sarcoma, and metastatic bronchogenic carcinoma tumors. They incorporated Principal Component Analysis (PCA), an effective feature extraction technique, with the discrete wavelet transform (DWT) before classification. The ensuing evaluations showcased remarkable performance across all metrics.

Paul et al. [9] focused on 989 axial images, intending to simplify the neural network procedure by omitting the incorporation of three distinct axes with redundant diagnostic information. Both fully connected networks and CNNs were utilized for classification. When trained with axial data, the neural network achieved an impressive accuracy of 91.43% employing five-fold cross-validation, indicating its classification precision. On the other hand, Ari et al. [10] presented a three-step method. The initial phase contained preprocessing, where nonlocal means and local smoothing techniques decreased noise. In the subsequent step, the extreme learning machine with local receptive fields (ELM-LRF) was employed to classify cranial MR images as benign or malignant. Finally, image processing techniques segmented tumor areas in the third phase.

In this research, Abiwinanda et al. [11] aimed to train a CNN model to identify the three most prevalent forms of brain malignancies: gliomas, meningiomas, and pituitary tumors. They developed the simplest conceivable CNN architecture, consisting of one layer each of convolution, max-pooling, and flattening, followed by a complete connection from a single hidden layer. Using the basic architecture and no previous region-based segmentation, the study attains a maximum validation accuracy of 84.19%. Afshar et al. [12] have recently included newly generated CapsNets to alleviate the problem with CNNs that fail to properly exploit spatial interactions. Since the relationship between the tumor and the neighboring tissue is a crucial sign of tumor kind. Because of this, a specialized version of the CapsNet architecture for classifying brain tumors is proposed, including the tumor's coarse borders as additional inputs inside its pipeline to sharpen its attention.

However, Khan et al. [13] proposed a method comprising three critical phases: preprocessing, brain tumor segmentation applying k-means clustering, and benign/malignant tumor identification via a fine-tuned VGG19 model. This method has been assessed employing the BraTS 2015 benchmark dataset. Furthermore, they proposed synthetic data augmentation to expand the training dataset size, which consequently enhanced the classification accuracy. Yahyaoui et al. [14] offer a new semantic approach by fusing 2D and 3D MRI data in this study. Preprocessing, categorization, and fusion are the three stages that make up the whole system. To classify 2D brain data, the DenseNet model is used, and the 3D-CNN model was created specifically for 3D brain scans. Authors relied on a domain-specific ontology to accomplish the fusion of the output classes.

Furthermore, Murthy et al. [15] deployed the Optimized Convolutional Neural Network with Ensemble Classification (OCNN-EC) for tumor image classification. This deep learning approach encompasses an ensemble classifier containing a Deep Neural Network (DNN), an autoencoder, and a Support

Vector Machine (SVM). This ensemble replaces the completely connected layer once the ACV-DHOA has optimized the count of convolutional layers and hidden neurons. In this study, Latif et al. [16] argue that features from the MR images are extracted using a deep CNN network and then input into a support vector machine classifier. This research uses the BraTS dataset to categorize Gliomas into many classes. The suggested method attained a remarkable 96.19% accuracy.

Several recognized limitations in MRI-based brain tumor diagnosis were clear in light of available studies. Notably, many existing models struggle with accuracy problems, the difficulties of adding data to MRI datasets, and the complexity of managing various MRI data sources. We established BrainTumorNet to close these gaps and improve the diagnostic capability. The purpose of our proposed model is to achieve higher performance in the detection of brain tumors from MRI images. BrainTumorNet aims to establish a new standard in both accuracy and reliability in the field of brain tumor diagnosis, further enhanced by blockchain technology for secure and transparent data administration.

III. METHODOLOGY

This section highlights the main methods we employed during our research, including image preprocessing, deep feature extraction, blockchain integrity, and deep learning algorithms. The process of tumor detection is presented in Fig. 1.

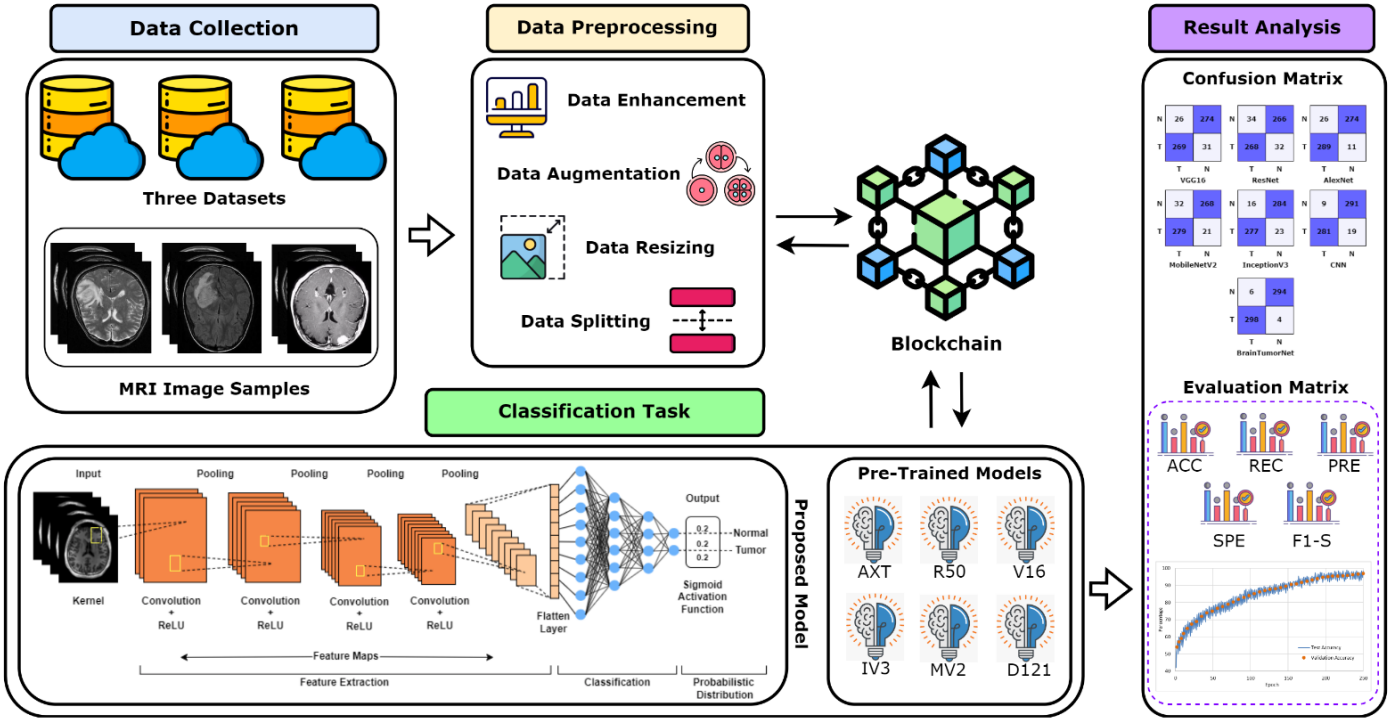


Fig. 1. The entire system of the proposed model BrainTumorNet.

A. Data Collection:

For the study, three publicly available datasets are used. The datasets are image datasets that contain MRI images of the brain. The detailed description of the employed dataset is stated below:

1) Dataset 1 (DT1): Br35H: Brain Tumor Detection 2020 [17] is the first dataset used in the study. It is an MRI image dataset. The dataset consists of 1500 images positive for brain tumors and 1500 images of normal brains. The normal and tumor class dataset samples are shown in Fig. 2(a).

2) Dataset 2 (DT2): The BraTS 2019 [18] is the second dataset used in this study. The dataset contains 1500 MRI images of the brain with tumor and 1500 MRI images of the normal brain. The sample images of this dataset are presented in Fig. 2(b).

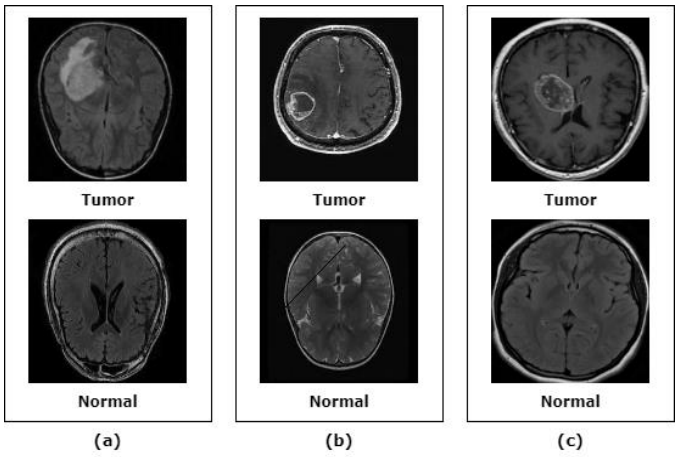


Fig. 2. Data sample of the three datasets (a) DT1 (b) DT2 (c) DT3.

3) Dataset 3 (DT3): The third dataset used in the study is available in the Kaggle data repository [19]. This is an unbalanced image dataset. It contains 170 MRI images of normal brain and 230 images of brain with tumor. The sample images of the normal and tumor class of DT3 are presented in Fig. 2(c).

B. Data-preprocessing:

1) Clache: Contrast limited adoptive histogram equalization (CLAHE) is used to enrich the image quality. CLAHE improves the contrast of the MRI image. Images are normalized using this technique, which also highlights finer information for Machine Learning (ML) classifiers to pick up on. By boosting contrast locally, the CLAHE method gets beyond the limitations of traditional, global approaches. Critical hyper-parameters for this technique are the tile size and clip limit. Multiple permutations of all these settings are examined before the ideal values of tileGridSize (8, 8), and the clip limit is settled on (3.7). Algorithm 1 displays the process of CLAHE.

Algorithm 1: CLAHE Working Process

Procedure CLAHE (Image I, ClipLimit c, GridSize g)
- $I_{gray} \leftarrow$ Convert I to grayscale
- $CLAHE_{object} \leftarrow$ Initialize CLAHE with c and g
- $I_{processed} \leftarrow$ Apply $CLAHE_{object}$ on I_{gray}
- Save $I_{processed}$ as 'clahe_output.jpg'
- **Return** $I_{processed}$

End

2) Data augmentation: The datasets, DT1 and DT2 are well balanced datasets. However, DT3 is a highly imbalanced dataset and consists of a small amount of data. Since ML models require a fairly large dataset to train and produce efficient performance result. To do this, we leverage the ImageDataGenerator class in the Keras library to generate high-quality images for the expansion of our training data. Table I has the image-making parameters. After augmentation, the final count of the DT3 is 500 images of normal brain and 500 images of brain tumor. Algorithm 2 represents the procedure of augmentation.

TABLE I. ATTRIBUTES OF IMAGEDATAGENERATOR

Attributes	Values
Shear_range	0.3
Zoom_range	0.2
Vertical flip	True
Horizontal flip	True
Rescale	1/255

Algorithm 2: Augmentation for Balancing Image

Procedure AugmentData
- Define ImageDataGenerator attributes
- Function LOAD_IMAGES (directory)
  - **Return** images from *directory*
  - End
- *NormalImages* ← LOAD_IMAGES (normal_dir)
- *TumorImages* ← LOAD_IMAGES (tumor_dir)
- **While** count of normal_images ! 500 **do**
  - Augment 'normal' images
- End
- **While** count of tumor_images ! 500 **do**
  - Augment 'tumor' images
- End
- Return Count

End

3) Resize image: For uniformity's sake, we resize every picture in the collection to 128×128 pixel, since the original dimensions of images in the three datasets (DT1, DT2, and DT3) vary substantially.

4) Dataset splitting: The datasets are divided into training and validation subsets. The division is done at a ratio of 80:20, where 80% of each data belongs to the training set and the remaining 20% to the test set. Table II shows the data division of each set.

TABLE II. DATA DISTRIBUTION OF THE DATASETS

Dataset	Training set	Test set	Total
DT1	2400	600	3000
DT2	2400	600	3000
DT3	800	200	1000

C. Blockchain Framework for Secure Data Management in BrainNet

Our proposed system incorporates blockchain architecture to ensure effective and secure data retrieval and sharing procedure designed for managing medical MRI data. Many institutions may cooperatively exchange and store MRI information to improve the BrainTumorNet model's detection abilities without compromising patient data integrity. Our work emphasizes the MRI data retrieval and sharing procedures inside our method, based upon the multi-organization blockchain designs mentioned by [20], [21].

1) Blockchain-based MRI data retrieval process: In the framework of our study, every participating data source/hospital provides MRI data, storing it as a unique transaction inside the blockchain network. Before this data is placed into the blockchain, it undergoes a preprocessing stage. Initially, the MRI images are improved using the CLAHE approach, which increases the contrast and overall visibility of tumor locations in the images. Following this, Data Augmentation methods are applied to extend the dataset intentionally. This assures increased variety and assists in training the BrainTumorNet model more robustly.

Retrieving this data from the blockchain nodes hinges on two key parameters: the distance between the nodes (d) and the unique ID of the hospital (ID). Each hospital denoted as x_i, is assigned a unique ID, determined in part by its distance d_{ij} from another hospital x_j. This can be represented as,

$$ID_i = f(x_i, d_{ij}) \quad (1)$$

where, f is a function computing the ID based on the hospital's attributes and its distance from other hospitals.

The blockchain system maintains log tables that register these unique IDs, ensuring that data integrity is maintained at all times. When BrainTumorNet needs to access specific MRI datasets for its diagnostic tasks, it retrieves the relevant data from the corresponding hospitals using the retrieval function $R(ID_i)$. This function returns the data x_i if ID_i exists in the log table, otherwise it returns to error.

Moreover, the neighborhood distance between two sources, x_i and x_j, plays a critical role in efficient data access. The distance d_{ij}, can be determined by,

$$d_{ij} = g(l(x_i), l(x_j)) \quad (2)$$

where, g is the distance function and $l(x)$ represents the location of source x.

This research effort aims to establish a new benchmark in secure, transparent, and collaborative medical diagnostics, laying the foundation for further advancements in this interdisciplinary field by combining the robustness of blockchain technology with the diagnostic prowess of BrainTumorNet and rigorous preprocessing steps.

D. Deep Learning Models for Classification

Our proposed model's major objective is to automatically identify people who have brain tumors while decreasing classification time and increasing accuracy. For the purpose of finding brain tumors utilizing multiples MRI datasets, we proposed a novel, reliable and robust CNN model BrainTumorNet. To establish the most effective transfer learning strategy for the classification assignment, six pre-trained models including VGG16, ResNet50, AlexNet, MobileNetV2, InceptionV3, and DenseNet121 are tested. The significant characteristics and some essential properties of the selected deep CNN models are compiled in Table III The next section includes an entire discussion of the model utilized in this study.

1) AlexNet: This model is consisting of five convolution layers and three fully linked layers. certain number of convolution layers are succeeded by the max-pooling layer (1, 2, and 5 layers). The ReLU nonlinearity is applied to the output of every fully connected and convolutional layer. Each of the connected layers has 4096 neurons [22]. During training, neurons are "turned off" with a predefined probability to prevent data over-adjustment using a regularization technique known as dropout [23].

TABLE III. PROPERTIES OF THE DEEP LEARNING MODELS EMPLOYED IN THIS RESEARCH

Model	Input Shape	Custom Input Shape	Parameters	Size (MB)
VGG16	224×224	224×224	138×10^6	552
ResNet50	224×224	224×224	25.6×10^6	102
AlexNet	227×227	224×224	60×10^6	240
MobileNetV2	224×224	224×224	3.5×10^6	14
InceptionV3	229×229	224×224	23.8×10^6	95
DenseNet121	224×224	224×224	8×10^6	32
BrainTumorNet	222×222	-	2.16×10^6	10

2) ResNet50: ResNet50 [24] is a 50-layer Convolutional Neural Network (CNN) composed of 48 fully connected layers, one max pooling layer, and one average pooling layer. It's capable of performing up to 3.8×10^9 floating-point computations. To resolve the vanishing gradient issue prevalent in traditional CNNs and expedite the training process, ResNet50 employs a spectrum of convolutional filters of varying sizes [25]. With fewer filters, ResNet's operates more promptly. This architecture is trained using approximately 23 million parameters. The network is designed to receive images where the height, breadth, and channel dimensions are multiples of 32.

3) VGG16: The Visual Geometric Group is referred to as VGG [26]. Simonay and Zimmerman [27] created the VGG model. VGG employs 3×3 convolutional layers that are layered on top of one another and become deeper over time. Max pooling layer is responsible for reducing the volume size. Afterwards, a softmax classifier is followed by two completely connected layers with a total of 4096 nodes each [27].

4) InceptionV3: At the ImageNet Recognition Challenge, Google introduced Inception version 3 [28]. The Auxillary Classifiers contain a label smoothing classifier, a factorized 7×7 convolution classifier, a batch norm classifier, an RMSProp optimizer, and a downscaling classifier for extracting and augmenting data from label sequences. The InceptionV3 model's training time is shortened by substituting bigger convolutions for smaller ones. Several optimization methods may be used to remove constraints and make an InceptionV3 model more flexible. The model is designed employing max pooling, convolutions, concatenations, dropouts, and fully-connected layers.

5) MobileNetV2: MobileNetV2 [38] is built upon an inverted residual structure, with residual connections

interconnecting its bottleneck layers. The intermediate expansion layer utilizes lightweight depth-wise convolutions to provide non-linear feature filtering. Following the initial convolutional layer with 32 filters, MobileNetV2 incorporates 19 residual bottleneck layers, resulting in a total of 53 layers for the network. The model has been pre-trained using over a million images from the ImageNet database, classified into 100 distinct categories. As a result, the network has gathered an enormous number of features from a diverse multitude of images.

6) DenseNet121: The Dense Convolution Network is a deep learning model that employs feedforward to link each layer to all subsequent layers [29]. DenseNet has (L(L+1))/2 direct lines compared to L connections for conventional L-layer CNNs. A feature map may be found in every layer of the model. Each layer's feature map serves as the following layer's input. It allows for the most information to be sent throughout the network by linking all levels directly to one another. DenseNet's primary benefits are a large reduction in parameter count, prevention of gradient runaway, improvement of feature diffusion, and encouragement of feature reuse. DenseNet needs fewer parameters than conventional CNN since the feature map is not repeatedly trained. Additionally, DenseNet uses regularization to lessen the possibility of overfitting. Each of the four dense blocks in DenseNet121 has six, twelve, twenty-four, and sixteen convolution blocks.

7) BrainTumorNet (Proposed Model): BrainTumorNet model is both small in size and computationally effective, improving performance across various datasets. This model is precisely designed for CNN that processes grayscale images with dimensions of 222×222×1 to identify brain tumors from MRI scans. The network is designed with four distinct blocks that have been optimized for feature extraction. Each of these blocks begins with a Conv2D layer, which is statistically stated by,

$$I'_{x,y,k} = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} I_{x-i,y-j,k} . K_{i,j,k} \qquad (3)$$

where, $I_{x,y,k}$ is the input feature map at position (x, y) for the k[th] channel. $K_{i,j,k}$ represents the kernel or filter at position (i, j) for the k[th] channel and $I'_{x,y,k}$ is the output of feature map after convolution at position (x, y) for the k[th] channel.

Following the convolutional transformation, there is a MaxPooling layer, which is presented as,

$$P_{x,y,k} = max_{i=0}^{s-1} \, max_{j=0}^{s-1} \, I_{sx+i,sy+j,k} \qquad (4)$$

Here, $P_{x,y,k}$ is the pooled output at position (x, y) for the k[th] channel, and s denotes the pooling window size.

Each block ends with a batch normalization layer, stabilizing the activations and ensuring the features remain standardized. As the depth of the network increases, the spatial dimensions are progressively reduced, encapsulating more complex and sensitive patterns essential to tumor identification. Once the entire spatial panorama has been thoroughly evaluated, the extracted features are flattened into a 1D tensor, covering a size of 10816. This tensor then runs across two dense layers, described by,

$$y = Wx + b \qquad (5)$$

Here, y is the output of the dense layer, W symbolizes the weight matrix, x represents the input to the dense layer, and b is the bias term.

The end of BrainTumorNet's activities is encapsulated in its last output layer, which is equipped with a sigmoid activation function. This design option provides binary classification capabilities, properly denoting whether the input MRI shows the existence or absence of a tumor. Table IV demonstrates the architecture of the proposed model BrainTumorNet. Fig. 3 illustrates a visual output of the proposed model.

This proposed model for detecting brain tumors using MRI images surpasses prior approaches through feature extraction and processing efficiency improvements. Increasing filters on convolutional layers are utilized to identify detailed features, which are essential for identifying brain tumor characteristics. Convolutional layers employ a balanced layout of 1×1 stride and 3×3 kernel size to optimize image processing, followed by max pooling. This arrangement minimizes the loss of information while preserving computational efficiency. Integrating BatchNormalization enhances learning stability, facilitating fast and consistent training. The model reaches the highest point with a binary classification output layer that is designed to ensure precise tumor detection. This architectural design signifies an important improvement in the accurate detection of brain tumors by addressing the distinct difficulties associated with MRI image analysis.

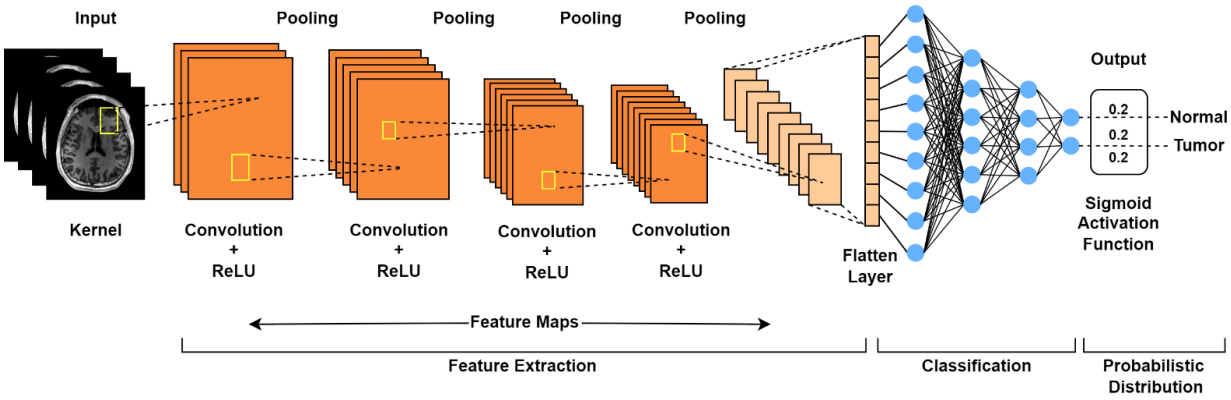


Fig. 3. The architecture of the proposed model BrainTumorNet.

TABLE IV. DESCRIPTION OF BRAINTUMORNET ARCHITECTURE LAYERS USED FOR BRAIN TUMOR DETECTION

Operation	Layer Name	No. of Filters	Stride Size	Kernel Size	Padding Size	No. of Channels	Input Shape	Output Shape
Input Image	Input Layer	-	-	-	-	3	222 × 222 × 1	-
Convolution	Convolution 2D	32	1×1	3×3	Same	32	222 × 222 × 1	222 × 222 × 32
Pooling	Maxpooling	-	2×2	2×2	Valid	32	222 × 222 × 32	111 × 111 × 32
Normalization	BatchNormalization	-	-	-	-	32	111 × 111 × 32	111 × 111 × 32
Convolution	Convolution 2D	32	1×1	3×3	Same	32	111 × 111 × 32	111 × 111 × 32
Pooling	Maxpooling	-	2×2	2×2	Valid	32	111 × 111 × 32	55 × 55 × 32
Normalization	BatchNormalization	-	-	-	-	32	55 × 55 × 32	55 × 55 × 32
Convolution	Convolution 2D	64	1×1	3×3	Same	64	55 × 55 × 32	55 × 55 × 64
Pooling	Maxpooling	-	2×2	2×2	Valid	64	55 × 55 × 64	27 × 27 × 64
Normalization	BatchNormalization	-	-	-	-	64	27 × 27 × 64	27 × 27 × 64
Convolution	Convolution 2D	64	1×1	3×3	Same	64	27 × 27 × 64	27 × 27 × 64
Pooling	Maxpooling	-	2×2	2×2	Valid	64	27 × 27 × 64	13 × 13 × 64
Normalization	BatchNormalization	-	-	-	-	64	13 × 13 × 64	13 × 13 × 64
Flattening	Flatten	-	-	-	-	-	13 × 13 × 64	10816
Fully Connected	Dense	128	-	-	-	-	10816	128
Fully Connected	Dense	64	-	-	-	-	128	64
Output	Dense	2	-	-	-	-	64	2

E. Hyperparameters Tuning:

The principal objective of this research is to develop the most efficient model BrainTumorNet for classifying brain MRI data. Hyperparameters are a group of factors that have the ability to impact the model's training process and provide the best outcomes [30-32]. These parameters include the volume of epochs, batch size, image size, optimizers, activation function, learning rate, decay rate, dropout rate and regularizer. During the experiment, we conducted several trials before settling on batch size, learning rate, regularization factor, etc. Different pre-trained models, including VGG16, ResNet50, AlexNet, MobileNetV2, InceptionV3, and DenseNet121, are used to execute the proposed brain tumor detection. Using a variety of optimizers, each model was assessed for 250 epochs. Each model is first tuned using Keras-tune to obtain the appropriate hyperparameter ranges. We employ the widely utilized grid search strategy for parameter tuning. Table V displays the parameters after tuning used during model training.

F. Evaluation Matrix

Accuracy, Specificity, Recall, Precision, and F1-score are some of the performance metrics calculated to assess the models' efficacy. Accuracy measures the rate of a model produce accurate predictions. The relevant predictions of positive classes are determined by calculating precision. Efficiency in predicting the negative class from the whole set of classes is measured by specificity. Contrarily, recall is the proportion of true positive classes that were anticipated. The F1-score measures how well specificity and recall are combined. Eq. (6) through Eq. (10) below express the parameters.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

$$Specificty = \frac{TN}{TN+FP} \quad (9)$$

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (10)$$

Here, True Positive (TP) is the proportion of positive predictions that turned out to be accurate. True Negative (TN) is the accurately anticipated negative images. False Negative (FN) represents the count of positive photos that were incorrectly labeled as negative. Similarly, False Positive (FP) represents the count of negative data that were incorrectly labeled as positive.

TABLE V. THE FINAL HYPERPARAMETERS USED TO TRAIN THE MODELS

Model	No. of Epochs	Batch Size	Image Size	Optimizers	Activation Function	Learning Rate	Decay Rate	Dropout Rate	Regularizer
VGG16	250	64	224×224	Adam	Softmax	0.000001	1e-3	0.2	1e-4
ResNet50	250	64	224×224	SGD	ReLU	0.0001	1e-4	0.2	1e-4
AlexNet	250	64	224×224	Adagrad	ReLU	0.00001	1e-2	0.2	1e-4
MobileNetV2	250	64	224×224	SGD	Softmax	0.1	1e-4	0.2	1e-4
InceptionV3	250	64	224×224	Adam	Sigmoid	0.001	1e-3	0.2	1e-4
DenseNet121	250	64	224×224	Adam	ReLU	0.0000001	1e-2	0.2	1e-4
BrainTumorNet	250	128	222×222	RMSProp	Sigmoid	0.01	1e-5	0.2	1e-4

IV. Experimental Results

A. Experimental Setup

We utilized the Keras 2.10.0, TensorFlow 2.0, and Python 3.7 programming languages to execute the proposed models and produce results. Visualization was done using the Seaborn and Matplotlib packages. System specifications include AMD Ryzen 7 running at 3.90 GHz, 32 GB of RAM, a MD Radeon RX 580 series GPU, and a Windows 10 setup.

B. Result Analysis

Six different transfer learning models and novel BrainTumorNet model were employed in the study to detect brain tumor from MRI data. Furthermore, the seven models are employed on the three datasets containing MRI images of the brain. The purpose of the study is to identify a robust model that can classify MRI data to diagnose brain tumor. The models were run for 250 epoch and the outcome of each epoch are recorded for all three datasets. Using Eq. (6) through Eq. (10), we can calculate the performance parameters of each model and so assess how well each model performs.

In Fig. 4 we can see the results of the performance indicators for DT1. Among the transfer learning models, it is seen, the model CNN consistently displays high performance, with an accuracy of 95.34%. A precision of 96.14% was attained, along with 95.78% recall, 96.32% specificity, and 95.95% f1-score. It is followed by InceptionV3 with specificity of 92.55%, similarly the model falls well short of perfection in accuracy (93.5%), precision (92.17%), recall (95.78%) and f1-score (93.13%). Likewise, the other models, including VGG16, AlexNet, MobileNetV2, and ResNet50, all have subpar accuracy (90.47%, 93.75%, 91.24% and 89.02%, respectively). Contrary to the transfer learning models, the proposed model performs with exceptionally high measures. The accuracy of the proposed model BrainTumorNet is 98.66%. Similarly, the specificity is 98.59% and the f1-score is 97.69%.

Fig. 5 displays the outcomes of DT2's key performance metrics. The CNN model shows the highest classification efficiency among the transfer learning models with an accuracy of 94.76% and an f1-score of 94.98%. The accuracy score of VGG16, ResNet50, AlexNet, MobileNetV2 and InceptionV3 are 85.78%, 91.45%, 92.01%, 89.71% and 90.83% respectively. Performance-wise, the proposed BrainTumorNet model is much superior to transfer learning methods with an accuracy of 97.17%.

The results of DT3 are recorded and presented in Fig. 6. Similar to the other datasets, the CNN model has the best classification efficiency among the transfer learning models in DT3 with an accuracy of 92.5% and an f1-score of 92.79%. Similar to the previous datasets, in DT3, the BrainTumorNet shows the highest efficiency. The classification accuracy of the model is 94.24%. The precision, recall, specificity, and f1scores are 96.34%, 95.06%, 94.59%, and 95.69%, respectively.

The performance matrix shows that the proposed fine-tuned model performs consistently with the highest accuracy over the three datasets. However, the classification accuracy of the three datasets varies. The datasets DT1 and DT2 achieve accuracy of 98.66% and 97.17%, respectively. These accuracies are quite similar to the accuracy achieved from DT3, which is 94.24%.

The datasets DT1 and DT2 have higher counts of data compared to DT3. When employed on DT3, the models have less data to train on, so their performance suffers. On the contrary, when the BrainTumorNet model is trained and validated on DT1, it achieves the highest performance efficiency.

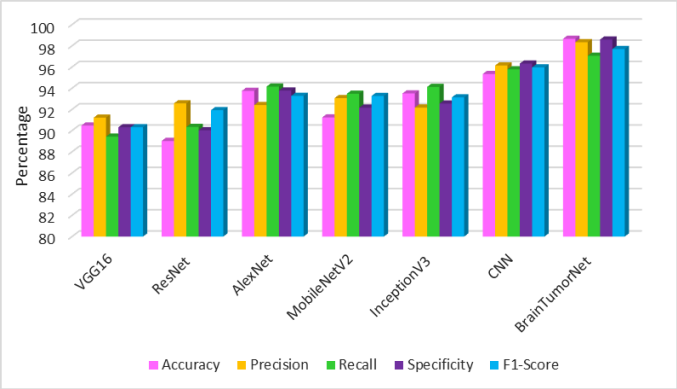


Fig. 4. Perfromance of the models for DT1.

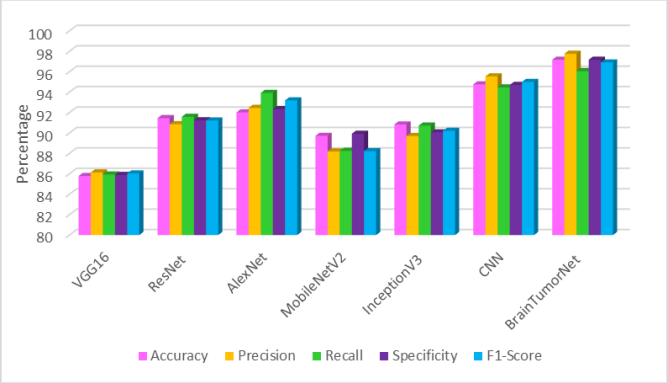


Fig. 5. Performance of the models for DT2.

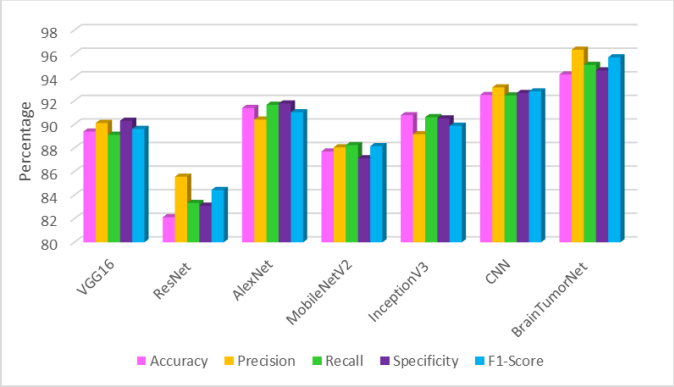


Fig. 6. Perfromance of the models for DT3.

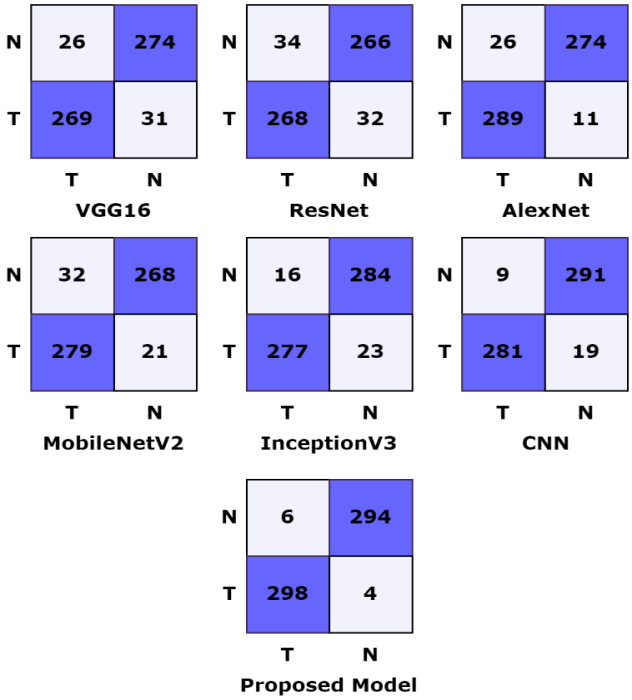


Fig. 7. Confusion Matrix of DT1 for the employed models.

In Fig. 7, we present the confusion matrix created by the seven models on the dataset DT1. In the confusion matrix, T refers to data with brain tumors, and N refers to data from normal brains. The confusion matrix was built from the validation set of DT1 consisting of 600 data. It can be observed from the confusion matrix that the proposed model BrainTumorNet correctly predicts the most data and has the lowest incorrect predictions.

In Fig. 8(a), the progress of the performance accuracy of the proposed model (BrainTumorNet) on DT1 is illustrated. The performance of the model on 250 epochs is presented during both training and validation of the model. It can be observed that in the 1st epoch, the model starts with a very low training accuracy of 42% and validation accuracy of 54%. However, with each progressing epoch, the model's accuracy increases rapidly. On the final epoch, the model provides the highest training accuracy of 97.83% and validation accuracy of 98.66%.

Likewise, in Fig. 8(b), the progress of the loss of the proposed model (BrainTumorNet) over 250 epochs is presented. On the 1st epoch, the model demonstrates a high training loss of 4.75% and a validation loss of 4.33%. Over the increased epoch, the loss rate gradually decreases with consistency. On the final epoch, the model achieves the lowest training loss of 0.207% and validation loss of 0.135%.

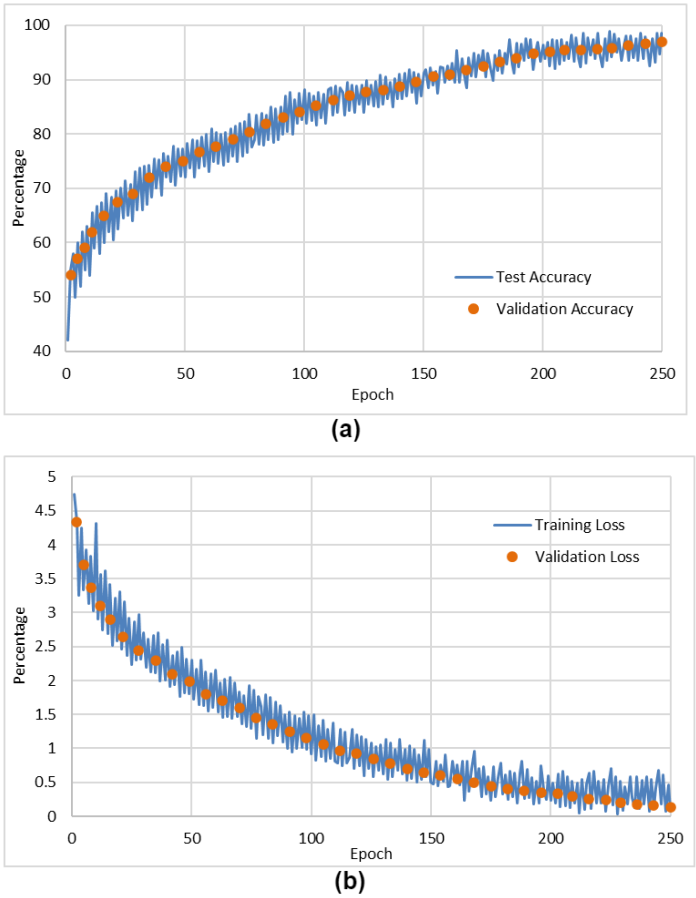


Fig. 8. (a) Accuracy and (b) Loss of proposed model on DT1 per epoch.

V. State-of-the-Art Comparison

This paper introduces BrainTumorNet, a complex convolutional neural network (CNN) specifically designed for classifying images of brain tumors, to address the crucial difficulty of reliably identifying brain tumors using Magnetic Resonance Imaging (MRI). Advanced preparation techniques were used to ensure the highest level of data quality, including CLAHE for image improvement, data augmentation for assuring dataset variety, and blockchain integration for secure and traceable data administration. By recording accuracy scores of 98.66%, 97.17%, and 94.24% across three datasets during testing, BrainTumorNet demonstrated its outstanding abilities and established a new standard when compared to other pre-trained models. This study raises the standard for MRI-based brain tumor identification and gives professionals a crucial diagnostic tool. Table VI thoroughly evaluates BrainTumorNet's performance in relation to other existing models.

TABLE VI. STATE-OF-THE-ART COMPARISON OF PROPOSED MODEL

Authors	Data Type	Methods	Accuracy
Khan et al. [13]	MRI	Fine-tune VGG19	90.03%
Yahyaoui et al. [14]	MRI	DenseNet	92.06%
Febrianto et al. [33]	MRI	CNN	93%
Afshar et al. [34]	MRI	CapsNet	90.89
Anaraki et al. [35]	MRI	Shallow CNN	94.20
Rehman et al. [36]	MRI	3D CNN	92.67%
Sajjad et al. [37]	MRI	VGG19	90.67%
Banik et al.	MRI	BrainTumorNet	98.66%
			97.17%
			94.24%.

VI. CONCLUSION

BrainTumorNet, a CNN model developed to detect brain tumors from MRI images, was demonstrated in this research article. Its performance was improved through a rigorous data preparation process that included CLAHE and data augmentation. Brain TumorNet's unique incorporation of blockchain technology assures MRI data management that is highly secure and identifiable, thereby developing confidence and facilitating collaborations focused on data integrity. Assessed through the utilization of a wide variety of metrics such as accuracy, specificity, recall, precision, f1-score, and confusion matrix, BrainTumorNet demonstrated its ability to perform by attaining accuracy rates of 98.66%, 97.17%, and 94.24% on three separate datasets. Furthermore, it outperformed six pre-trained deep learning models. Despite infrequent misclassifications, its overall efficacy represents a significant development in the field of medical imaging. It is believed that integrating deep learning with blockchain will bring about an important change in perspective in healthcare management and brain tumor detection in the future.

REFERENCES

[1] K. R. Bhatele and S. S. Bhadauria, ''Machine learning application in glioma classification: Review and comparison analysis,'' Arch. Comput. Methods Eng., vol. 29, pp. 247–274, Apr. 2021.

[2] M. Y. B. Murthy, A. Koteswararao, and M. S. Babu, ''Adaptive fuzzy deformable fusion and optimized CNN with ensemble classification for automated brain tumor diagnosis,'' Biomed. Eng. Lett., vol. 12, no. 1, pp. 37–58, Feb. 2022.

[3] R. Mehrotra, M. A. Ansari, R. Agrawal, and R. S. Anand, ''A transfer learning approach for AI-based classification of brain tumors,'' Mach. Learn. with Appl., vol. 2, Dec. 2020, Art. no. 100003.

[4] S. Grampurohit, V. Shalavadi, V. R. Dhotargavi, M. Kudari, and S. Jolad, ''Brain tumor detection using deep learning models,'' in Proc. IEEE India Council Int. Subsections Conf. (INDISCON), Oct. 2020, pp. 129–134.

[5] G. Raut, A. Raut, J. Bhagade, J. Bhagade, and S. Gavhane, ''Deep learning approach for brain tumor detection and segmentation,'' in Proc. Int. Conf. Converg. Digit. World Quo Vadis (ICCDW), Feb. 2020, pp. 1–5.

[6] T. C. Hollon, B. Pandian, A. R. Adapa, E. Urias, A. V. Save, S. S. S. Khalsa, D. G. Eichberg, R. S. D'Amico, Z. U. Farooq, S. Lewis, and P. D. Petridis, ''Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks,'' Nature Med., vol. 26, no. 1, pp. 52–58, Jan. 2020.

[7] Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., & Fan, Y. (2018). A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. Medical image analysis, 43, 98-111.

[8] Mohsen, H., El-Dahshan, E. S. A., El-Horbaty, E. S. M., & Salem, A. B. M. (2018). Classification using deep learning neural networks for brain tumors. Future Computing and Informatics Journal, 3(1), 68-71.

[9] Paul, J. S., Plassard, A. J., Landman, B. A., & Fabbri, D. (2017, March). Deep learning for brain tumor classification. In Medical Imaging 2017: Biomedical Applications in Molecular, Structural, and Functional Imaging (Vol. 10137, pp. 253-268). SPIE.

[10] Ari, A., & Hanbay, D. (2018). Deep learning based brain tumor classification and detection system. Turkish Journal of Electrical Engineering and Computer Sciences, 26(5), 2275-2286.

[11] Abiwinanda, N., Hanif, M., Hesaputra, S. T., Handayani, A., & Mengko, T. R. (2019). Brain tumor classification using convolutional neural network. In World congress on medical physics and biomedical engineering 2018 (pp. 183-189). Springer, Singapore.

[12] Afshar, P., Plataniotis, K. N., & Mohammadi, A. (2019, May). Capsule networks for brain tumor classification based on MRI images and coarse tumor boundaries. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1368-1372). IEEE.

[13] Khan, A. R., Khan, S., Harouni, M., Abbasi, R., Iqbal, S., & Mehmood, Z. (2021). Brain tumor segmentation using K - means clustering and deep learning with synthetic data augmentation for classification. Microscopy Research and Technique, 84(7), 1389-1399.

[14] Yahyaoui, H., Ghazouani, F., & Farah, I. R. (2021, July). Deep learning guided by an ontology for medical images classification using a multimodal fusion. In 2021 International Congress of Advanced Technology and Engineering (ICOTEN) (pp. 1-6). IEEE.

[15] Murthy, M. Y. B., Koteswararao, A., & Babu, M. S. (2022). Adaptive fuzzy deformable fusion and optimized CNN with ensemble classification for automated brain tumor diagnosis. Biomedical engineering letters, 12(1), 37-58.

[16] Latif, G., Ben Brahim, G., Iskandar, D. A., Bashar, A., & Alghazo, J. (2022). Glioma Tumors' classification using deep-neural-network-based features with SVM classifier. Diagnostics, 12(4), 1018.

[17] Kaggle. Avialble Online: https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection?select=Br35H-Mask-RCNN (Accessed on 1 August 2023).

[18] Kaggle. Avialble Online: https://www.kaggle.com/datasets/aryanfelix/brats-2019-traintestvalid?select=dataset (Accessed on 1 August 2023).

[19] Kaggle. Avialble Online: https://www.kaggle.com/datasets/mhantor/mri-based-brain-tumor-images (Accessed on 1 August 2023).

[20] Maymounkov P, Mazieres D. Kademlia: A peer-to-peer informa- ' tion system based on the XOR metric. In: Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). (2002). doi: 10.1007/3-540-45748-8_5.

[21] Rahman A, Islam MJ, Montieri A, Nasir MK, Reza MM, Band SS, et al. SmartBlock-SDN: an optimized blockchain-SDN framework for resource management in IoT. IEEE Access. (2021) 9:28361–76. doi: 10.1109/ACCESS.2021.3058244.

[22] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems; Curran Associates, Inc.: New York, NY, USA, 2012; pp. 1097–1105.

[23] Pérez-Pérez, B. D., Garcia Vazquez, J. P., & Salomón-Torres, R. (2021). Evaluation of convolutional neural networks' hyperparameters with transfer learning to determine sorting of ripe medjool dates. Agriculture, 11(2), 115.

[24] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

[25] Shamrat, F. J. M., Azam, S., Karim, A., Islam, R., Tasnim, Z., Ghosh, P., & De Boer, F. (2022). LungNet22: A Fine-Tuned Model for Multiclass Classification and Prediction of Lung Disease Using X-ray Images. Journal of personalized medicine, 12(5), 680.

[26] Zaccone, G.; Karim, M.R. Deep Learning with TensorFlow: Explore Neural Networks and Build Intelligent Systems with Python; Packt Publishing Ltd.: Birmingham, UK, 2018.

[27] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.

[28] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.

[29] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[30] Shamrat, F. J. M., Azam, S., Karim, A., Ahmed, K., Bui, F. M., & De Boer, F. (2023). High-precision multiclass classification of lung disease through customized MobileNetV2 from chest X-ray images. Computers in Biology and Medicine, 155, 106646.

[31] Shamrat, F. J. M., Akter, S., Azam, S., Karim, A., Ghosh, P., Tasnim, Z., & Ahmed, K. (2023). AlzheimerNet: An effective deep learning based proposition for alzheimer's disease stages classification from functional brain changes in magnetic resonance images. IEEE Access, 11, 16376-16395.

[32] Tasnim, Zarrin, et al. "Classification of breast cancer cell images using multiple convolution neural network architectures." International Journal of Advanced Computer Science and Applications 12.9 (2021).

[33] Febrianto, D. C., Soesanti, I., & Nugroho, H. A. (2020, March). Convolutional neural network for brain tumor detection. In IOP Conference Series: Materials Science and Engineering (Vol. 771, No. 1, p. 012031). IOP Publishing.

[34] Afshar, P., Plataniotis, K. N., & Mohammadi, A. (2019, May). Capsule networks for brain tumor classification based on MRI images and coarse tumor boundaries. In ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 1368-1372). IEEE.

[35] Anaraki, A. K., Ayati, M., & Kazemi, F. (2019). Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms. biocybernetics and biomedical engineering, 39(1), 63-74.

[36] Rehman, A., Khan, M. A., Saba, T., Mehmood, Z., Tariq, U., & Ayesha, N. (2021). Microscopic brain tumor detection and classification using 3D CNN and feature selection architecture. Microscopy Research and Technique, 84(1), 133-149.

[37] Sajjad, M., Khan, S., Muhammad, K., Wu, W., Ullah, A., & Baik, S. W. (2019). Multi-grade brain tumor classification using deep CNN with extensive data augmentation. Journal of computational science, 30, 174-182.

[38] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

Proposal of a Machine Learning-based Model to Optimize the Detection of Cyber-attacks in the Internet of Things

Cheikhane Seyed, Jeanne roux BILONG NGO, Mbaye KEBE
Laboratory LIRT- Polytechnic Superior School, University Cheikh Anta DIOP (UCAD), Dakar, 10200, Senegal

***Abstract*—In this article, we propose a model to optimize the detection of attacks in IoT. IoT network is a promising technology that connects living and non-living things around the world. Despite the increased development of these technologies, cyber-attacks remains a weakness, making it vulnerable to numerous cyber-attacks. Of course, automatic computer intrusion detection systems are deployed. However, it does not make it possible to mobilize the full potential of Machine Learning. Our approach in this maneuver consists of offering a means to select the least expensive ML method in terms of learning in order to optimize the prediction of threats to introduce IoT objects. To do this, we make modular design based on two layers. The first module is a canvas containing the different methods most used in ML such as supervised learning method, unsupervised learning method and reinforcement learning method. The second module introduces a mechanism to measure the learning cost linked to each of these methods in order to choose the least expensive one in order to quickly and efficiently detect intrusions in IoT objects. To prove the validity of the proposed model, we simulated it using the Weka tool. The results obtained illustrate the following behaviors: The classification quality rate is 93.66%. This last result is supported by a classification consistency rate of 0.882 (close to unity 1) demonstrating a trend towards convergence between observation and prediction.**

***Keywords*—IoT; Machine learning; cyber-security; detection of attacks; weka tool; classification quality and consistency**

I. INTRODUCTION

IoT refers to a network of smart objects around the world via the Internet, allowing them to collect and exchange data without any human interference [1]. However, security in the Internet of Things (IoT) is a major concern because it is susceptible to cyber-attacks like any other network given the proliferation of connected devices and the massive data collection they perform [2].

Intrusion Detection System (IDS) is an effective technique for detecting cyber-attacks in any network. IDS detects cyber-attacks efficiently and quickly at fog nodes compared to the cloud [3]. The IoT network consists of connections between different types of smart objects, ranging from supercomputers to tiny devices that may have very little computing power. It is therefore difficult to secure this type of network and cyber security is therefore a major challenge.

Faced with this new mode of operation of the strike chain, traditional network security is no longer suitable. Certainly, computer systems for automatic intrusion detection are deployed [4, 5]. Most of the latest IDS are based on machine learning algorithm for training and detection of cyber-attacks on the network. However, their conceptual deficiency does not make it possible to mobilize the full potential of Machine Learning.

The problem that arises is that these IDS do not emphasize the impact linked to the learning cost for the ML method used. This can slow down the attack prediction process by choosing an ML method that is inappropriate for the context and environment of IoT objects.

In this context, we propose an approach to boost the detection of attacks by choosing the optimal method in terms of learning cost to provide a prediction of attacks that is fast, efficient and reflects reality.

The contribution consists of proposing a new process framework for integrating ML techniques. This process is based on three pillars. The first is the dynamic dimension of the cyber-attacks chain. This problem is addressed by proposing an updatable dataset in terms of sampling and scoring of variables. The second is the competition process of different existing ML methods. The third is the introduction of cost-sensitive learning using a risk-based cost matrix.

The remainder of this article is organized as follows: Related work is given in Section II, Section III delves into Approach and method, then Section III deals with the proposed modeling, then Section V deals with the methods and materials. Section VI dedicated to the results and discussions and finally we present the conclusion and the perspectives of our work in Section VII.

II. RELATED WORK

Intrusion detection systems are built based on data collected and trained using supervised, semi-supervised and unsupervised learning methods [6]. This article proposes to evaluate the performance of intrusion detection systems over the long term. The objective is to be able to detect still unknown zero-day attacks.

On the other hand, a summary on the analysis of security threats, issues and solutions for Cloud computing uses machine learning algorithms [7]. They are used to overcome security issues of cloud computing in supervised, unsupervised, semi-supervised and hardened modes.

The Internet of Connected Things in the industrial domain (I-IoT) is also an active area of research and is the subject of several studies. The problem of low detection rates and high proportions of false alarms is addressed in this article [7]. The sole objective of this work is to detect and stop cyber-attacks. Concerns about the costs and impact of this detection are not the focus.

The article in [8] makes an important contribution to solving the problem of security of connected objects. An in-depth analysis of the literature is assigned to them. The articles cited in this study certainly differ in their aims and objectives. Some of them approach the question from the reasonable angle of the technical constraints intrinsic to the IoT, notably storage, memory and energy.

Other authors in [9, 10] introduce the notions of layered architecture with or without integration of techniques such as machine learning, artificial intelligence and cryptography. The contextualization of the security issue of connected objects remains reactive and corrective. However, the proposed solutions do not seem to be part of an innovative and proactive methodology.

In the article [11], the authors have access to a review of the typology of anomalies, detection layers, context and methodology. What emerges is an overly simplistic view of anomaly classification. All attacks are classified into a single anomaly category, resulting in only four anomaly types. This represents more than half of the population. In addition, the type of attack is not well specified. More than 90% of articles do not consider context. This further weakens the robustness of the proposed solutions.

The articles in [12,13] shows the need to focus on learning methods for Cyber security in IoT Networks, the quality of the data used and the importance of security issues in free decision-making. This last point is crucial with regard to the cognitive dimension of the proposed solution.

In the paper [14], the authors proposed a cyber-attack detection framework for IoT using the voting-based ensemble learning approach. This idea of ensemble learning like Random Forest (RF) is good, but the process does not include risk-based thinking although it achieves over 97% accuracy.

The authors in [15] proposed several approaches based on recurrent neural networks (RNN) using long short-term memory (LSTM), auto encoders and multi-layer perceptron. For the authors [16], deep learning LSTM is applied with the resampling of an imbalanced dataset. Here again, no risk prioritization or cost discrimination was applied. The overall accuracy rate of 99% did not resolve the impact of individual cyber-attacks.

In the work [17], it is recognized that the IoT cyberspace is like an “unsecured Internet of Things” and that emerging technologies (machine learning, block chain) are key solutions. This survey reveals that there are many problems associated with the use of machine learning techniques. Topics include dataset accuracy and versioning.

However, these articles do not seem to draw attention to the innovative approach of modular analysis and security segregation based on learning costs which prompted us to focus in this research study on the impact of learning cost of ML methods thus influencing the effectiveness of attack predication in IoT.

III. Approach and Methodology

The majority of detection systems are based on classical linear structure learning models (see Fig. 1). That is, we draw a dataset S from an operational IoT network. This data set and the underlying variables describe a particular state of a functional system. Critical security factors may constantly change over time horizon t, with respect to the dynamics of cyber threats. The resulting M-model becomes inappropriate for tracking cyber-attacks due to short-term mutations in the threat process called block chain.

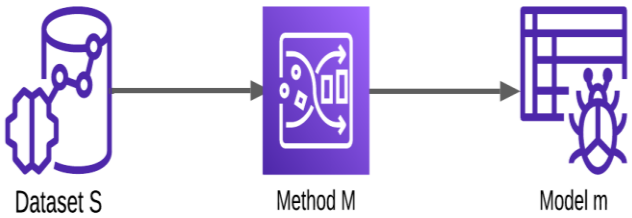


Fig. 1. Classic intrusion detection model.

Viable, reliable and agile machine learning that streamlines operations and strengthens businesses is a very serious and time-consuming task [18]. Indeed, this approach to cyber security modeling does not call into question the improvement in algorithmic performance delivered by a one-way learning method. As a result, we will not be able to capitalize on the differential advantage offered by a range of different learning methods. Additionally, it should be noted that omitting the impact of the learning process and attack detection would reduce the quality of the model. The reason is that cyber-attacks differ in terms of their impact on the entire company and in terms of the resources used to neutralize them. These conceptual cognitions constitute the very foundation of our motivation to propose a new type of approach to modeling cyber security issues. We seek to achieve essentially the following objectives.

On the one hand, the goal of our approach is to define what represents an acceptable algorithmic methodology. It allows the learning engine to have access to a set of machine learning methods to increase the algorithmic space. The reason is to match the most salient issues that need to be resolved. This will increase the quality of intrusion detection and the cyber security of connected objects. Additionally, the outcome of this expected performance depends on the raw data sampling process. The quality criteria for this sampling include not only the size of the data set but also the relationship between the descriptive variables called inter-correlation. This last parameter should ideally be reduced to zero. In addition, the delay in capturing network data has a great advantage. This will keep the dataset up to date and consistent with threat level and complexity.

On the other hand, the objective is to follow a risk-based approach when analyzing the cyber-security attack chain or kill chain. This means that cyber risk must be identified first. Then,

the assessment of this risk is based on its probability and its total impact on the system. The end goal of this process is to prioritize each risk category in terms of cost weighting. This will lead to a framework enabling the integration of security objectives, taking into account the risk tolerance level of security councils, and at the same time reduce the costs induced by security and insecurity related to the Internet of Things.

The ultimate goal of this approach is to develop an optimized model that recovers all the drawbacks linked to the state of the art of cyber-security and ensures enhanced cyber-security of connected objects. This model should compensate for the shortcomings of the approaches discussed in the state of the art.

IV. Proposed Modeling

The motivation for proposing a new intrusion detector optimization model lies in the fact that NIDS attempts to apply the same intrusion filters regardless of the risk policies in place. Since zero risk is unrealistic, it is essential to control its assessment and level of acceptance.

The ultimate goal of this approach is to develop a model allowing you to choose among Machine Learning methods (supervised, unsupervised and reinforcement), the optimal method to effectively detect intrusions in IoT objects. This choice is essentially based on the cost-sensitive technique thus minimizing the learning impact of attacks and intrusions in IoT objects.

Upgrading the generic classification and detection model involves redefining the methodology. To do this, we imagined the creation of two functional layers at the conceptual level (see Fig. 2). This brings us to modular programming of the detection engine. On the one hand, the first module serves to design the algorithmic component of the methodology in order to integrate a wide range of learning methods. On the other hand, the second module models the security-cost component of the methodology. This allows the security manager to control the acceptable level of risk in relation to the typology of cyber-attacks. In this way, the most optimal classification method will be chosen. This is the least expensive method in terms of negative impact.

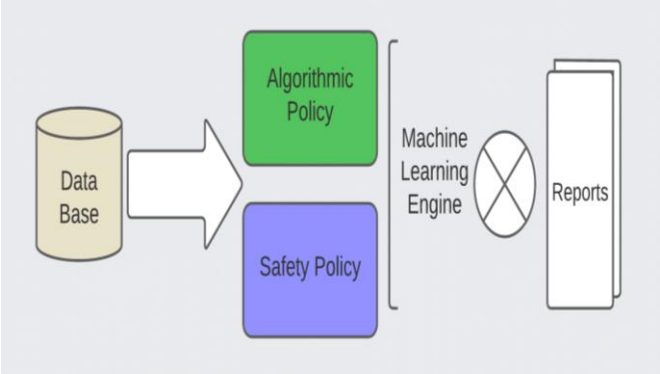


Fig. 2. Macro-learning optimization method for cyber-attacks.

We find ourselves in an optimization automation process with the possibility of acting on the algorithmic parameters and the security cost.

Achieving the previously set optimization goals will be a prerequisite for achieving the optimized cyber security logic model (see Fig. 3).

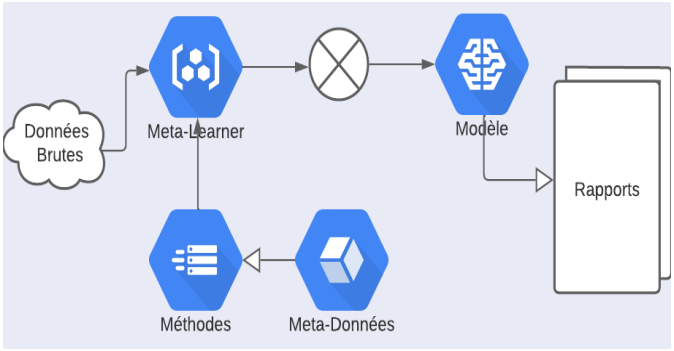


Fig. 3. Optimized cyber-security logic model (olm).

This process takes place in two successive phases. First, we introduce the concept of Meta-Learning on which algorithmic policy is based. Next, we present the cost-aware learning technique. This will make it possible to implement the entity's security policy in terms of intrusion detection and cost control.

Meta-learning corresponds to what we could call macro learning. This involves understanding the behavior of multiple learning methods. The objective is to collect metadata consisting of performance values and algorithmic parameters associated with the methods.

This approach makes it possible to nest or encompass several learning methods within a single Canvas. We know that the quality of an algorithm includes not only how well it predicts reality, but also how quickly it is executed. This is the basic principle of the meta-learning process.

In the second block dedicated to security policy, we propose to introduce the notion of learning costs (impacts). The goal of classical learning is to minimize the errors generated by the difference between prediction and observation. Since not all errors have the same cost or impact, we will use the cost-sensitive learning technique. The fundamental principle of cost-aware learning is that the learning engine is informed about the cost or impact of intrusion detection scenarios.

At the formal level of the description of the proposed model, we are faced with an operational research problem. On the one hand, it involves defining an objective function, which takes into account the resource constraint and aims to maintain a level of algorithmic performance and to reduce the expression of costs.

Let n be the bisquare dimension matrix, the confusion matrix M (Mij) and the cost matrix C (Cij). The objective function F should be the aggregation of all effects that trigger resource consumption. This includes the cost of detection training and the cost of missed detection of an intrusion. F will be the scalar product of M and C:

$$F = (M * C) = \sum_{\substack{i=1 \\ j=1}}^{n} \left((M_{i,j} * C_{i,j}) \right)$$

Since the first diagonal corresponds to well-classified items, their cost is identical and can be normalized to Cij=1:

$$F = \sum (\mathrm{M_{i,j}} * \mathrm{C_{i,j}}) + \sum (M_{i,i}) \ , \qquad i \neq j$$

The cost of well-classified elements related to correct detection is $\sum(\mathrm{M_{i,i}})$. This is simply a computational cost. Thus, the remaining quantity $\sum(\mathrm{M_{i,j}} * \mathrm{C_{i,j}})$ of F corresponds to the cost linked to the impact of cyber-security. The linear optimization of the objective function is obtained by:

$$min \sum (M_{i,j} * C_{i,j}), \qquad i \neq j$$

V. Methods and Materials

A. Dataset

To conduct the proposed work, we used the latest DDoS attack dataset CICIDS2018 [19]. Most DDoS attack datasets have many limitations, such as missing relevant data and redundancy, which are unreliable. The CICIDS2018 datasets contain up-to-date real-world working network like data. This dataset was collected for five consecutive days with many different cyber-attacks as well as normal data. This dataset contains the latest network data with and without attack, very close to real network data. This dataset is unbalanced, so we balanced it using a duplication method because it seriously affects the training of the deep learning method and therefore testing. This work is applied in an environment containing a 32-bit Intel Core-i5 processor with 8 GB of RAM in a Windows 10 environment.

B. Confusion Matrix

The confusion matrix is a predictive analysis tool in machine learning. It is also known as an error matrix, and is used to evaluate the performance of a machine-learning model based on classification, which aim to predict a categorical label for each input instance [20].

We can also say that the confusion matrix is a summer table of the number of correct and incorrect predictions produced by a classifier for binary classification tasks.

The matrix displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by the model on the test data (see Table I).

- True Positive (TP): when the actual value is Positive and predicted is positive.
- True Negative (TN): when the actual value is Negative and prediction is Negative.
- False Positive (FP): When the actual is negative but prediction is Positive. Also known as the Type 1 error
- False Negative (FN): When the actual is Positive but the prediction is Negative. Also known as the Type two errors.

TABLE I. Confusion Matrix Basic Metrics

Techniques	Observation	
Learning (Predicted Class)	(True Positives) TP	(False Positive) FP
	(False Negative) FN	(True Negative) TN

C. Performance Metrics

The performance of proposed deep learning models for the detection of DDoS attack is measured by standard matrices as Accuracy, Recall and Precision. The definition and the equation for the same is given below:

Accuracy: An indicator makes it possible to measure the proportion of well-classified individuals relative to the entire population examined. It is obtained using the following equation:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}$$

Precision: An indicator of false alarms. It allows us to answer the following question. What proportions of positive identifications were actually correct? It is obtained using the following equation:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Recall: A metric that characterizes detection failures. This failure results in the presence of false negatives. It is obtained using the following equation:

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

D. Weka Tool

The physical implementation of the IoT cyber-security model requires the use of hardware and software resources. In order to produce an Optimized Physical Cyber-security Model (OPCM), we opted for open source software that is well known in the scientific research community. This is Weka and its applications.

WEKA provides implementations of learning algorithms that you can easily apply to your database. It also includes a variety of tools for transforming datasets. These include algorithms for discretization and sampling. It can also be used to pre-process a dataset, integrate it into a learning scheme, and analyze the resulting classifier and its performance [21].

In this article, we use the tool for apply learning methods to a dataset and analyze its output to learn more about the data. The objective is to verify the performance of our cyber security mechanism in effectively predicting attacks hitting the IoT.

VI. Results and Discussions

The evaluation study of the proposed model is based on a set of tests obtained after simulating the model in the Weka tool. These data will be provided as input values to the prediction function. The results of this operation will be compared with the corresponding observation values.

Fig. 4. Prediction (star series) of cyberattacks using confusion matrix.

As these results show in Fig. 4, we notice that almost 2863 packets are denial of service (see DINAIL*). This prediction is confirmed by the observations of the test set. In addition there are 1928 packets recognized as benevolent (see Normal*) both in prediction and in observation. We find the errors made by the model for a column of observations outside the first diagonal. There were 1628 predictions ARP spoofing (see SPOOF*). The curves show that the prediction (star series) of cyber-attacks is very close to the values from observation.

To assess the validity of the model, several indicators can measure these objectives at the same time. We have the holistic statistical estimates, which evaluate the overall performance of the model. These indicators reflect the quality, shortcomings and consistency of the learning process.

Correctly Classified Instances	29306	93.6623 %
Incorrectly Classified Instances	1983	6.3377 %
Kappa statistic	0.882	
Mean absolute error	0.0254	
Root mean squared error	0.1592	
Relative absolute error	12.0258 %	
Root relative squared error	49.0447 %	
Total Number of Instances	31289	

Fig. 5. Performance statistics.

In Fig. 5, the results obtained illustrate the following behaviors: The classification quality rate is 93.66%. This rate shows a high level of conformity between predictions and observations. This result is supported by a classification consistency rate of 0.882 (close to unity 1), demonstrating a trend towards convergence between observation and prediction. This deduces the accuracy and performance of the model evaluated.

We then proceed to deepen our assessment of the model's validity using other metrics such as the model's sensitivity, specificity and precision.

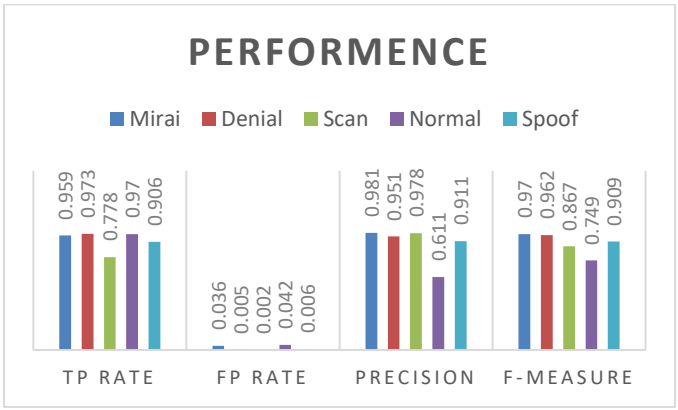


Fig. 6. Learning assessment metrics.

From these test results (see Fig. 6), we deduce that the sensitivity varies between 74% and 97%, while the specificity of the model is between 95.8% and 99.8%. The false alarm rate (false positives) is therefore between 0.2% and 4.2%.

VII. Conclusion

Our exploration has revealed major IoT security risks. The physical security of IoT is challenged in remote sites. Hardware and software upgrades and updates are critical. This is a major constraint to the scale of the threat.

This threat is accentuated by the availability of tools for researching and exploiting vulnerabilities in the IoT system. This represents obvious cyber security challenges.

In this design, special attention is paid to the cybernetic strike chain. To adapt to this, we opted for an optimized detection model. This optimization is based on algorithmic and security policies.

This integrates the potential of algorithmic methods and the reduction of learning costs. To implement it, macro-learning (Meta learning) and discriminated cost methods are used. The programming is carried out on the Weka of the machine-learning platform.

References

[1] B. Mazon-Olivo and A. Pan, (2022) "Internet of Things: State-of-the-art, Computing Paradigms and Reference Architectures," in IEEE Latin America Transactions, vol. 20, no. 1, pp. 49-63, Jan. 2022, doi: 10.1109/TLA.2022.9662173.

[2] K. Yang, Y. Zhang, X. Lin, Z. Li and L. Sun, (2022) "Characterizing Heterogeneous Internet of Things Devices at Internet Scale Using Semantic Extraction," in IEEE Internet of Things Journal, vol. 9, no. 7, pp. 5434-5446, 1 April1, 2022, doi: 10.1109/JIOT.2021.3110757.

[3] I. Dutt, S. Borah and I. K. Maitra, "Immune System Based Intrusion Detection System (IS-IDS): A Proposed Model," in IEEE Access, vol. 8, pp. 34929-34941, 2020, doi: 10.1109/ACCESS.2020.2973608.

[4] W. Iqbal, H. Abbas, M. Daneshmand, B. Rauf and Y. A. Bangash, (2020) "An In-Depth Analysis of IoT Security Requirements,

Challenges, and Their Countermeasures via Software-Defined Security," in IEEE Internet of Things Journal, vol. 7, no. 10, pp. 10250-10276, Oct. 2020, doi: 10.1109/JIOT.2020.299765.

[5] Z. Shi, S. He, J. Sun, T. Chen, J. Chen and H. Dong, (2023) "An Efficient Multi-Task Network for Pedestrian Intrusion Detection," in IEEE Transactions on Intelligent Vehicles, vol. 8, no. 1, pp. 649-660, Jan. 2023, doi: 10.1109/TIV.2022.3166911.

[6] S. Yao et al., "Deep Learning for the Internet of Things," in Computer, vol. 51, no. 5, pp. 32-41, May 2018, doi: 10.1109/MC.2018.2381131. T.-H. Chua et I. Salam (2022), « Evaluation of Machine Learning Algorithms in Network-Based Intrusion Detection System », p. 31 symmetry 2023, https://doi.org/10.3390/sym15061251.

[7] U. A. Butt et al., (2020) « A Review of Machine Learning Algorithms for Cloud Computing Security », Electronics, vol. 9, no 9, Art. no 9, sept. 2020, doi: 10.3390/electronics9091379.

[8] I. A. Khan, M. Keshk, D. Pi, N. Khan, Y. Hussain, et H. Soliman, (2022) « Enhancing IoT networks protection: A robust security model for attack detection in Internet Industrial Control Systems », Ad Hoc Networks, vol. 134, p. 102930, sept. 2022, doi: 10.1016/j.adhoc.2022.102930.

[9] P. Williams, I. K. Dutta, H. Daoud, et M. Bayoumi, (2022) « A survey on security in internet of things with a focus on the impact of emerging technologies », Internet of Things, vol. 19, p. 100564, août 2022, doi: 10.1016/j.iot.2022.100564.

[10] A. Kumar and A. Bansal, (2019) "Software Fault Proneness Prediction Using Genetic Based Machine Learning Techniques," 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), Ghaziabad, India, 2019, pp. 1-5, doi: 10.1109/IoT-SIU.2019.8777494.

[11] F. H. Sifat, R. Mahzabin, S. Anjum, A. -A. Nayan and M. G. Kibria, (2022) "IoT and Machine Learning-Based Hypoglycemia Detection System," 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET), Chittagong, Bangladesh, 2022, pp. 222-226, doi: 10.1109/ICISET54810.2022.9775890.

[12] M. R. Diana, P. Tobón, D. Múnera (2023) « Anomaly classification in industrial Internet of things: A review », In Intelligent Systems with Applications, vol. 18, p. 200232, mai 2023, doi: 10.1016/j.iswa.2023.200232.

[13] Naji, M., Zougagh, H. (2023). Deep Learning Models for Cybersecurity in IoT Networks. In: El Ayachi, R., Fakir, M., Baslam, M. (eds) Business Intelligence. CBI 2023. Lecture Notes in Business Information Processing, vol 484 . Springer, Cham. https://doi.org/10.1007/978-3-031-37872-0_3.

[14] A. Oseni et al., "An Explainable Deep Learning Framework for Resilient Intrusion Detection in IoT-Enabled Transportation Networks," in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 1, pp. 1000-1014, Jan. 2023, doi: 10.1109/TITS.2022.3188671.

[15] G. S. Mahara and S. Gangele, "Fake news detection: A RNN-LSTM, Bi-LSTM based deep learning approach," 2022 IEEE 1st International Conference on Data, Decision and Systems (ICDDS), Bangalore, India, 2022, pp. 01-06, doi: 10.1109/ICDDS56399.2022.10037403.

[16] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," IEEE transactions on neural networks and learning systems, vol. 28, pp. 2222-2232, 2017.

[17] N. Ahmad, R. P. George, R. Jahan and S. Hussain, "Integrated IoT and Block chain for secured access and managing education data," 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, 2022, pp. 1201-1204, doi: 10.1109/ICICICT54557.2022.9917643.

[18] A. Abeshu and N. Chilamkurti, "Deep learning: the frontier for distributed attack detection in Fog-to-Things computing," IEEE Communications Magazine, vol. 56, pp. 169- 175, 2018.

[19] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in ICISSP, 2018, pp. 108- 116.

[20] J. L. Garcia-Balboa, M. V. Alba-Fernandez, F. J. Ariza-López and J. Rodriguez-Avi, "Homogeneity Test for Confusion Matrices: A Method and an Example," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 2018, pp. 1203-1205, doi: 10.1109/IGARSS.2018.8517924.

[21] Ian H. Witten, Mark A. Hall, et Eibe Frank, « The WEKA workbench ». Consulté le: 20 Octobre 2023. [En ligne]. Disponible sur: https://waikato.github.io/weka-wiki/documentation.

Construction of an Intelligent Evaluation Model of Yield Risk Based on Empirical Probability Distribution

Zhou Yanru[1], Yang Jing[2]
Chaohu University, School of Mathematics and Big Data, Chaohu 238000, China[1]
Hefei Technology College, School of Architectural Engineering, Chaohu 238000, China[2]

***Abstract*—In order to improve the accuracy of yield risk evaluation, an intelligent evaluation model of yield risk based on empirical probability distribution is constructed. The dimensionality reduction method of risk factor based on principal component analysis is adopted. After adjusting the multiple data dimensions of risk factors that affect the rate of return to a unified dimension, the cluster-based evaluation index screening method is used to build the evaluation index set that best reflects the risk of the rate of return; The index weight vector equation method based on entropy weight and information entropy is used to set the evaluation index weight. Through the comprehensive evaluation model based on the empirical probability distribution of risk indicators, the empirical probability distribution information of risk indicators at all levels is analyzed, and the risk level of yield is intelligently evaluated. The research structure shows that the model can effectively evaluate the level of return risk and provide an effective reference for preventing and controlling investment return risk.**

Keywords*—Empirical probability distribution; yield; risk intelligence evaluation; principal component analysis; clustering; weight

I. INTRODUCTION

The characteristics of the investment itself, the complexity of the market environment and the risk management ability of investors will greatly affect the project's return on investment [1]–[3]. In order to obtain the highest return, investors should evaluate the investment risk of the product in the early stage. Because the risk always accompanies the return, and the high return must be accompanied by the high risk, so once the risk occurs, it may give investors a devastating blow [4], [5] The greater the risk of the activity is, the greater the loss of the final result if the decision is wrong, and vice versa. Risk cannot be completely avoided, but rational choice can minimize risk [6]. When conducting risk evaluation, investors should consider the source and use of funds. When examining the use of funds, that is, the investment of projects, they should also comprehensively consider the risks of financing and the overall market environment [7].

According to the analysis of the existing risk evaluation models, Authors used the NPV analysis model and @ risk software to assess the economic benefits and risks of China's carbon capture, utilization and storage projects in the context of carbon neutrality. Although the evaluation effect is effective, it is limited by the completeness of software functions. If the software is abnormal, whether the evaluation accuracy meets the standards remains to be tested [8]. Researchers built a risk evaluation model for overseas mining investment based on the structural power theory. Firstly, they built an evaluation index system of mining investment risk with the safety structure, production structure, financial structure and knowledge structure as the criterion level; then, the Topsis method and grey correlation analysis method were used to build a grey correlation risk evaluation model to complete the effective evaluation of overseas mining investment risk. However, this model does not analyze the problem of data dimension and indicator overlap, and the evaluation ability needs to be optimized [9]. Authors used the case analysis method to identify the investment risk of overseas railway construction projects and built a risk index system. Relevant methods establish the risk assessment model of overseas railway construction project investment. But the evaluation accuracy of this model is limited by the training effect of the neural network [10].

Although the above methods have made some progress, the accuracy of rate of return risk evaluation is low, and there are problems in evaluating the risk level of rate of return. Therefore this study focuses on the intelligent evaluation of yield risk. After reducing the dimension of risk factors, the clustering algorithm is introduced to cluster risk factors and build the optimal yield risk index system. On the basis of determining the weight of each index, empirical probability distribution theory is introduced to build an intelligent evaluation model of yield risk based on empirical probability distribution. In this model, the dimension reduction method of risk factors based on principal component analysis is adopted. After adjusting the data dimensions of various risk factors that affect the rate of return to a unified dimension, the evaluation index set that best reflects the rate of return risk is constructed by the evaluation index screening method based on clustering. The index weight vector formulation method based on entropy weight and information entropy is used to set the evaluation index weight. Through the comprehensive evaluation model based on the empirical probability distribution of risk indicators, the empirical probability distribution information of risk indicators at all levels is analyzed, and the risk level of return rate is intelligently evaluated. After the experimental test, the conclusions are as follows: (1) After reducing the dimension of the data of the influencing factors of yield risk,

the data dimension is obviously controlled in a unified range, which ensures the regularity of the data; The test results of cross-correlation coefficient of yield risk assessment results show that the cross-correlation coefficient is 1, and there is a significant correlation between the yield risk assessment results and the actual yield risk level; Compared with other models, the evaluation effect is the best, which can accurately evaluate the risk level of return, and the evaluation results are in line with the reality.

II. Intelligent Evaluation Model of Yield Risk

According to the classification of systematic risk and non-systematic risk, the classification of risk types of return on investment is shown in Fig. 1.

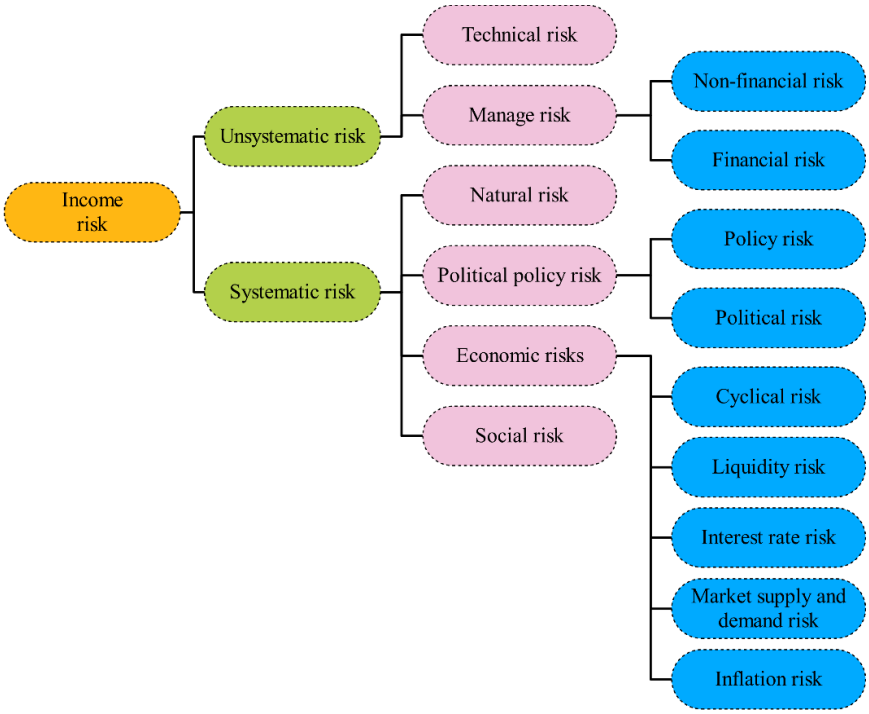


Fig. 1. Classification of risk categories of return on investment.

A. The Dimensionality Reduction Method of Risk Factor Based on Principal Component Analysis

According to the risk types described in Section II, many risk factors affect the rate of return. These risk factors are highly related to economic data, and the data dimensions are inevitably different [11], [12]. If such factors are directly used in the rate of return risk evaluation, it will increase the number of evaluation tasks. As a statistical analysis method for data dimensionality reduction, the principal component analysis method can transform multiple dimensions of the rate of return risk factors into a unified dimension on the premise of retaining most of the original information so as to improve the efficiency of data analysis [13], [14].

Principal component analysis (PCA) is a statistical analysis method with the main method of reducing data dimensions. On the premise of losing little original information, it transforms multiple influencing factors into several comprehensive factors (principal components) by calculating covariance, explains the internal structure of multiple influencing factors through a few principal components, catches the main contradictions and reduces the number of variables, and achieves the purpose of data compression and improving the efficiency of analysis [15], [16]. The main steps of the dimensionality reduction method for risk factors based on principal component analysis are as follows:

1) Standardize the data of yield risk factors. The risk factor data of the original yield is standardized to eliminate the impact of the yield risk factor data dimension and order of magnitude [17]. The standardization equation is:

$$a'_{ij} = \frac{a_{ij}-\beta_j}{R_j} \quad (1)$$

In the formula, a'_{ij} is the risk factor data of the standardized rate of return; a_{ij} is the original factor data; β_j and R_j are the sample mean and standard deviation of the jth yield risk factor; i=1,2,3,...,n, j=1,2,3,...,m, and are the number of samples and the number of factors.

2) Calculate the correlation coefficient matrix of standardized factor data [18]. The correlation coefficient matrix $S = (s_{ij})_{m\times m}$ is the m -order symmetric matrix, and the correlation coefficient s_{ij} represents the degree of correlation between the i -th factor and the j -th factor. The calculation equation of s_{ij} is:

$$s_{ij} = \frac{\sum_{h=1}^{n}(a_{hi}-\beta_i)(a_{hj}-\beta_j)}{\sqrt{\sum_{h=1}^{n}(a_{hi}-\beta_i)^2}\sqrt{\sum_{h=1}^{n}(a_{hj}-\beta_j)^2}} \quad (2)$$

In the formula, a_{hi} and a_{hj} are the h standardized data of the i -th and j -th risk factors respectively. β_i is the mean value of the second risk factor sample.

3) For the characteristic vector v_i of the yield risk factor, $\sum_{j=1}^{m} v_{ij}^2 = 1$ is required, where v_{ij} represents the j -th component of the characteristic vector v_i[19], [20].

4) Calculate the variance contribution rate and determine the principal component. Variance contribution rate F_i represents the proportion of variance of principal component G_i in the total variance, and its calculation equation is:

$$G_i = \frac{\varepsilon_i}{\sum_{i=1}^{m} \varepsilon_i} \times S_{ij} \tag{3}$$

Generally, the first h principal components of the cumulative contribution rate $\sum_{i=1}^{h} G_i \geq 85\%$ can contain most of the original factor information [21].

5) Calculate the principal component load (principal component coefficient matrix). The relationship between the principal component load value z_{ij} and the eigenvector v_i is:

$$\Omega_{ij} = \frac{v_{ij}}{\sqrt{\varepsilon_i}} \times G_i \tag{4}$$

6) Determine the dimension of risk evaluation factors of return rate [22]. Principal component load value Ω_{ij} and variance contribution rate F_i of the principal component jointly determine the final dimension of yield rate risk assessment factors. Then:

$$\Psi_j = \frac{\sum_{i=1}^{h} |\Omega_{ij}| G_i}{\sum_{j=1}^{h} \sum_{i=1}^{h} |\Omega_{ij}| G_i} \tag{5}$$

In the formula, Ψ_j is the dimension of the j -th yield risk factor.

B. Cluster-based Evaluation Index Screening Method

There is usually "coverage" and "overlap" of information between evaluation indicators. In order to obtain more accurate and objective evaluation results, this paper first determines and eliminates the indicators whose information is covered and then adjusts the weight of evaluation indicators according to the amount of information overlap.

Determining the evaluation index set usually includes two stages: rough selection and simplification of the evaluation index. The rough selection indicator set mainly considers the comprehensiveness of the evaluation indicators and is determined by the experts in the field and the evaluators through consultation; the selected indicator set mainly considers the representativeness of the evaluation indicators, which can be determined by statistical analysis methods such as correlation analysis based on the evaluation indicator value. There is no relevant quantitative method for the rough selection of evaluation indicators. Here, only a few main principles to be followed are given:

1) Purpose principle. To select evaluation indicators, it must first clarify the purpose of the evaluation. The evaluation indicators concerned vary with different purposes.

2) The principle of comprehensiveness. This principle needs to be followed in the rough selection stage of evaluation to ensure that the information contained in the evaluation index set reflects the effectiveness of weapons and equipment as fully and comprehensively as possible. As a result, there is also a phenomenon of coverage and overlap between evaluation indicators.

3) Principle of independence. This is a slightly conflicting principle with the principle of comprehensiveness. In the process of rough selection of evaluation indicators, it is necessary to compromise with the principle of comprehensiveness. It is to avoid the coverage and overlap of evaluation indicators as much as possible and ensure that certain characteristics of weapon equipment effectiveness will not be repeatedly reflected in multiple evaluation indicators.

4) Feasibility principle. The selected evaluation index must have a clear meaning, which can not only be understood and recognized by most people but also determines the evaluation index value based on sufficient and reliable data.

In this section, in the risk factor set after dimension reduction in Section II(A), the "very close" risk indicator factors can be determined through cluster analysis. At this time, these indicators can be considered to cover each other. Here, the distance between the yields risk evaluation indicators A_i and A_j is given as follows:

$$D_{ij} = 1 - \frac{\sum_{h=1}^{m} \left(A_i^h - \bar{A}_i\right)\left(A_j^h - \bar{A}_j\right)}{\sqrt{\left[\sum_{h=1}^{m} \left(A_i^h - \bar{A}_i\right)\left(A_j^h - \bar{A}_j\right)\right]^2}} \tag{6}$$

In the formula, D_{ij} represents the distance between indicators A_i and A_j; $\bar{A}_i = \sum_{h=1}^{m} A_i^h / m$, $\bar{A}_j = \sum_{h=1}^{m} A_j^h / m$; m is the number of indicators.

It is assumed that the risk evaluation index of return rate, class O_N, is obtained by combining class O_H and class O_Z, and the distance between it and class O_I is defined by the middle distance method as follows:

$$d_{nI} = \sqrt{\frac{1}{2} d_{HI}^2 + \frac{1}{2} d_{ZI}^2 - \frac{1}{4} d_{HZ}^2} \times D_{ij} \tag{7}$$

where, d_{nI} represents the distance between the class O_N and O_I; d_{HZ} represents the distance between classes O_H and O_I; d_{ZI} represents the distance between classes O_Z and O_I; d_{HI} represents the distance between classes O_H and O_Z.

If O_N and O_I only contain one yield risk evaluation index A_i and A_j, there are:

$$d_{nI} = D_{ij} \tag{8}$$

Further, the evaluation index screening method based on clustering is as follows:

1) Set the distance threshold σ between the evaluation indicators, and treat the yield risk evaluation indicator $A_1, A_2, \ldots, A_m$ as m different categories;

2) Calculate the distance between classes according to Eq. (6) to Eq. (8);

3) Determine whether the minimum distance $d_{nI} = d_{min}$ is less than σ;

4) If yes, merge classes O_H and O_Z into a new class O_N, and skip to step (2), otherwise execute step (5);

5) Output evaluation index class O_N.

Assume that O_N contains the set of yield risk evaluation indicators $\{l_1, l_2, \ldots, l_z\} \subseteq \{A_1, A_2, \ldots, A_m\}$, and z is the number of yield risk evaluation indicators included in O_N. At this point, it can be considered that $\{l_1, l_2, \ldots, l_z\}$ has basically covered each other, and it is further necessary to select the most representative yield risk evaluation index from them. Since the complex correlation number S_i reflects the degree of correlation between l_i and other indicators, the larger the value is, the higher the degree of coverage between l_i and other indicators is; that is, the more representative $l_i$is. The complex correlation coefficient between l_i and other indicators are as follows:

$$S_i = \sqrt{\frac{1}{z-1}\left(\sum_{j=1}^{1} s_{ij}^2 - 1\right)}, i \in z \tag{9}$$

where, s_{ij} is the correlation coefficient between indicators.

Further, the complex correlation coefficient corresponding to the representative index l_s of O_N is $S_s = \underset{1 \le i \le z}{max} S_i$.

After the reduction of yield risk indicators, the set of yield risk evaluation indicators obtained is $\{\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_u\} \in \{A_1, A_2, \ldots, A_m\}$, u is the number of yield risk evaluation indicators after the reduction, and the set of indicators $\{\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_u\}$ is the evaluation indicator system that best reflects the yield risk after the reduction.

C. Index Weight Vector Equation Method based on Entropy Weight and Information Entropy

1) Equation ion of weight vector in index layer considering data entropy weight: Entropy weight is an indicator to measure the degree of information provided by indicator data. By evaluating the degree of variation of data, we can measure the impact of the indicator data on the final rate of return risk evaluation results [23].

Firstly, the following steps are adopted to equation the weight vector between indicators within each indicator layer:

a) Through the expert scoring method, it can get the importance matrix $B(t) = \left(b_{ij}\right)_{m \times m}$ between the indicators in the k -th index layer of $\{\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_u\}$, where $k = 1,2, \ldots, u$, b_{ij} is the importance between the i -th index and the j -th index. The maximum eigenvalue and eigenvector of $B(t)$ are calculated as shown in Eq. (10).

$$\gamma(k) \frac{1}{\hat{A}_u} \sum_{1}^{\hat{A}_u} \frac{(B(t) \cdot \varpi(k))}{b_i}_{max} \tag{10}$$

In the formula, $\varpi(k) = (\varpi_1, \varpi_2, \ldots, \varpi_u)^T$ is the approximate eigenvector of the k -th index layer, which is the initial weight. A consistency check is performed on $\gamma(k)_{max}$, and if it is satisfied, it can proceed to the next step; otherwise, return to step (1) to re-evaluate $B(t)$ [24].

b) The information entropy evaluation is carried out for the jth index of the k -th index layer in $\{\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_u\}$. The evaluation method is as follows (11):

$$f(k,j) = \frac{-\sum_{z=1}^{Q(k,j)} q(k,j,z) \ln\left(2, q(k,j,z)\right)}{ln\left(Q(k,j)\right)} \tag{11}$$

In the formula, $f(k,j)$ is the information entropy of the j -th index in the k -th index layer; $Q(k,j)$ is the number of factor data of the j -th index of the k -th index layer; $q(k,j,z)$ is the satisfaction index of the z -th index factor data in the k -th index layer.

c) The improvement of $\varpi(k)$ taking into account information entropy is shown in Eq. (12):

$$\bar{\varpi}(k) = \varpi(k)^{T} f(k) \tag{12}$$

In the formula, $\bar{\varpi}(k)$ is the weight vector of the k -th index layer taking into account the information entropy; $f(k)$ is the information entropy vector of the k -th index layer; It can normalize the above equation:

$$\bar{\varpi}(k,j) = \frac{\bar{\varpi}(k,j)}{\sum_{j=1}^{m} \bar{\varpi}(k,j)} \tag{13}$$

In the formula, $\bar{\varpi}(k,j)$ is the weight of the j -th index in the k -th index layer after normalization.

2) Formulation of weight vector of criterion layer considering weight information entropy

However, the impact of different criteria levels on the final rate of return risk evaluation results is different, which cannot be reflected by the traditional expert scoring method and needs to be improved by using weight information entropy [25].

The formulation steps of the weight vector of the criterion layer considering the weight information entropy are as follows:

a) Based on the expert scoring method and the feature vector method [26], the weight vector between the criteria layers is obtained as $\varpi_{m \times m}$, and the specific method is the same as the weight vector between the indicators within the specified criteria layers.

b) The entropy weight is reflected in the criterion layer as follows: the weight information entropy of different evaluation indicators in the criterion layer, and the weight information entropy of the k -th criterion layer is calculated as shown in Eq. (14):

$$f'(k) = \frac{-\sum_{j=1}^{m(k)} \bar{\varpi}(k,j) \ln\left(2, \bar{\varpi}(k,j)\right)}{ln\left(m(k)\right)} \tag{14}$$

where, $f^{'}(k)$ is the weight information entropy of the k -th criterion layer; $m(k)$ is the number of indicators in the k -th criterion layer.

c) The weight vector $\varpi_{m \times m}$ between the criteria layers is improved by using the weight information entropy, as shown in equation (15).

$$\bar{\varpi}(k) = \varpi(k)^{T} f'(k) \tag{15}$$

where, $\bar{\varpi}(k)$ is the improved weight vector of the criterion layer. To sum up, the k -th criterion layer can be obtained, and the comprehensive weight of the j -th index is shown in Eq. (16):

$$\varpi(k,j) = \bar{\varpi}(k)f'(k) \quad (16)$$

D. Comprehensive Evaluation Model Based on the Empirical Probability Distribution of Risk Indicators

In order to comprehensively evaluate the return risk of each proposed investment project, we adopted an expert scoring method based on questionnaire. Considering that the rate of return risk index system consists of multiple groups of indicators, and each group of indicators contains different numbers of factors, a questionnaire scoring table is tailored for each project.

The design of the questionnaire carefully considered various risk indicators to ensure that quantitative and qualitative data were covered, so that experts could comprehensively and accurately assess risks. Specially invited 20 experts from the industry to participate in the grading. These experts have profound academic background and rich practical experience in related fields, and their grading has high authority and reference value. In the questionnaire, experts need to tick the corresponding level according to the data collected by the index system and their cognition of the project risk level. This process ensures that every expert can score according to a unified standard, thus increasing the objectivity and accuracy of scoring.

After this process, 20 answers from different experts will be obtained for each project. These answers provide valuable data to construct the empirical probability distribution of risk indicators. Through statistical analysis, we can understand the distribution of various risk indicators and the overall risk assessment of the project by experts. In order to ensure the validity and reliability of the questionnaire, a strict validity test is carried out. The content validity test is used to ensure that the questions and options in the questionnaire can fully reflect the rate of return risk index system. Secondly, the structural validity test is carried out, and the factor structure of the questionnaire data is analyzed to ensure that the measured results of the questionnaire are consistent with the expected risk factor structure.

1) Calculate the empirical probability distribution of risk indicators at each level:

$$q_{1ij}(Y_h) = \frac{1}{20}\varpi(k,j) \quad (17)$$

In the formula, $q_{1ij}(Y_h)$ refers to the probability distribution column of the risk indicators at the criterion level among the risk indicators at each level; Y_h is the risk level. Similarly, there are similar expressions for risk indicators at the indicator level, namely:

$$q_{2ij}(Y_h) = \frac{1}{20}\bar{\varpi}(k,j) \quad (18)$$

In the formula, $q_{2ij}(Y_h)$ represents the probability distribution column of the risk indicators at the indicator level among the risk indicators at each level.

Among the risk indicators at all levels, the empirical probability distribution of the j -th risk factor of the i -th risk indicator at the criterion level is:

$$G_{1ij}(y) = \sum_{h=1}^{5} q_{1ij}(Y_h) \quad (19)$$

At this time, the total empirical probability distribution of risk indicators at the indicator level can be expressed as:

$$G_{1i}(y) = \sum_{j=1}^{\bar{\varpi}(k,j)} G_{1ij}(y) \quad (21)$$

In the formula, $\bar{\varpi}(k,j)$ is the comprehensive weight of risk indicators at the indicator level.

2) Based on the empirical probability distribution of risk indicators at all levels, the risk level of the rate of return is evaluated:

$$Q^{(1)} = \left(q_{ij}(Y_1), q_{ij}(Y_2), q_{ij}(Y_3), q_{ij}(Y_4), q_{ij}(Y_5)\right) \quad (22)$$

In the formula, $q_{ij}(Y_1)$ is the empirical probability distribution column data of indicators in risk level 1. The total value of empirical probability distribution is 1. It is mainly used to judge the degree of risk level based on the proportion of empirical probability distribution of risk indicators in each risk level. In the issue of income risk evaluation, the risk level is mainly divided into five levels, namely, lower risk, low risk, medium risk, high risk and higher risk.

III. Experimental Analysis

In order to analyze the use effect of the model in this paper, the risk level of the return rate of two enterprises invested by a private equity fund is evaluated. A comprehensive data set is needed to train and verify the model, and the data set of yield risk is selected as the experimental data set, which is collected from publicly available financial databases, including historical yield data of financial markets such as stocks, bonds and futures. Collect GDP growth rate, inflation rate, interest rate, etc. from economic databases or official institutions, and collect income, profits, assets, etc. from company financial reports or public databases.

Data preprocessing is an important step to build the model. Firstly, data cleaning is carried out to remove the repeated, abnormal and wrong data points in the data set to ensure the accuracy and reliability of the data. Secondly, in order to eliminate the influence of dimension and range in feature data, feature normalization is carried out, so that different features can be compared and calculated fairly. Finally, the missing values are filled by interpolation or regression to ensure the integrity and continuity of the data. These pretreatment measures can improve the efficiency and accuracy of model training, and provide a solid data foundation for building an intelligent evaluation model of return risk based on empirical probability distribution. Table I is the information on yield risk indicators constructed by this model.

Table II shows the setting details of the weight of the yield risk evaluation indicators in this model:

The model in this paper analyzes the investment risk of two enterprise projects using the risk evaluation of the rate of return. Table III is the empirical probability distribution column of risk indicators of Enterprise 1 and Enterprise 2.

TABLE I. INFORMATION ON YIELD RISK INDICATORS

Criterion layer	Indicator layer	Factor
Financial risk	Solvency	Current ratio
		Quick ratio
		Asset-liability ratio
		Liabilities/EBIT
	Profitability	Net interest rate of equity
		Profit margin of main business
		Net asset interest rate
	Operational capacity	Total asset turnover
		Inventory cycle rate
		Accounts receivable turnover rate
	Cash payment ability	Free cash flow
		Cash flow debt ratio
		Sales cash ratio
	Growth ability	Sales growth rate
		Profit growth rate in the past two years
Non-financial risk	Industry risk	Political and economic fluctuation risk
		Policy and regulatory risks
		Industry life cycle risk
	Market risk	Sales risk
		Supply chain risk
		Market competition risk
		Market development risk
	Product risk	Property right risk
		Product substitution risk
		Technical environmental risk
		Product technical risk
		Product economic risk

TABLE II. WEIGHT SETTING DETAILS OF YIELD RISK EVALUATION INDICATORS

Criterion layer	Indicator layer	Weight	Factor	Weight
Financial risk	Solvency	0.140	Current ratio	0.273
			Quick ratio	0.296
			Asset-liability ratio	0.176
			Liabilities/EBIT	0.255
	Profitability	0.312	Net interest rate of equity	0.487
			Profit margin of main business	0.315
			Net asset interest rate	0.198
	Operational capacity	0.179	Total asset turnover	0.312
			Inventory cycle rate	0.302
			Accounts receivable turnover rate	0.386
	Cash payment ability	0.114	Free cash flow	0.4
			Cash flow debt ratio	0.302
			Sales cash ratio	0.298
	Growth ability	0.255	Sales growth rate	0.491
			Profit growth rate in the past two years	0.509
Non-financial risk	Industry risk	0.258	Political and economic fluctuation risk	0.322
			Policy and regulatory risks	0.24
			Industry life cycle risk	0.438
	Market risk	0.271	Sales risk	0.288
			Supply chain risk	0.227
			Market competition risk	0.251
			Market development risk	0.234
	Product risk	0.471	Property right risk	0.204
			Product substitution risk	0.194
			Technical environmental risk	0.154
			Product technical risk	0.253
			Product economic risk	0.195

TABLE III. THE EMPIRICAL PROBABILITY DISTRIBUTION OF RISK INDICATORS OF TWO ENTERPRISE PROJECTS

Risk level	Item 1	Item 2
1	0.450	0.173
2	0.347	0.290
3	0.151	0.263
4	0.052	0.143
5	0.000	0.131

As shown in Table III, the empirical probability distribution column value of Project 1 accounts for the largest proportion in risk level 1, followed by risk level 2, and 0.00 in risk level 1, while the empirical probability distribution column value of Project 2 accounts for the largest proportion in risk level 2, followed by risk level 3. There are empirical probability distribution columns in five risk levels, indicating that the risk is greater than Project 1.

Fig. 2 and Fig. 3 show the details of the data distribution dimensions before and after the dimensionality reduction of the data of the factors influencing the yield risk.

As shown in Fig. 2 and Fig. 3, the data dimension of the model in this paper is significantly different before the dimensionality reduction of the data of the factors influencing the yield risk. If such data is directly used in the intelligent risk evaluation, it will increase the difficulty. However, after reducing the dimension of the data of the factors influencing the yield risk in this model, the data dimension is obviously controlled in a unified range, ensuring the regularity of the data.

Whether the risk evaluation results are credible when the model is used to assess the project yield risk of two enterprises is tested, and the cross-correlation coefficient reflects the test results. The analysis method of correlation number Γ_{ij} is:

$$\Gamma_{ij} = \frac{cov\left(Q_i^{(1)}, Q_j^{(1)}\right)}{Q_i^{(1)} Q_j^{(1)}} \tag{22}$$

In the formula, $Q_i^{(1)}$ and $Q_j^{(1)}$ are the standard deviation between the risk level of the i -th and j -th project rate of return and the risk level of the actual rate of return. The value of Γ_{ij} is 0, indicating that there is an error in the yield risk evaluation result; If the value of Γ_{ij} is close to 1, it indicates that there is a significant correlation between the yield risk evaluation result and the actual yield risk level. Then the cross-correlation coefficient of the yield risk evaluation results of the model in this paper is shown in Fig. 4.

As shown in Fig. 4, the cross-correlation coefficient test results of the yield risk evaluation results of the model in this paper show that the cross-correlation number Γ_{ij} is 1, indicating that the yield risk evaluation results are significantly correlated with the actual yield risk level, and the evaluation results are reliable.

In order to highlight the mining effect of the model in this paper, it compares the model in reference [8], model in reference [9] and model in reference [10] to determine whether the model in this paper has application advantages. Fig. 5 shows the comparison test results of the four models.

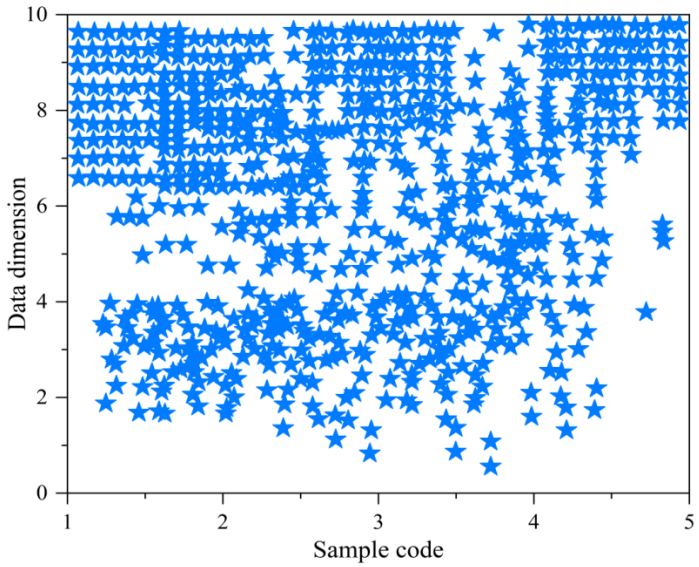


Fig. 2. Before dimensionality reduction of the data of yield risk influencing factors.

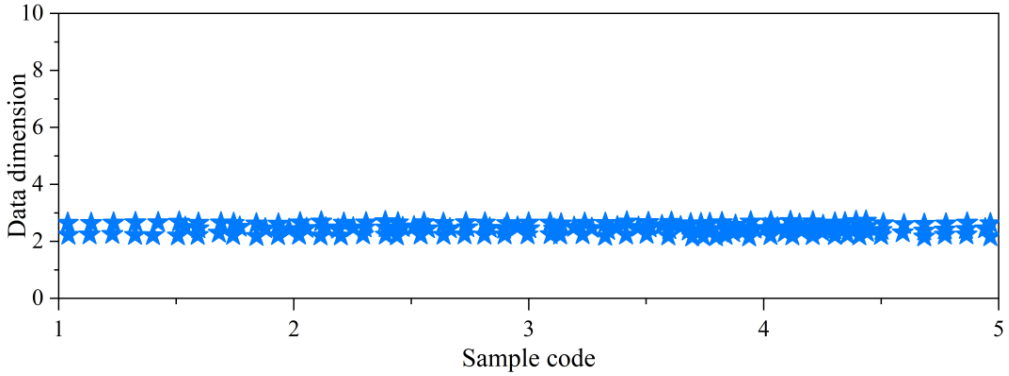
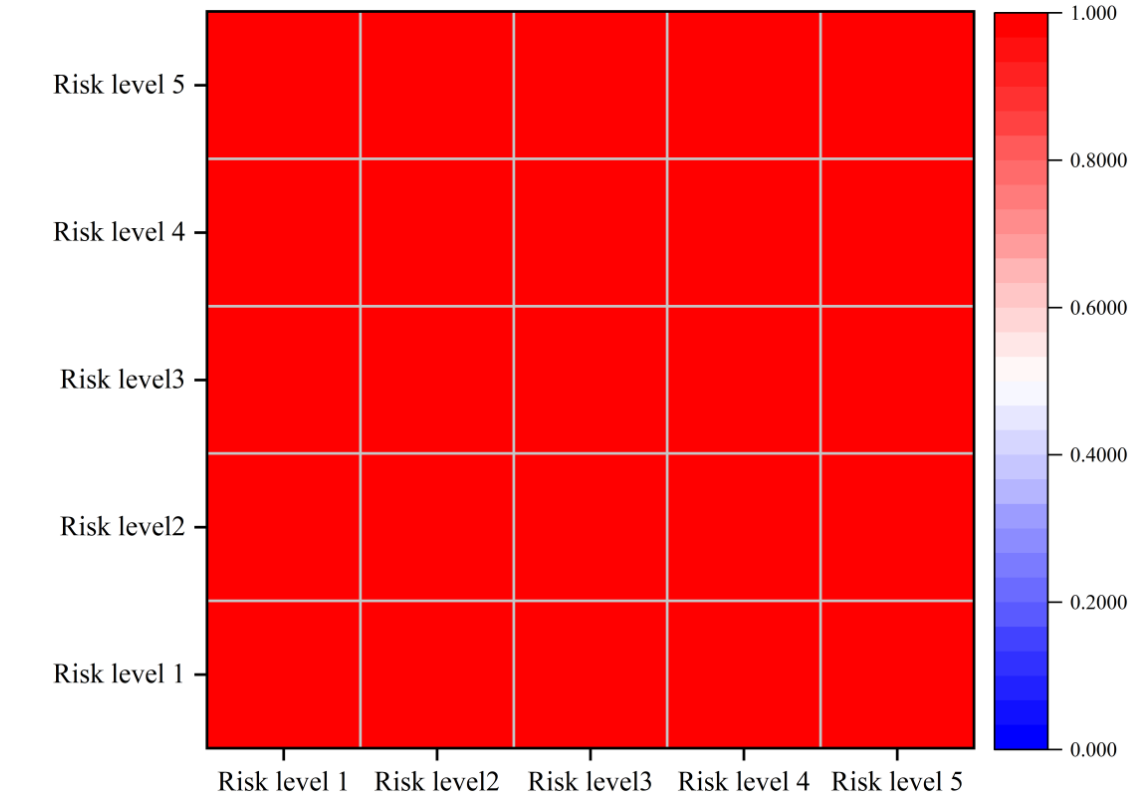
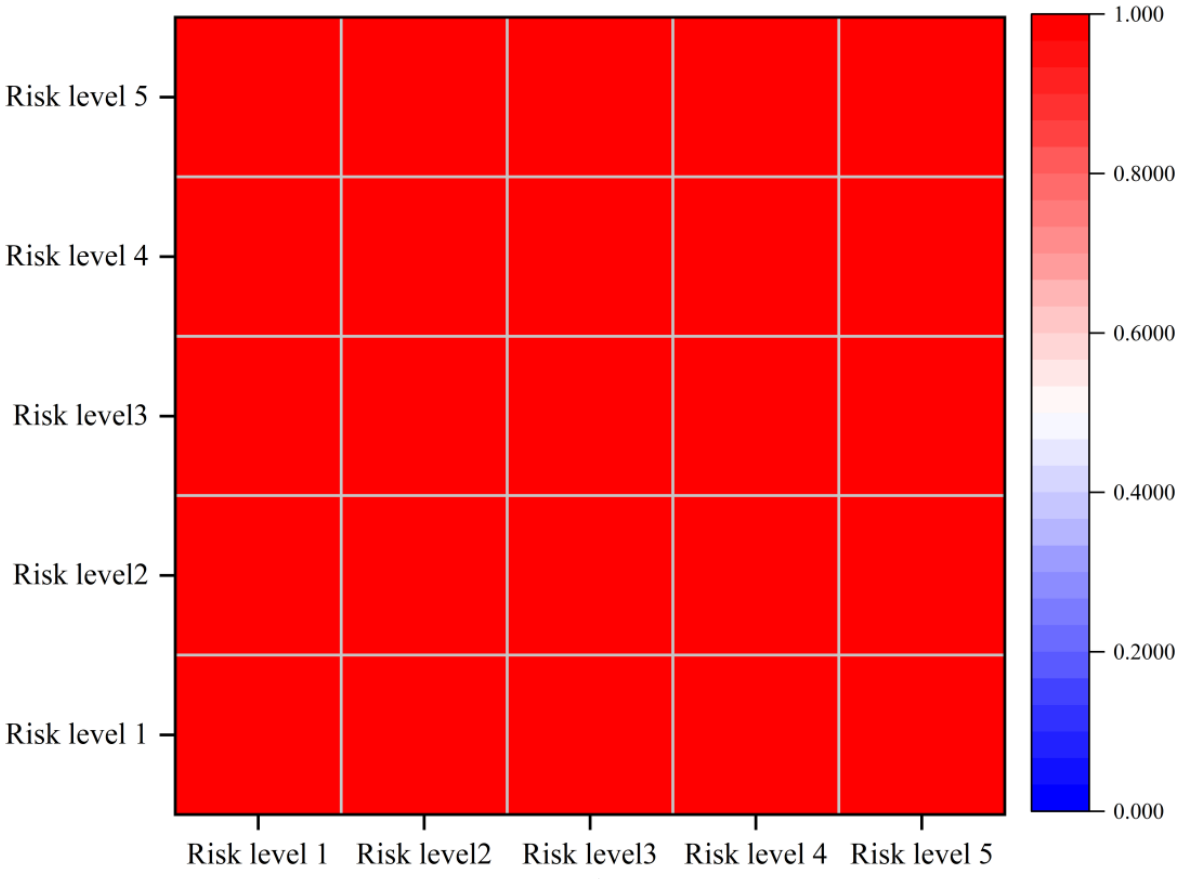


Fig. 3. After the dimensionality reduction of the data on the influencing factors of yield risk.



(a) Project of enterprise 1



(b) Project of enterprise 2

Fig. 4. The test results of the cross-correlation coefficient of the yield risk evaluation results of this model.

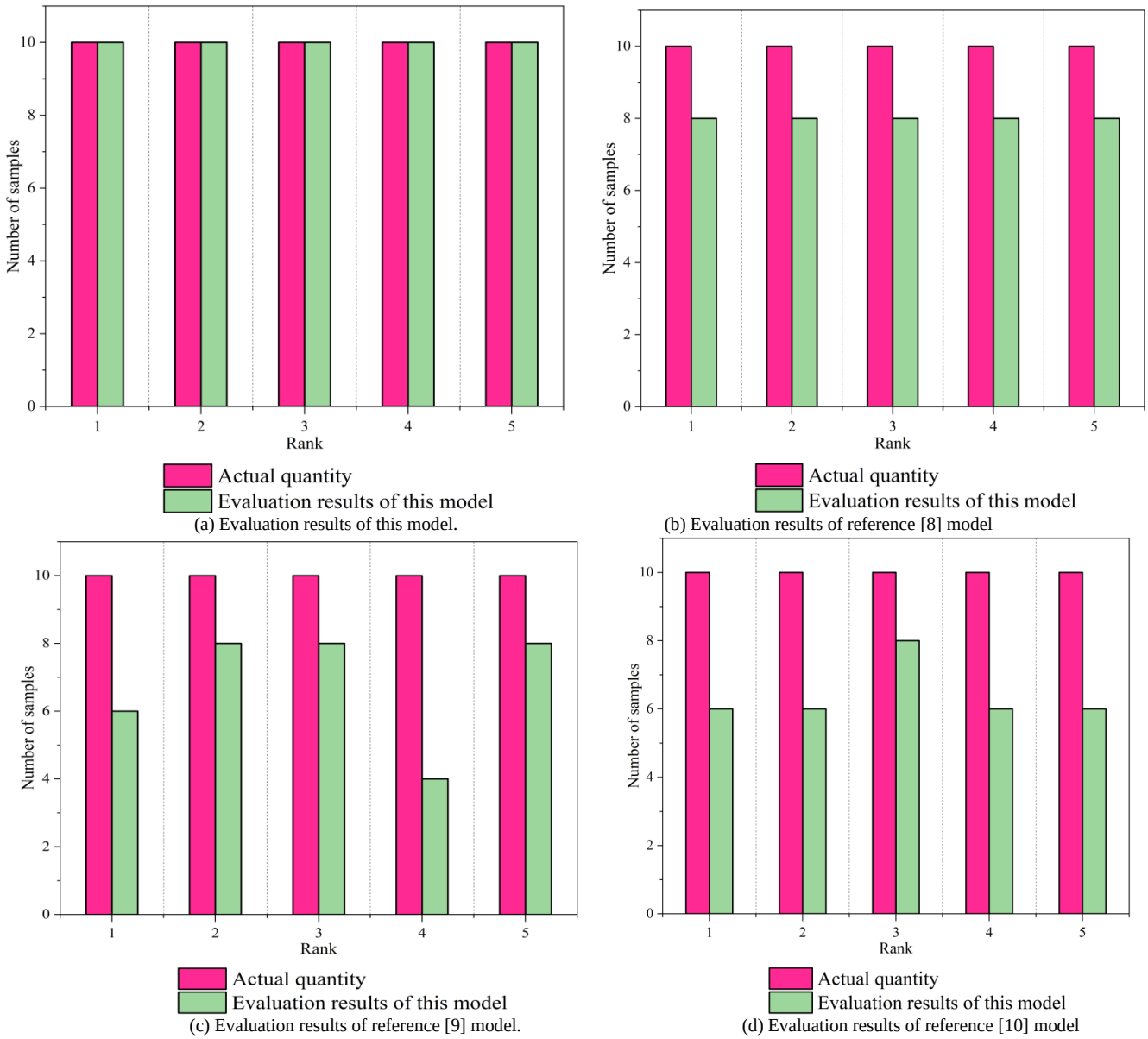


Fig. 5. Comparative test results of four models.

As shown in Fig. 5, after the model in this paper, the model in reference [8], the model in reference [9] and the model in reference [10] evaluate the same yield risk target, the evaluation effect of the model in this paper is the best, which can accurately evaluate the yield risk level, and the evaluation results are in line with the reality. However, the evaluation results of the model in reference [8], the model in reference [9], and the model in reference [10]are biased. The reason is that the model in this paper can reduce the dimension of risk indicators and screen indicators before evaluation so as to ensure the rationality of indicators.

IV. Conclusion

This paper constructs an intelligent evaluation model of yield risk based on empirical probability distribution, comprehensively analyzes the yield risk from multiple perspectives, analyzes it from financial and non-financial perspectives, constructs an intelligent evaluation index of yield risk, and designs a comprehensive evaluation model based on the empirical probability distribution of risk indicators. After in-depth performance testing in the experiment, the test conclusions are as follows:

1) After the dimensionality reduction of the data of the factors influencing the yield risk in the model of this paper, the data dimension is obviously controlled in a unified range, ensuring data regularity.

2) The cross-correlation coefficient test results of the yield risk evaluation results of the model in this paper show that the correlation number is 1, the yield risk evaluation results have a significant correlation with the actual yield risk level, and the evaluation results are reliable.

3) Compared with other models, the model in this paper has the best evaluation effect and can accurately evaluate the

risk level of the rate of return. The evaluation results are in line with reality.

For the future research work of the intelligent evaluation model of return risk based on empirical probability distribution, the following are some prospects:

1) Establish a more accurate model: Future research can explore a more accurate model and introduce market sentiment and macroeconomic indicators to improve the prediction accuracy of the model.

2) Consider the nonlinear relationship: In fact, there may be a nonlinear relationship between stock returns, and future research can explore how to consider the nonlinear relationship to better describe and predict the return risk.

3) Consider the change of market conditions: The change of market conditions is one of the important factors affecting the rate of return. Future research can explore how to adjust model parameters according to different market conditions and provide corresponding risk assessment.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing of interests.

AUTHORSHIP CONTRIBUTION STATEMENT

Yang Jing: Writing-Original draft preparation

Conceptualization, Supervision, Project administration.

Zhou Yanru: Conceptualization, Methodology, Supervision

AVAILABILITY OF DATA AND MATERIALS

On Request

REFERENCES

[1] K. M. Frias, D. L. Popovich, D. F. Duhan, and R. F. Lusch, "Perceived market risk in new ventures: A study of early-phase business angel investment screening," Journal of Macromarketing, vol. 40, no. 3, pp. 339–354, 2020.

[2] J. K. Hammitt and L. A. Robinson, "Introduction to special issue on risk assessment, economic evaluation, and decisions," Risk Analysis, vol. 41, no. 4, pp. 559–564, 2021.

[3] F. He, Y. Li, T. Xu, L. Yin, W. Zhang, and X. Zhang, "A data-analytics approach for risk evaluation in peer-to-peer lending platforms," IEEE Intell Syst, vol. 35, no. 3, pp. 85–95, 2020.

[4] E. F. Drabo et al., "A Social-Return-On-Investment Analysis Of Bon Secours Hospital's 'Housing For Health'Affordable Housing Program: Study evaluates the broader social, environmental, and economic benefits of Bon Secours Hospital's Housing for Health program.," Health Aff, vol. 40, no. 3, pp. 513–520, 2021.

[5] H. San Martín, B. Hernández, and Á. Herrero, "Social consciousness and perceived risk as drivers of crowdfunding as a socially responsible investment in tourism," J Travel Res, vol. 60, no. 1, pp. 16–30, 2021.

[6] D. Li, J. Bi, and M. Hu, "Alpha-robust mean-variance investment strategy for DC pension plan with uncertainty about jump-diffusion risk," RAIRO-Operations Research, vol. 55, pp. S2983–S2997, 2021.

[7] Q. Wu, Y. Gao, and Y. Sun, "Research on Probability Mean-Lower Semivariance-Entropy Portfolio Model with Background Risk," Math Probl Eng, vol. 2020, pp. 1–13, 2020.

[8] N. Wei, S. Liu, Z. Jiao, and X. Li, "A possible contribution of carbon capture, geological utilization, and storage in the Chinese crude steel industry for carbon neutrality," J Clean Prod, vol. 374, p. 133793, 2022.

[9] H. Shuifeng, D. Yating, and L. I. Hongdan, "Risk evaluation of China's overseas mining investment based on structural power theory," China Mining Magazine, vol. 30, no. 10, pp. 24–31, 2021.

[10] Y. Changwei, L. Zonghao, G. Xueyan, Y. Wenying, J. Jing, and Z. Liang, "Application of BP neural network model in risk evaluation of railway construction," Complexity, vol. 2019, 2019.

[11] R. Castro and J. Tapia, "Adding a social risk adjustment into the estimation of efficiency: the case of Chilean hospitals," Quality Management in Healthcare, vol. 30, no. 2, pp. 104–111, 2021.

[12] A. Garratt and I. Petrella, "Commodity prices and inflation risk," Journal of Applied Econometrics, vol. 37, no. 2, pp. 392–414, 2022.

[13] L. Lu, A. Gavin, F. J. Drummond, and L. Sharp, "Cumulative financial stress as a potential risk factor for cancer-related fatigue among prostate cancer survivors," Journal of Cancer Survivorship, vol. 15, pp. 1–13, 2021.

[14] J. Safitri, S. Suyanto, M. L. Taolin, and S. L. Prasilowati, "Inclusion of interest rate risk in credit risk on bank performance: Evidence in Indonesia," JRAP (Jurnal Riset Akuntansi dan Perpajakan), vol. 7, no. 01, pp. 13–26, 2020.

[15] W. Li et al., "Characteristic of five subpopulation leukocytes in single-cell levels based on partial principal component analysis coupled with Raman spectroscopy," Appl Spectrosc, vol. 74, no. 12, pp. 1463–1472, 2020.

[16] Y. Zhang et al., "Infrared image impulse noise suppression using tensor robust principal component analysis and truncated total variation," Appl Opt, vol. 60, no. 16, pp. 4916–4929, 2021.

[17] P. J. Atkins and M. Cummins, "Improved scalability and risk factor proxying with a two-step principal component analysis for multi-curve modelling," Eur J Oper Res, vol. 304, no. 3, pp. 1331–1348, 2023.

[18] F. B. Salling, N. Jeppesen, M. R. Sonne, J. H. Hattel, and L. P. Mikkelsen, "Individual fibre inclination segmentation from X-ray computed tomography using principal component analysis," J Compos Mater, vol. 56, no. 1, pp. 83–98, 2022.

[19] K. A. Asha, L. E. Hsu, A. Patyal, and H. M. Chen, "Improving the quality of FPGA RO-PUF by principal component analysis (PCA)," ACM J Emerg Technol Comput Syst, vol. 17, no. 3, p. 3442444, 2021.

[20] Z. Nie et al., "Using a single sensor for bridge condition monitoring via moving embedded principal component analysis," Struct Health Monit, vol. 20, no. 6, pp. 3123–3149, 2021.

[21] X. Chen, L. Wang, and Z. Huang, "Principal component analysis based dynamic fuzzy neural network for internal corrosion rate prediction of gas pipelines," Math Probl Eng, vol. 2020, pp. 1–9, 2020.

[22] T. Alharbi, "Pulse-shape discrimination of internal α-contamination in LaBr3: Ce detectors by using the principal component analysis," Journal of Instrumentation, vol. 15, no. 06, p. P06010, 2020.

[23] J. Zhang, "A study on mental health assessments of college students based on triangular fuzzy function and entropy weight method," Math Probl Eng, vol. 2021, pp. 1–8, 2021.

[24] J. Huan, D. Ma, W. Wang, X. Guo, Z. Wang, and L. Wu, "Safety-state evaluation model based on structural entropy weight–matter element extension method for ancient timber architecture," Advances in Structural Engineering, vol. 23, no. 6, pp. 1087–1097, 2020.

[25] P. Liu and Y. Li, "An improved failure mode and effect analysis method for multi-criteria group decision-making in green logistics risk assessment," Reliab Eng Syst Saf, vol. 215, p. 107826, 2021.

[26] M. Jia et al., "Network Optimization of CNT Yarn Sensor Based on NNIA Algorithm in Damage Monitoring of 3D Braided Composites," Materials, vol. 15, no. 23, p. 8534, 2022.

Enhancing Question Pairs Identification with Ensemble Learning: Integrating Machine Learning and Deep Learning Models

Salsabil Tarek[1*], Hatem M. Noaman[2], Mohammed Kayed[3]
Computer Science Department, Faculty of Computer Science, Nahda University, Beni-Suef 62511, Egypt[1]
Computer Science Department, Faculty of Computers and Artificial Intelligence,
Beni-Suef University, Beni Suef 62511, Egypt[2, 3]

***Abstract*—The effectiveness of machine learning (ML) and deep learning (DL) models on the Quora question pairs dataset is investigated in this study. ML models, including AdaBoost, reached 73.44% test accuracy, while ensemble learning approaches enhanced outcomes even further, with the Hard-Voting Ensemble achieving 76.13%. DL models, such as FCN, demonstrated test accuracy of 81% with cross validation. These findings contribute to natural language processing by demonstrating the potential of ensemble learning for ML models and the DL models' detailed pattern-capturing capacity.**

***Keywords*—Ensemble learning; natural language processing; deep learning; machine learning**

I. INTRODUCTION

Ensemble learning has grown increasingly popular as an efficient machine learning technique to increase accuracy and predictability in predictions. At its core, ensemble learning entails merging multiple models into a more accurate predictor [1]. There are various techniques for producing multiple models simultaneously. Bagging is one such approach [2], where multiple models are trained on random subsets of training data for multiple models to share a set of results. Another technique, called boosting [3], allows models to be trained sequentially to address errors from prior models. A third technique called stacking [4] uses multiple models trained on identical data and employs a meta-model as a bridge to integrate their outputs. Stacking is a method by which multiple models are trained to find an ideal way of combining their predictions. Predictions from base models serve as input into an intermediate-level model which then learns how to weigh and combine them to produce a final prediction.

Ensemble learning offers numerous advantages over single models. It reduces the risk that an overfitted model becomes too complex and learns noise instead of patterns in data; and captures more patterns. Ensemble learning can also enhance the predictability and stability of predictions by creating a model less sensitive to small fluctuations in data. Furthermore, ensemble learning combines all models' strengths for improved accuracy in forecasts. Ensemble learning in machine learning refers to the practice of combining multiple models into one to increase accuracy and robustness. This involves training multiple models on one dataset using different initializations or hyperparameters for their training sessions. Ensemble learning's central concept is that combined models will outperform individual ones due to being better at capturing more patterns while avoiding overfitting. Ensemble learning in natural language processing has yielded excellent results for a range of tasks such as text classification and sentiment analysis. Ensemble learning can easily manage various data types - textual as well as structured data - making it applicable to many NLP tasks. Selecting the optimal ensemble method and configuration can be a complex process that involves extensive experimentation and evaluation. Furthermore, training multiple models may prove too expensive a prospect; hence ensemble learning has proven an indispensable asset to NLP applications. Recent machine learning studies have demonstrated the power of ensemble learning over individual classifiers when it comes to improving performance. Ensemble learning has had a considerable effect on machine learning applications, leading to its widespread usage across various domains such as text classification. [5- 8]. Deep neural networks (DNN), one of the cornerstones of machine learning, have emerged as a formidable force over recent years. DNNs have contributed to advancing natural language processing and text classification techniques. Speech recognition, object detection, visual object recognition, and object identification all benefit. [9]. Deep learning techniques differ from classical machine learning in that they automatically identify and extract complex features without manually creating them [10]. Deep learning employs multiple network architectures to address problems, including feed-forward neural nets, convolutional networks, and recurrent networks [11]. Recently, many attempts have been made to combine DNNs and ensemble methods to enhance prediction performance. To develop ensemble deep learning, one of the easiest and simplest approaches is integrating deep learning directly into existing ensemble learning methods. Most attempts focus on creating weighted-average models of deep learning models; studies have demonstrated that ensembles incorporating DNNs outperform individual DNNs for classification tasks [12].

The Quora Question-Pairs dataset is one of the most frequently utilized resources for identifying question pairs. With over 400,000 questions identified as either duplicates or not duplicates, this dataset offers an excellent way to pinpoint question pairings. Duplicate questions must be identified to reduce redundancy on search engines, forums, and question-answering software. Unfortunately, due to its wide range of languages and question structures, this task is no easy feat. This

study investigates the efficacy of ensemble learning methods in recognizing question pairs from the Quora Question Pairs data set. Various ensemble methods were compared to improve the accuracy of machine learning models as well as deep learning models. Current experiments demonstrate how ensemble learning can be accomplished by combining models such as logistic regression, random forests, and XGBoost with deep learning models like convolutional neural networks or long-short-term memory networks. This study will conduct further investigations of hyperparameters and assess the interpretability of ensemble models, contributing to an expanding body of research in ensemble learning and natural language processing (NLP), providing insight into optimal ensemble methods and configurations to identify question pairs; these results may also have applications in search engines, online forums and question answering software systems.

This research performs comparative experiments on the dataset using the most popular ensemble techniques; weighted and vote ensemble methods, which contribute to and encompass the development of ensemble learning algorithm, extensive experimentation with various machine learning models, performance evaluation, comparative analysis against existing methods, and an exploration of prediction strategies within the context of ensemble learning for semantic similarity. To that purpose, the following are the primary contributions of the paper:

In current study, comprehensive trials were carried out by cross-validating training several machine learning models, followed by precise evaluations. This strategy demonstrates the ensemble methods underlying potential for assessment the performance of models on the dataset of Quora question pairings. Ensemble learning approaches using a varied variety of machine learning models had not been investigated on the Quora question pairs dataset.

This analysis using Current Ensemble Techniques is expanded to evaluate the effectiveness of ensemble approach with a number of well utilized ensemble techniques. The competitiveness and benefits in the context of semantic similarity are shown in this comparative analysis.

This research covers a comparison by applying deep learning to the Quora question pairs dataset which enables us to illustrate deep learning's performance and natural benefits in dealing with the issues given by question pair classification tasks that providing a complete evaluation of model performance and emphasizing the practical value of these gains in tackling real-world situations.

This paper is organized as follows. Section II presents the related works. Data preparation is shown in Section III. Section IV discusses the proposed model. While the results and discussion are given in Section V. Finally, Section VI concludes the current research.

II. Related Work

Several studies as shown in Table I, have investigated the effectiveness of different ensemble methods, such as bagging, boosting, and stacking, in improving the performance. Study by Dhakal et al. [13] focused on used Natural Language Processing to address the issue of question duplication in Q&A forums by using Deep Learning to determine whether question pairings are duplicates. Sharma et al. research [14] investigated the task of Natural Language Understanding (NLU) through the analysis of duplicate questions in the Quora dataset. They explored the dataset extensively and applied a variety of machine learning models, including linear and tree-based models. To overcome the duplicate question problem provided by the Quora dataset, they tried an enormous number and variety of machine learning models. A basic Continuous Bag of Words neural network performed the best, they also performed error analysis and discovered some subjectivity in the dataset's labelling.

Chandra and Stefanus [15] modelled the Quora question pairings dataset to find a related question; The assignment is a binary categorization. They attempted several methodologies and algorithms, as well as a distinct approach from earlier efforts. For XGBoost and CatBoost, they employed Bag of Words with Count Vectorizer and Term Frequency-Inverse Document Frequency with Unigram for feature extraction. Furthermore, they tested the WordPiece tokenizer, which considerably increases model performance, and they were able to get up to 97 percent accuracy. They tested Bag of Words with two boosting algorithms: Catboost and XGBoost. They also used simple LSTM and BERT to evaluate Quora Question Pairs. The results reveal that BERT outperformed the other models.

The goal of research done by Sharma et al. [16] was to determine whether a question pair is similar; they used a dataset provided by Quora on Kaggle to accomplish this. That dataset had four lakh records, which assisted us in training their models and obtaining the necessary outcomes. They employed Natural Language Processing knowledge and different classification and boosting techniques to determine which is more useful, then they examined the accuracy of various models to determine which method is best suited for the task. The same has been done with the aid of multiple graphs and tables to highlight the differences in the accuracy of various algorithms. It was critical to clean and pre-process the data before applying any algorithm. After that, they used techniques such as the Count Vectorizer with XG Gradient Boosting, the TF-IDF Vectorizer with XG Gradient Boosting, Logistic Regression, and Random Forest.

Anishaa et al. [17] proposed a novel approach by filtration of the Quora datasets using SQLite which takes one-quarter the time it takes to pre-process the same dataset using existing methodologies such as python functions. It concluded that XGBoost outperformed the other machine learning approaches discussed, it has also been discovered that pre-processing with SQLite has improved response time. To analyses and find the best model, they employed machine learning techniques such as Random Forest, Logistic Regression, Linear SVM (Support Vector Machine), and XGBoost. The error log loss functions (0.887, 0.521, 0.654, and 0.357) of the machine learning algorithms were analyzed and compared. XGBoost has the best performance among the other models.

Chandra and his colleagues [18] provided a technique for detecting duplicate question pairs in their study by dividing the

selected dataset in a 70:30 ratio. A technique known as random splitting was employed. It was discovered that if a feature timestamp for each question was given, then a time-based splitting might be utilized to partition the dataset, because the questions asked earlier differed greatly from the questions answered recently. Enabling this feature increases accuracy. The model uses the Glove pre-embedding to classify the questions. Features such as fuzzywuzzy help to achieve very minimal log loss. The log loss for the XGBoost model was 0.35, while the log loss for the Siamese LSTM model was 0.21.

Furthermore, Gontumukkala et al. [19] proposed a method to overcome two drawbacks of Quora as the occurrence of duplicate questions that cause ambiguity and insincere questions that lessen the value of the site by suggesting a strategy to address these two issues using Deep Learning (DL) and Natural Language Processing (NLP) approaches. Bi-directional Long Short-Term Memory (BiLSTM) and Bi-Gated Recurrent Unit (BiGRU) architectures with attention mechanisms were used for both problems, and Siamese Manhattan Long Short-Term Memory (MaLSTM) architectures were used for question pair identification. Five different word embeddings were used for each problem. When it comes to accuracy, precision, recall, and F1 Score, the models that have been used are performing well. For the classification of sincere questions, their model achieved the highest accuracy of 95% and the highest F1 score of 0.82 using FastText + BiLSTM + BiGRU. For the identification of Quora question pairs, their research work achieved the highest accuracy of 90% and the highest F1 score of 0.89 using Paraphrase-MiniLM-L6-v2 + Siamese MaLSTM.

Sendi et al. [20] introduced a transparent, deep ensemble classification method based on multiagent arguments. This approach leverages deep learning algorithms combined with argumentation to outperform traditional ensemble methods, providing explain-ability while meeting Explainable AI needs. Furthermore, Mohammed and Kora [21] proposed a novel ensemble meta-learning strategy which combines multiple classifiers was proposed.

Karlos and his colleagues [22] presented the proposed ensemble method outshone other ensemble methods on benchmark datasets in terms of performance. Furthermore, its meta-learner's performance was further improved by taking advantage of probability distributions for class labels. This paper describes an ensemble-based training scheme for binary classifying using random feature splitting.

Gonçalves et al. [23] assessed the effectiveness of a multi-view ensemble for full-text classification using different document sections as views. Results demonstrate its accuracy in classification accuracy and F1-score calculations. C4.5 serves as their meta-learner to implement support vector machine algorithms for stacking. For views creation, they utilized the OHSUMED full-text biomedical dataset; results from experiments demonstrate that multi-view techniques significantly improve text classification within biomedical text mining. Findings indicate that adding text from certain sections to datasets outperforms simply using titles and abstracts alone.

Haghighi and Omranpour [24] offered an ensemble classifier stacking model to recognize handwritten digits. Addressing different writing styles and structural similarities among digits, this model uses a convolutional network (CNN) paired with bidirectional long-short-term memory (BLSTM) to unify both methods. It utilizes the innovative use of image class probability vectors as input to the meta-classifier, further increasing accuracy with its deep-learning model through BLSTM's ability to learn vectors and arrays. Stacking ensemble classification helps reduce recognition errors by considering similarities between Persian/Arabic numbers and writing style variations. The model was tested on a large dataset consisting of 102.352 points from 102.352 classes of Persian/Arabic data. It achieved high accuracy rates of 99.98% for the training set and 99.39% for the test set. These results demonstrate enhanced performance compared with convolutional neural network experiments and previous research.

Araque et al. [25] investigated ways of improving performance using both deep learning techniques and traditional surface approaches for Sentiment Analysis. Deep learning offers advantages over surface approaches in terms of automatic feature extraction and richer representation abilities. This paper features six contributions; as an initial task, a deep-learning-based sentiment classifier using word embedding is constructed as a baseline solution. Second, two ensemble techniques combine the baseline with other surface classifications commonly employed for Sentiment Analysis. Thirdly, they introduce two models that leverage data from multiple sources by combining surface and deep features. Fifthly, a taxonomy that classifies all proposed models is presented. Seven datasets from microblogging and movie reviews domains are utilized to conduct various experiments that compare performance between proposed models and baseline deep learning systems. An F1-Score analysis verifies the performance of their proposed models.

A study done by Onan et al. [26] implemented a multi objective voting scheme for sentiment analysis that uses optimization. The ensemble method incorporates a static classifier, majority voting errors, forward search, and multi objective differentiation evolution algorithm. Base learners include Bayesian log regression, naive Bayes (linear discriminant analysis), logistic regression, and support vector machine while the current method outshone ensemble learning techniques in various classification tasks. Ankit and Saleena [27] offered a Twitter Sentiment Analysis which detects sentiments and opinions within tweets. To achieve accurate classification of tweets they selected an accurate classifier. They consider common base classifiers such as Naive Bayes and Random Forest, SVMs, and Logistic Regression as base classifiers. An ensemble classifier combining all these classifiers is then proposed to improve performance and accuracy.

Convolutional Neural Networks (CNN) [28] used to identify the semantic similarity of questions using the Quora question pairs dataset. Glove pre-trained word embedding applied to identify the semantic similarity between queries. This word embedding vector is fed into CNN, and the results are compared to Siamese Neural Networks. The model achieved an accuracy of 79%. Wang et al. [29] used the Stack Overflow dataset to investigate three deep learning algorithms

to identify duplicate questions: DQ-CNN, DQ-RNN, and DQ-LSTM, which are based on CNN, RNN, and LSTM, respectively. Six distinct question groups are used to evaluate the effectiveness of DQ-CNN, DQ-RNN, and DQ-LSTM. Except for the Ruby question group, their experimental results reveal that DQ-LSTM outperforms DupPredictor, Dupe, DupePredictorRepT, and DupeRep in terms of recall-rate@5, recall-rate@10, and recall-rate@20.

TABLE I. SUMMARY OF THE PREVIOUS WORKS

MODEL	Accuracy	Precision	Recall	F1-score
Supervised Machine Learning Algorithm [13]				
Random Forest [13]	0.741	-	-	-
Logistic Regression [13]	0.677	-	-	-
Decision Tree [13]	0.683	-	-	-
Support Vector Machine [13]	0.542	-	-	-
K Nearest Neighbors[13]	0.719	-	-	-
Multinomial Naive Bayes [13]	0.673	-	-	-
Most frequent class [14]	63.1	-	-	-
LR with Unigrams[14]	75.4	-	-	63.8
LR with Bigrams[14]	79.5	-	-	70.6
Linear LR with Trigrams[14]	80.8	-	-	71.8
LR with Trigrams, tuned[14]	80.1	-	-	71.5
SVM with Unigrams[14]	75.9	-	-	63.7
SVM with Bigrams[14]	79.9	-	-	70.5
SVM with Trigrams[14]	80.9	-	-	72.1
Tree-Based Decision Tree[14]	73.2	-	-	65.5
Random Forest[14]	75.7	-	-	66.9
Gradient Boosting[14]	75.0	-	-	66.5
CBOW[14]	**83.4**	-	-	**77.8**
LSTM[14]	81.4	-	-	75.4
LSTM + Attention [14]	81.8	-	-	75.5
BiLSTM[14]	82.1	-	-	76.2
BiLSTM + Attention [14]	82.3	-	-	76.4
CV-XGBoost [15]	68.09	-	-	-
CV-CatBoost [15]	74.66	-	-	-
TF-IDF-XGBoost [15]	69.14	-	-	-
TF-IDF CatBoost [15]	**75.39**	-	-	-
XGB [16]	0.79	0.80	0.80	-
Logistic regression [16]	0.74	0.75	0.73	-
SGDC [16]	0.74	0.73	0.75	-
Random forest [16]	0.83	0.81	0.82	-
XGBoost [18]	69%	0.79	0.69	0.73
Paraphrase- MiniLM-L6-v2 + Siamese MaLSTM [19]	90%	0.85	0.94	0.89
LSTM [30]	83.8%	-	-	0.79
BiLSTM + Frame-GBDT [31]	87.92%	-	-	-
Neural Networks + Multi-head Attention [32]	86.83%	0.84	0.81	0.82
Siamese LSTM [33]	82.77%	0.79	0.70	0.75
XG Boost [34]	81%	-	-	-
BERT Model [35]	80%	-	-	-

III. DATASET

In this study, experiments were conducted on the Quora Question Pairs dataset, which is widely used for question pairs identification tasks. The dataset consists of 404,290 question pairs, each identified by a unique ID. For each question pair, the dataset provides the question IDs (qid1 and qid2), the actual text of the first question (question1), the actual text of the second question (question2), and a binary label indicating whether the questions are duplicates (1) or not duplicates (0).

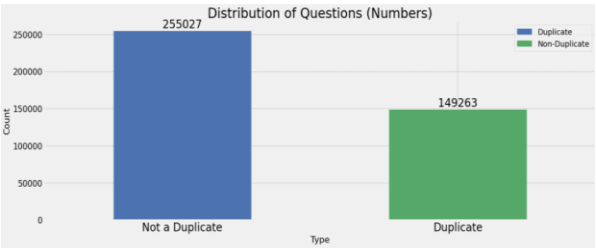


Fig. 1. The distribution of questions in the QQP dataset.

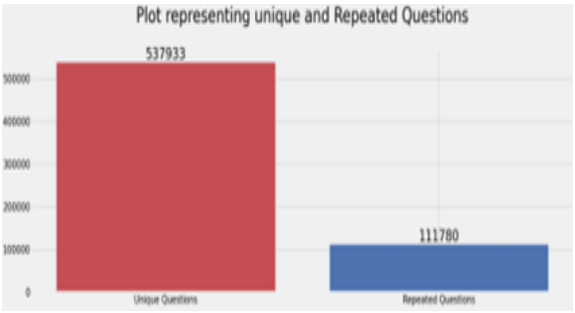


Fig. 2. Unique and repeated questions in the QQP dataset.

Fig. 1 shows a notable aspect of the dataset is the distribution of question pairs. Approximately 63.08% of the pairs are labeled as not similar or non-duplicates, while 36.92% are labeled as similar or duplicates. This class imbalance should be taken into consideration during the data preprocessing and model training stages.

One interesting idea were explored during the data preparation stage is the identification of unique and repeated questions as shown in Fig. 2. By analyzing the dataset, discovered that 98% of the questions occur only once, implying that most questions do not repeat themselves. The dataset also revealed that the maximum number of times a question is repeated is fifty.

Understanding the distribution and uniqueness of questions provides valuable insights into a dataset, aiding in designing appropriate preprocessing techniques and sampling strategies.

Cross-validation was used to assess the performance of models on the Quora Question Pairs dataset. The code snippet illustrates how to select and initialize various machine learning algorithms, such as Gaussian Naive Bayes (GNB), Logistic Regression (LR), Stochastic Gradient Descendant, Decision Tree (DT), Random Forest AdaBoost Extra Trees. Deep Learning classifiers included Fully Connected Networks (FCN), LSTM Bidirectional LSTM.

As part of models' assessment, stratified 10-fold cross-validation were used to compare models. This technique ensures each fold maintains an equal distribution of classes to that found in the original dataset and minimizes potential bias. Accuracy scores were used to judge each classifier's performance; files containing this information allow for thorough comparisons among them.

Cross-validation provides us with a powerful way to evaluate the performance of simple models for question pair identification using the Quora dataset. Reliable estimates of each model's accuracy allow us to make informed choices about suitability for ensemble learning; the results of which will inform future hyperparameter tuning and model selection steps leading to an ideal ensemble model to identify question pairs.

IV. PROPOSED MODEL

This paper proposes a model of a schematic representation architecture that uses ensemble approaches on the QQP dataset to improve the accuracy and reliability of question pair similarity prediction as provided in Fig. 3. This architecture integrates their predictions to provide strong and comprehensive similarity evaluations, with the goal of capitalizing on the strengths of varied base models. The ensemble technique tries to increase predictive accuracy while also establishing a more robust foundation for question pair analysis by synthesizing multiple perspectives on question relatedness. This architecture advances the subject of question similarity assessment within the context of the Quora dataset by integrating several modelling strategies and evaluating their combined efficacy.

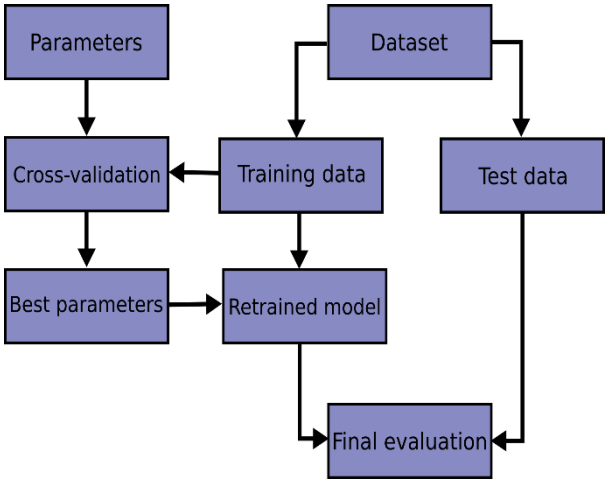


Fig. 3. Proposed general model architecture.

At this stage, text data was processed in various ways to extract features that would aid in the modeling process. The text was initially tokenized using the Tokenizer module of TensorFlow Keras preprocessing modules. This process converted sentences to integer sequences according to word indexes; tokenized sequences were then either extended or reduced until reaching 25 words long. Pre-trained GloVe word embeddings were employed to capture semantic information present in text documents, using word-to-vector maps provided. By iterating over the word index, a matrix of word

embeddings was produced, initially initialized with zeroes before gradually being filled up by GloVe embeddings' word vectors. This word embedding matrix was employed to transform text data, using tokenized and padded question data to fit neural network input requirements for model training. These features were denoted q1_data (original data) and q2_data (transformed version of data), then stored for training use.

Ensemble Learning with Machine Learning Classifiers Approach

Ensemble learning techniques are powerful tools that can increase model performance by combining predictions from different base models into one cohesive prediction. In this research study [36], voting ensemble and weighted average were employed as ensemble techniques.

The Voting Ensemble Technique involves aggregating individual model predictions to produce a final result. A hard voting technique was employed, in which the model with the highest number of votes was selected to predict class. This collective decision-making enabled more effective performances compared to any one model within an ensemble.

Soft voting was utilized, taking into account the probabilities associated with each class prediction and adding up each label's predicted probabilities to determine which class had the highest predicted probability. This method also factored in confidence levels associated with each model prediction to enhance ensemble predictive abilities.

Simple averaging the ensemble techniques was used as a simple averaging approach or weighted averaging. This approach averages the predicted values across different base models using element-wise averages and was utilized in both classification and regression tasks to provide more accurate prediction models from diverse models.

This section presents an in-depth overview of the machine learning (ML) classifiers utilized in this ensemble framework to identify question pairs. These algorithms include Gaussian Naive Bayes (GNB), Logistic Regression (LR), Stochastic Gradient descent, Decision Tree Random Forest, and Gradient Boosting Machine.

Gaussian Naive Bayes, a probabilistic classification method that utilizes Bayes' theorem and assumes independence of features, is derived from Bayesian Naive Bayes. Logistic Regression, on the other hand, is a linear model widely used to estimate probabilities associated with binary outcomes using logarithmic functions. Stochastic Gradient Descent optimizer model parameters iteratively using random subsets from the training data, while Decision Tree creates an interwoven tree-like structure by recursively partitioning features into feature splits. Random Forest employs ensemble averaging to combine multiple decision trees into an ensemble for improved prediction performance, while AdaBoost trains weak learners iteratively by assigning greater weights to instances that have been misclassified, and Extra Trees creates a group of random decision trees to improve generalization. Gradient Boosting Machine creates an ensemble by continuously adding models that correct previous errors.

The dataset was examined using an ensemble learning approach, employing various machine learning models. The dataset was divided into two sets for analysis - X_train (training) and X_test (testing), with target variables created as copies of these two groups Y_train and Y_test, respectively.

For evaluation purposes, widely popular classifiers were trained such as Gaussian Naive Bayes, Logistic Regression, Stochastic Gradient descent, Decision Tree, and Random Forest classifiers as well as AdaBoost Extra Trees Gradient Boosting Machine and Random Forest to compare models' performances against each other. Each of the classifiers employed its respective hyperparameters and training algorithms to train on training data before being deployed on testing data to predict target variables.

An ensemble prediction was created by combining predictions from each model using a simple average technique and then comparing these predictions against labels to calculate accuracy scores. Furthermore, performance metrics such as precision, recall, and F1 scores; specificity loss logs; ROC scores; Cohen's Kappa coefficient of correlation values were calculated and recorded.

Voting Classifier was used to implement a voting-based ensemble. Two variations, hard and soft voting ensembles were explored; hard voting uses majority voting to combine individual classifier predictions; while soft voting used probabilities weighted according to each classifier's confidence in its prediction. Both ensembles were evaluated on the accuracy, classification reports, and confusion matrices for evaluation. A data frame (score) was produced to summarize the results, detailing each model and ensemble's performance metrics.

A. Deep Learning Approach

The current approach also integrates Deep learning (DL) classifiers such as Fully Connected Networks, Long Short-Term Memory, and Bidirectional LSTM into its repertoire. FCN is an architecture of neural networks with fully connected layers; typically used for classification tasks. LSTM is a recurrent neural network type capable of modeling long-term dependencies within sequential data, while Bidirectional LSTM adds context information from past and future inputs by processing sequences both forwards and backward simultaneously. This ensemble framework harnessed their complementary properties and strengths in combination with each other to increase accuracy and robustness for question pair identification tasks. Finally, experimental results were presented as well as evaluating their performance.

As part of the current experimental setup, this study analyzed this dataset using deep learning methods, demonstrating several neural network models including Fully Connected Network, Long Short-Term Memory network (LSTM), and Bidirectional LSTM models. a Time Distributed Layer was used, which employs deep learning architecture to classify questions as duplicates or not. Below is an outline of its implementation and evaluation process.

Initializing all variables and data structures. Next, the dataset was split into two sets - training and testing. Within the training set there can also be further subdivided into five folds

to facilitate cross-validation. The current model was then built using Keras' functional API for flexible architecture consisting of two input layers for every pair of questions posed to it.

The embedding layer transforms words into dense vectors of fixed size for every question, using pre-trained embeddings of words to capture semantic data. Output from this embedding is fed into a Time Distributed Layer followed by a Max Pooling operation to produce fixed-length representations of each question, before being concatenated together and passed through several dense layers such as batch normalization and dropout regularization to reach completion.

The final layer is a dense layer with a sigmoid activation function. This layer produces a score that indicates the likelihood of two questions being identical.

A 5-fold cross-validation approach is used to evaluate the model. Once trained using a set number of epochs and callbacks are implemented to save weights based on validation accuracy, then tested against a test set to evaluate accuracy and loss of prediction.

The algorithm incorporates visual elements like confusion matrices and learning curves to get a general sense of how well its model is performing. Other metrics, such as precision, recall, F1 scores and specificity metrics were used to evaluate the classification performance of the models.

TABLE II. UTILIZING CROSS VALIDATION FOR COMPARATIVE ANALYSIS OF SIMPLER MODELS ACCURACIES

MODEL	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Naive Bayes	0.67119	0.67324	0.67324	0.67409	0.67698	0.67641	0.67446	0.67347	0.67273	0.66998
Logistic Regression	0.70775	0.70867	0.70860	0.70701	0.71157	0.71078	0.71273	0.70778	0.71071	0.70859
Stochastic Gradient Descent	0.71214	0.71026	0.71344	0.71365	0.72298	0.71269	0.71492	0.70891	0.71121	0.71615
Decision Tree	0.72171	0.72114	0.72160	0.72051	0.72415	0.72191	0.72212	0.72583	0.72068	0.72290
Random Forest	**0.77569**	**0.77728**	**0.77803**	**0.77545**	**0.78092**	**0.77473**	**0.77837**	**0.77918**	**0.78042**	**0.77826**
AdaBoost	0.73220	0.73676	0.72765	0.73408	0.73429	0.73389	0.74032	0.73421	0.73672	0.73933
Extra Trees	0.76806	0.77068	0.77414	0.76866	0.77598	0.76883	0.77569	0.77402	0.77600	0.77378
Gradient Boosting Machine	0.75418	0.75188	0.75435	0.75654	0.75813	0.75283	0.75537	0.75590	0.75685	0.75251

TABLE III. CROSS-VALIDATION RESULTS: MEAN VALUES AND STANDARD DEVIATIONS OF MODEL PERFORMANCE

Algorithm	Cross Validation Means	Cross Validation Errors
Naive Bayes	0.67358	0.00201
Logistic Regression	0.70942	0.00180
Stochastic Gradient Descent	0.71364	0.00372
Decision Tree	0.72226	**0.00156**
Random Forest	**0.77783**	0.00197
AdaBoost	0.73494	0.00344
Extra Trees	0.77258	0.00304
Gradient Boosting Machine	0.75486	0.00195

FCN model architecture features dense layers with different activation functions and dropout layers to prevent overfitting, created using an Adam optimizer with binary cross-entropy function and dropout layers as dropout layers to avoid overfitting. The LSTM architecture employed a recurrent network with LSTM cells. Data was reshaped according to input requirements for an LSTM model and dense layers were added similar to an FCN; dropout layers were also implemented to enhance generalization. Finally, this model was constructed and trained using an optimization algorithm, loss function, and FCN model as its training environment. Implementing the Bidirectional-LSTM Model An additional bidirectional layer was added to an LSTM architectural model for training of Bidirectional-LSTM model, enabling it to capture data from past and future timesteps while increasing understanding of temporal dependencies. Finally, an LSTM was used as the training medium.

V. RESULTS

A. Evaluation Metrics

- Accuracy, which stands as one of the most fundamental, and intuitive evaluation metrics as it measures the ratio of correctly predicted instances over the total number of evaluated instances as shown in formula (1). It signifies the overall correctness of a model's predictions. While accuracy serves as a valuable initial assessment, it may not be the sole determinant of a model's performance, particularly when dealing with imbalanced datasets where one class predominates over others [37].

$$\boldsymbol{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- Error Rate, which quantifies the proportion of incorrectly predicted instances within the dataset and provides a clear picture of misclassifications which is the ratio of incorrectly predicted instances over the total number of evaluated instances as shown in formula (2) and is particularly relevant in scenarios where false positives or false negatives bear substantial consequences [37].

$$\boldsymbol{Error\ Rate} = \frac{FP+FN}{TP+TN+FP+FN} \quad (2)$$

- Precision [37], which accentuates the accuracy of positive predictions and quantifies the proportion of true positive predictions (correctly identified positive instances) relative to the total number of positive predictions (comprising true positives and false positives) as shown in formula (3).

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (3)$$

- Recall (sensitivity or the true positive rate) assesses a model's capability to correctly identify all relevant instances from a dataset [37]. From the proportion of true positive predictions in relation to the total number of actual positive instances (encompassing true positives and false negatives) as shown in formula (4).

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (4)$$

TABLE IV. COMPARISON TABLE OF ML BASED MODELS

	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	Specificity	Matthew Correlation Coefficient	Cohen Kappa	ROC Score	Loss Log
Naive Bayes	0.51866	0.5175	0.42993	0.9409	0.59017	0.26966	0.25797	0.1689	0.605	17.39107
Logistic Regression	0.63082	0.63075	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.5	13.3092
Stochastic Gradient Descent	0.5944	0.5908	0.46496	0.7177	0.56432	0.51652	0.2283	0.21049	0.617	14.74902
Decision Tree	0.68679	0.67717	0.57045	0.509	0.53799	0.77561	0.29225	0.29114	0.642	11.6358
Random Forest	0.69127	0.68942	0.6417	0.3598	0.46106	0.88239	0.28845	0.26644	0.621	11.19449
AdaBoost	0.77972	0.73435	0.66278	0.5712	0.61359	0.82987	0.41553	0.41288	0.701	9.57503
Extra Trees	0.70752	0.70098	0.65113	0.4098	0.50298	0.87147	0.32135	0.30464	0.641	10.7777
Gradient Boosting Machine	0.69441	0.69088	0.6805	0.307	0.42308	0.91563	0.28832	0.25117	0.611	11.14189

- F1-score, as shown in formula (5) represents the harmonic mean of precision and recall. This metric strikes a balance between precision and recall, offering a consolidated score that accounts for both false positives and false negatives [37].

$$\boldsymbol{F1-score} = \frac{\boldsymbol{2\times Precision\times Recall}}{\boldsymbol{Precision-Recall}} \quad (5)$$

B. Ensemble Learning on ML Results

At this research facility, an in-depth evaluation was conducted to select the ideal model. To do so, several classifiers were used such as Gaussian Naive Bayes (GNB), Logistic Regression (LR), Stochastic Gradient descent Decision Tree Random Forest AdaBoost Extra Trees to select an effective one.

As part of this evaluation of each classifier's performance, first conducted a stratified cross-validation with 10-folds to assess their performance and select the most accurate models for further study, cross-validation scores for each fold were then computed as illustrated in Table II which offers a side-by-side comparison of the variability and average performance of several machine learning methods. The best mean accuracy is demonstrated by Random Forest and Extra Trees, although numerous other algorithms, including Logistic Regression and Decision Tree, also perform consistently.

Calculation and summarizing the mean and standard deviation of cross-validation results for each model as provided in Table III.

After selecting classifiers, an in-depth analysis was conducted using grid search to fine-tune their hyperparameters. This involved optimizing each model's parameters using cross-validation with the GridSearchCV feature; hyperparameters were chosen carefully based on empirical evidence and prior knowledge for each classifier.

After optimizing hyperparameters, every classifier was trained using the entire training dataset. Standard scaling was applied both during training and testing to ensure unbiased evaluation, predictions were made using training data, predictions were made from both sources simultaneously while accuracy and computational time were recorded.

Reports were prepared on the optimal settings and scores for each classifier, along with grid scores from parameter tuning processes, to give an insight into performance variations between hyperparameter combinations. After using independent test data to assess the generalization abilities of classifiers, their generalization abilities were evaluated using various performance metrics. The results of the performance metrics of machine learning models as presented in Table IV reveal that Naive Bayes obtained moderate accuracy but demonstrated a trade-off between precision and recall.

Table V illustrated the additional results presents the performance metrics of ensemble learning techniques applied to the previously mentioned machine learning models. The two ensemble methods used are Simple Average and Voting (both Hard-Voting and Soft-Voting).

Simple Average, Hard-Voting Ensemble, and Soft-Voting Ensemble were among the ensemble approaches used to enhance overall prediction accuracy. The results of applying ensemble learning techniques to the previous machine learning models show improved performance compared to individual models. The Simple Average ensemble achieved good accuracy and precision, but its recall rate was relatively low. The Hard-Voting Ensemble achieved the highest train accuracy and a balanced performance between precision and recall. The

Soft-Voting Ensemble showed a good overall performance, with higher recall but slightly lower precision compared to the Hard-Voting Ensemble. Both ensemble methods demonstrated better performance metrics compared to the individual models, suggesting the effectiveness of combining multiple models in improving predictions.

By leveraging the strengths of individual models and combining their predictions, ensemble learning techniques have the potential to enhance the performance of machine learning models.

TABLE V. ENSEMBLE ON SIMPLE BASE MODELS

	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	Specificity	Matthew Correlation Coefficient	Cohen Kappa	ROC Score
Simple Average	0.74303	0.69804	0.73163	0.28782	0.41312	0.93819	0.30955	0.25852	0.61301
Hard-Voting Ensemble	0.79982	0.76133	0.67724	0.67563	0.67643	0.81150	0.48737	0.48737	0.74357
Soft-Voting Ensemble	0.78274	0.75000	0.63556	0.75703	0.69100	0.74588	0.48896	0.48374	0.75145

In order to optimize these ensemble methods and determine their applicability to other datasets, additional analysis and experimentation are required. The Hard-Voting Ensemble outperformed the individual models in terms of accuracy, precision, memory, and discriminating ability. The Soft-Voting Ensemble exhibited good accuracy and recall as well, but with somewhat lower precision than the Hard-Voting Ensemble. Individual models were outperformed by both ensemble techniques, with the Soft-Voting Ensemble having the greatest ROC score, suggesting higher discriminating abilities.

C. Deep Learning Results

The provided results present the performance metrics of deep learning models applied to the Quora question pairs dataset. The evaluated models consist of Fully Connected Network (FCN), Long Short-Term Memory (LSTM), and Bidirectional LSTM (Bi-LSTM).

The results in Table VI show how three deep learning models, FCN, LSTM, and Bi-LSTM, performed in a classification test. The FCN model obtained 68.47% train accuracy and 68.54% test accuracy, suggesting consistent performance across training and testing periods.

Moreover, through the application of cross-validation, the initial test accuracy of the FCN model, which stood at 68.54%, was significantly improved to 0.81 % as demonstrated in Table VII.

1) Deep learning using cross validation: The results given in Table VIII showed the performance metrics of the FCN model when cross validation was done for a binary classification task. The evaluation of the model was carried out in two classes, which were labeled as 0 and 1. Precision, recall and F1 score were computed for each class.

Results from a binary classification model that was applied to a dataset with two separate classes labelled as 0 and 1. The outcomes are performance measures for a binary classification model for the classes labelled 0 and 1. The model obtains 86% accuracy, 84% recall, and an F1 score of 85% for class 0. It achieves 73% accuracy, 77% recall, and a 75% F1 score for class 1. Collectively, these measures show that the model performs very well at categorizing cases into class 0, while also achieving better for class 1 examples, even though with significantly lower precision.

VI. DISCUSSION

The outcomes and performance metrics of different machine learning and deep learning classifiers in the current study are analyzed to evaluate their effectiveness in predicting target classes.

The results of the performance metrics of machine learning models reveal that Naive Bayes obtained moderate accuracy but demonstrated a trade-off between precision and recall. These results are in consistence with [13] where they proposed to use of ANN's minimal cost architecture and the selection of highly dominating attributes from the questions make it an excellent model for detecting duplicate questions and subsequently finding high-quality replies to queries in Q&A forum. They obtained 0.673 % of accuracy from Multinomial Naive Bayes model.

Logistic Regression performed poorly; however Stochastic Gradient Descent displayed balanced precision and recall in agreement with [16] study article adopted by Sharma et al. who employed Natural Language Processing knowledge and different classification and boosting techniques to determine which is more useful. Then they examined the accuracy of various models to determine which method is best suited for the task. The same has been done with the aid of multiple graphs and tables to highlight the differences in the accuracy of various algorithms. By comparing the two questions, Sharma et al. were able to determine whether they were identical. They used a dataset provided by Quora to solve this hard challenge and trained multiple machine learning algorithms on four entries to determine whether two questions are identical or not. They used multiple techniques after cleaning and preparing the data as needed. First, they used logistic regression, which produced unsatisfactory results. Therefore, they attempted xG boosting with Count Vectorizer and TF-IDF Vectorizer, and they got an accuracy of more than 80%. With 125 trees, Random Forest produced the best results, yielding an accuracy of 83%, which is quite impressive. The performance of the Decision Tree and Random Forest models was comparable, with the latter obtaining slightly greater accuracy and specificity in consistence with results obtained by [13] and [14], which showed percent of specificity of 0.683% and 73.2% respectively. AdaBoost worked admirably, displaying strong accuracy, precision, recall, and discriminating abilities. Extra Trees and Gradient Boosting Machine performed rather well, with a trade-off between various models to determine which method is best suited for the task. The same has been done with the aid of multiple graphs and tables to highlight the differences in the accuracy of various algorithms. By comparing the two questions, Sharma et al. were able to determine whether they were identical. They used a dataset provided by Quora to solve this hard challenge and trained multiple machine learning algorithms on four entries to determine whether two questions are identical or not. They used multiple techniques after cleaning and preparing the data as needed. First, they used logistic regression, which produced unsatisfactory results. Therefore, they attempted XG boosting with Count Vectorizer and TF-IDF Vectorizer, and they got an accuracy of more than 80%. With 125 trees, Random Forest

produced the best results, yielding an accuracy of 83%, which is quite impressive. The performance of the Decision Tree and Random Forest models was comparable, with the latter obtaining slightly greater accuracy and specificity in consistence with results obtained by [13] and [14], which showed percent of specificity of 0.683% and 73.2% respectively. AdaBoost worked admirably, displaying strong accuracy, precision, recall, and discriminating abilities. Extra Trees and Gradient Boosting Machine performed rather well, with a trade-off between precision and recall. These results are in accordance with [14] and [15] where they investigated the task of Natural Language Understanding (NLU) by examining duplicate question identification in the Quora dataset. They conducted extensive investigation of the dataset and employed various machine-learning models, including linear and tree-based models. The researchers discovered that a simple Continuous Bag of Words neural network model outperformed more complex recurrent and attention-based models with accuracy of 83.4 %.

TABLE VI. COMPARISON TABLE OF DL BASED MODELS

	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	Specificity	Matthew Correlation Coefficient	Cohen Kappa	ROC Score
FCN	0.68469	0.68537	0.79293	0.20021	0.31970	0.96939	0.28151	0.20071	0.58480
LSTM	0.70532	0.70494	0.78143	0.27896	0.41114	0.95432	0.33279	0.26915	0.61664
Bi-LSTM	0.65856	0.66013	0.83982	0.09832	0.17603	0.98902	0.20726	0.10691	0.54367

TABLE VII. FCN MODEL WITH CROSS VALIDATION

Evaluation metrics	FCN Model
Train Accuracy	0.95
Test Accuracy	0.81
Precision	0.81
Recall	0.81
F1 Score	0.81
Specificity	0.84
Matthew Correlation Coefficient	0.6
Cohen Kappa	0.6

TABLE VIII. BINARY CLASSIFICATION RESULTS FOR CLASS 0 AND 1

Class	Precision	Recall	F1 Score
0	0.86	0.84	0.85
1	0.73	0.77	0.75

The machine learning models' performance on the Quora question pairs dataset produced diverse results as the Logistic Regression model had a challenging time classifying positive instances, showing limited success along with the Naive Bayes and Stochastic Gradient Descent models. The Gradient Boosting Machine and AdaBoost models achieved the highest train accuracies, outperforming the Decision Tree, Random Forest, and Extra Trees models. These results are in agreement with authors of [17] research, where their goal was to find the best machine learning technique for removing all duplicate questions and increasing user satisfaction. Using a real-time dataset, this work trained and tested four machine learning models to recognize duplicate inquiries. The raw dataset was discovered to be 7GB in size. PL/SQL was used to pre-process data before it was stored in the database. PL/SQL loads the full dataset only once, and data is acquired directly from the database whenever a query is conducted, making this procedure quick and efficient. In one hour, the complete dataset was cleaned and pre-processed effectively. While existing solutions use python methods available in python libraries to pre-process massive datasets, it takes four times as long as PL/SQL. Four distinct machine learning models were applied, and their results were evaluated to determine which model performed the best. Following execution, the error parameters referred from the log loss function for the random model, logistic regression model, linear SVM, and XGBoost are 0.887, 0.521, 0.654, and 0.357, respectively. Because efficiency is inversely related to error function, it can be concluded that XGBoost is the optimal model, delivering highest accuracy in the shortest amount of time, which is supplemented by the unique pre-processing procedures performed using PL/SQL, hence improving overall response time.

The results of the Simple Average ensemble had a middling accuracy but a better specificity, suggesting its ability to recognize negative events in agreements with [22], which offered the use of base ensemble consists of two participants in soft voting mode, but multiple classifiers combined into an ensemble method to improve predictive performance. In addition, the experimental results demonstrated by [26] and [27] showed that ensemble classifiers perform significantly better than standalone classifications or majority voting ensembles for sentiment classification purposes. Furthermore, that research explored how feature representation and preprocessing affect sentiment classification performance in consistent with current data results.

The Deep Learning model has a comparatively high accuracy of 79.29%, indicating its ability to categorize positive events reliably. The reduced recall of 20.02%, on the other hand, indicates that the model struggled to catch many positive events. When compared to the FCN model, the LSTM model performed better in terms of accuracy, precision, and recall. It obtained 70.53% train accuracy and 70.49% test accuracy, with a precision of 78.14% and a recall of 27.89% in consistency with [18] study, which used a Siamese LSTM to assess the semantic similarity of two queries in order to improve prediction. The Siamese network is an architecture composed of parallel neural networks, namely LSTM units, for the parallel processing of two questions, with each question passing through an Embedding Layer, an LSTM unit, and then a dense layer. Following that, the outputs of two networks were integrated and compared, yielding a similarity score reflecting

how similar two queries are. The log loss metric was used as the major statistic in this study to evaluate alternative models. The main addition is that the Siamese network is utilized to process two questions in parallel and find vector representations for each. The vectors produced by this technology enable more effective similarity detection than existing models. The GloVe word embedding method was used to determine the semantic similarity of two queries. As the basis model, a random classifier was developed, then logistic regression, linear SVM, and the XGBoost model were utilized to reduce log loss. Finally, a Siamese LSTM was proposed, which significantly minimizes the loss. The XGBoost model accurately identified 69% of question pairings as duplicate, resulting in a recall rate of 0.69. The precision rate was 0.79, and the F1- score was 0.73. Finally, as compared to individual models, ensemble techniques performed better in the classification challenge. The Hard-Voting Ensemble and Soft-Voting Ensemble performed better in terms of accuracy, precision, memory, and discrimination, highlighting the value of mixing various models. These findings extend machine learning approaches for categorization problems by emphasizing the potential benefits of ensemble methods in improving prediction performance. Furthermore, current results are in agreement with [26] study which proposed a model that identifies duplicate question pairs by integrating three word embedding feature extraction techniques (Google News Vector, FastText Crawl, and FastText Crawl Subword), which results in significantly higher accuracy than these embeddings independently. Furthermore, this study developed a novel Siamese MaLSTM model that uses the Manhattan distance to determine semantic similarity among questions with 95% accuracy, far outperforming previous studies. Looking closely at the manhattan values, the manhattan score classifies the question pairings more accurately than any other embedding in a blend of different word embedding predictions; the duplicate question score is nearly one, while the non-duplicate pair values are nearly zero.

VII. Conclusion

In this study, a thorough investigation of deep learning (DL) and machine learning (ML) models using the dataset of Quora question pairings. To ensure a reliable analysis, the dataset was put through ten folds of cross-validation. A variety of machine learning (ML) models were trained, such as Naive Bayes, Logistic Regression, Stochastic Gradient Descent, Decision Tree, Random Forest, AdaBoost, Extra Trees, and Gradient Boosting Machine, and evaluated the performance of each model using a variety of evaluation criteria.

These findings showed that the ML models performed at various levels. While models like Decision Tree, Random Forest, AdaBoost, Extra Trees, and Gradient Boosting Machine performed better, models like Naive Bayes and Logistic Regression had little success. Following that, ensemble learning strategies like Simple Average, Hard Voting, and Soft Voting were used to improve the performance of the ML models. These ensemble approaches significantly increased F1 scores, accuracy, and precision, demonstrating their efficacy in combining predictions from many models.

In addition, this study investigated how well DL models like the Fully Connected Network (FCN), Long Short-Term Memory (LSTM), and Bidirectional LSTM (Bi-LSTM) performed. The DL models' performance was assessed using precision, recall, and F1 scores after they were trained on the identical dataset of Quora question pairs. Cross-validation was used to evaluate the FCN model with two classes, and the results showed precision, recall, and F1 scores of 0.81 for both classes. These results show that the FCN model has performed well overall.

This study discovered that the ensemble learning techniques applied to the ML models produced competitive results when comparing their performance to that of the DL models. The DL models, on the other hand, showed off their capacity to identify intricate patterns and connections in the dataset. With high precision, recall, and F1 scores for both classes, the FCN model showed promise.

In summary, the current study emphasizes the effectiveness of using ensemble learning techniques to improve the performance of machine learning models. Moreover, this study has observed that deep learning models, the FCN model demonstrate great potential in accurately categorizing pairs of questions. These discoveries contribute to the progress of natural language processing. Offer valuable insights for enhancing question pair classification tasks. Moving forward, it would be beneficial to concentrate on refining these models exploring different architectures and examining their applicability to diverse datasets and real-world scenarios.

VIII. Future Work

In light of current study's results in employing ensemble learning techniques to enhance deep learning models, several avenues for future work emerge, broaden this ensemble learning approach to include a broader range of deep learning architectures, improve computational efficiency, and improve interpretability. Will also examine applications in domains with limited labelled data, assess generalization skills further, and develop adaptive ensemble weight techniques for dynamic data distributions.

Acknowledgment

The authors would like to thank all those who made contributions towards this work.

References

[1] A. Mohammed and R. Kora, “A comprehensive review on ensemble deep learning: Opportunities and challenges,” Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 2, pp. 757-774, Feb. 2023, https://doi.org/10.1016/j.jksuci.2023.01.014

[2] Sagi, O., & Rokach, L, “Ensemble learning: A survey,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 8, no. 4, pp.1249, 2018, https://doi.org/10.1002/widm.1249.

[3] Y. Freund and R. E. Schapire,“A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,” Journal of Computer and System Sciences, vol. 55, pp. 119- 139 , 1997.

[4] D. H. Wolpert, “Stacked Generalization,” Neural Networks, Vol. 5, pp. 241-259, 1992 0893-6080/92.

[5] J. Abellán and C. J. Mantas, “Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring,” Expert Syst Appl, vol. 41, no. 8, pp. 3825–3830, Jun. 2014, doi: 10.1016/j.eswa.2013.12.003.

[6] A. A. Aburomman and M. Bin Ibne Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system," Applied Soft Computing Journal, vol. 38, pp. 360–372, Jan. 2016, doi: 10.1016/j.asoc.2015.10.011.

[7] C. Catal, S. Tufekci, E. Pirmit, and G. Kocabag, "On the use of ensemble of classifiers for accelerometer-based activity recognition," Applied Soft Computing Journal, vol. 37, pp. 1018–1022, Dec. 2015, doi: 10.1016/j.asoc.2015.01.025.

[8] C. F. Tsai, Y. C. Lin, D. C. Yen, and Y. M. Chen, "Predicting stock returns by classifier ensembles," in Applied Soft Computing Journal, Mar. 2011, pp. 2452–2459. doi: 10.1016/j.asoc.2010.10.001.

[9] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553. Nature Publishing Group, pp. 436–444, May 27, 2015. doi: 10.1038/nature14539.

[10] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning."

[11] L. Deng and D. Yu, "Deep learning: Methods and applications," Foundations and Trends in Signal Processing, vol. 7, no. 3–4. Now Publishers Inc, pp. 197–387, 2013. doi: 10.1561/2000000039.

[12] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," Inf Sci (N Y), vol. 501, pp. 511–522, Oct. 2019, doi: 10.1016/j.ins.2019.06.011.

[13] A. Dhakal, A. Poudel, S. Pandey, S. Gaire and H. P. Baral, "Exploring Deep Learning in Semantic Question Matching," IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), Kathmandu, Nepal, pp. 86-91, 2018, doi: 10.1109/CCCS.2018.8586832.

[14] L. Sharma, L. Graesser, N. Nangia, and U. Evci, "Natural Language Understanding with the Quora Question Pairs Dataset," ArXiv, vol abs/1907.01041, 2019, Corpus ID: 195776066, https://doi.org/10.48550/arXiv.1907.01041.

[15] A. Chandra, and R. Stefanus, "Experiments on Paraphrase Identification Using Quora Question Pairs Dataset,"Computation and Language-arXiv, Jun 2020, https://doi.org/10.48550/arXiv.2006.02648.

[16] A.Sharma, S. Jha, S. Arora, S. Garg, and Sandeep Tayal, "Twin Question Pair Classification," Smart and Sustainable Intelligent Systems, WILEY Online Library, Chapter 16, Book Editors: N. Gupta, P. Chatterjee, and T. Choudhury, March 2021, https://doi.org/10.1002/9781119752134.ch16.

[17] V.K.R. Anishaa, P. Sathvika, and S. Rawat, "Identifying Similar Question Pairs Using Machine Learning Techniques," Indian Journal of Science and Technology, vol. 14, no. 20, pp. 1635-1641, 2021, https://doi.org/ 10.17485/IJST/v14i20.312.

[18] M. Chandra, A. Rodrigues, and J. George, "An Enhanced Deep Learning Model for Duplicate Question Detection on Quora Question pairs using Siamese LSTM," IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, pp.1-5, 2022, DOI: 10.1109/ICDCECE53908.2022.9792906

[19] S.S.T. Gontumukkala, Y.S.V. Godavarthi, B.R.R.T. Gonugunta, D. Gupta, and S. Palaniswamy, "Quora Question Pairs Identification and Insincere Questions Classification," 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, pp. 1-6, 2022, doi: 10.1109/ICCCNT54827.2022.9984492.

[20] N. Sendi, N. Abchiche-Mimouni, and F. Zehraoui, "A new transparent ensemble method based on deep learning," in Procedia Computer Science, Elsevier B.V., 2019, pp. 271–280. doi: 10.1016/j.procs.2019.09.182.

[21] A. Mohammed and R. Kora, "An effective ensemble deep learning framework for text classification," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 10, pp. 8825–8837, Nov. 2022, doi: 10.1016/j.jksuci.2021.11.001.

[22] S. Karlos, G. Kostopoulos, and S. Kotsiantis, "A soft-voting ensemble based co-training scheme using static selection for binary classification problems," Algorithms, vol. 13, no. 1, Jan. 2020, doi: 10.3390/a13010026.

[23] C. A. Gonçalves, A. S. Vieira, C. T. Gonçalves, R. Camacho, E. L. Iglesias, and L. B. Diz, "A Novel Multi-View Ensemble Learning Architecture to Improve the Structured Text Classification," Information (Switzerland), vol. 13, no. 6, Jun. 2022, doi: 10.3390/info13060283.

[24] F. Haghighi and H. Omranpour, "Stacking ensemble model of deep learning and its application to Persian/Arabic handwritten digits recognition," Knowl Based Syst, vol. 220, May 2021, doi: 10.1016/j.knosys.2021.106940.

[25] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," Expert Syst Appl, vol. 77, pp. 236–246, Jul. 2017, doi: 10.1016/j.eswa.2017.02.002.

[26] A. Onan, S. Korukoğlu, and H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification," Expert Syst Appl, vol. 62, pp. 1–16, Nov. 2016, doi: 10.1016/j.eswa.2016.06.005.

[27] Ankit and N. Saleena, "An Ensemble Classification System for Twitter Sentiment Analysis," in Procedia Computer Science, Elsevier B.V., 2018, pp. 937–946. doi: 10.1016/j.procs.2018.05.109.

[28] Universitas Gadjah Mada, Institute of Electrical and Electronics Engineers. Indonesia Section., and Institute of Electrical and Electronics Engineers, Proceedings, 2019 5th International Conference on Science and Technology (ICST) : 30-31, Eastparc Hotel, Yogyakarta, Indonesia. July 2019.

[29] L. Wang, L. Zhang, and J. Jiang, "Detecting Duplicate Questions in Stack Overflow via Deep Learning Approaches," in Proceedings - Asia-Pacific Software Engineering Conference, APSEC, Dec. 2019, vol. 2019-December, pp. 506–513. doi: 10.1109/APSEC48747.2019.00074.

[30] E. Dadashov, Elkhan, S. Sakshuwong, and K. Yu, "Quora Question Duplication," 2017, https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs.

[31] X. Zhang, X. Sun, and H. Wang, "Duplicate Question Identification by Integrating FrameNet With Neural Networks," Thirty-Second AAAI Conference on Artificial IntelligenceProceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[32] H. Zhang, and L. Chen, "Duplicate Question Detection based on Neural Networks and Multi-head Attention," International Conference on Asian Language Processing (IALP), pp. 13-18, 2019.

[33] Z. Imtiaz, M. Umer, M. Ahmad, S. Ullah, G. S. Choi,and A. Mehmood, "Duplicate Questions Pair Detection Using Siamese MaLSTM," IEEE Access, vol. 8, pp. 21932-21942, 2020.

[34] A. Chunamari, M. Yashas, A. Basu, D. K. Anirudh, , and C.S. Soumya,, "Quora question pairs using XG boost," Emerging Research in Computing, Information, Communication and Applications, pp. 715–721, 2021.

[35] H. T. Le, D. T. Cao, T. H. Bui, L. T. Luong and H. Q. Nguyen, "Improve Quora Question Pair Dataset for Question Similarity Task," RIVF International Conference on Computing and Communication Technologies (RIVF), pp. 1-5, 2021.

[36] T.G. Dietterich, "Ensemble Methods in Machine Learning. In: Multiple Classifier Systems," MCS Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, vol. 1857, 2000, https://doi.org/10.1007/3-540-45014-9_1.

[37] M. K. Elhadad, K. F. Li , and F. Gebali, "Detecting Misleading Information on COVID-19," IEEE Access, vol. 8, pp. 165201-165215, Sep.2020, doi: 10.1109/ACCESS.2020.3022867.

The Fusion Method of Virtual Reality Technology and 3D Movie Animation Design

Xiang Yuan, He Huixuan
Shijiazhuang University of Applied Technology, Animation Academy, Shijiazhuang City Hebei Province, 050031, China

***Abstract*—To further improve the design effect of 3D film and television animation, integrating virtual reality technology with 3D film and television animation design is studied. This method uses 3Ds Max software in virtual reality technology to build 3D film and television animation scenes by manual modeling. Based on the established 3D film and television animation scene, texture mapping is performed on it, and then the 3D film and television animation character model is established and simulated. After optimizing the established 3D scene and character model using the improved quadratic error measurement algorithm, the roaming interaction of 3D film and television animation scene is realized through Unity3D software, and the integration of virtual reality technology and 3D film and television animation design is realized. The experimental results indicate that the 3D film and television animation scene created using virtual reality technology is very realistic, which can effectively optimize the 3D film and television animation model. The number of path nodes is the least when the 3D film and television animation scene roams and interacts, which has a relatively significant application effect.**

Keywords*—Virtual reality technology; 3D film and television; Animation design; model optimization; roaming interaction

I. INTRODUCTION

3D animation technology is also known as 3D animation technology. 3D animation technology has recently advanced functions such as animation character modeling, material design adjustment, scene environment building, etc. [1]–[3]. It can use various visual communication techniques to reconstruct and simulate more complex spatial scenes, logical thinking, character representation, and other external forms and realize the scheme on the 3D software platform [4]–[6]. At present, virtual reality technology can be applied more and more widely [7], especially in the direction of virtual simulation modeling in 3D space. VR and 3D animation technology can be integrated through 3D integrated model reconstruction and virtual scene creation [8], [9]. In the next stage of the development of 3D modeling technology [10], virtual reality technology is bound to bring innovations and breakthroughs to 3D animation production technology. Therefore, integrating virtual reality technology and 3D film and television animation design will encourage researchers to move towards more in-depth technology integration products [11] and then break through the outdated 3D animation-producing technology to make 3D animation more authentic through virtual scene reproduction.

Abas A et al. [12] suggested a multi-scale transformation multi-focus image fusion method, which fused virtual reality technology and 3D images through a meta-heuristic optimization algorithm to achieve the integration of virtual reality technology and 3D film and television animation design. Mila B et al. [13] proposed an experimental method for social aspects of game virtual reality, which applies virtual reality technology to establish game virtual scenes and characters and realizes the integration of virtual reality technology and 3D film and television animation design. Tastan H et al. [14] proposed the method of building modeling using the handheld user interface and direct operation in immersive virtual reality. This method combines holistic virtual reality technology, portable user interface, and direct operation modeling technology to obtain 3D film and television animation scenes, realizing the integration of virtual reality technology and 3D film and television animation design. Jo, H et al. [15] proposed different visual environment reproduction methods and used this method to integrate virtual reality technology with 3D film and television animation design to obtain urban soundscape and landscape. Karagiannis P. et al. [16] proposed a fusion method using simulation and virtual reality technology. This method utilizes virtual reality technology to build scene and character models and then uses the simulation technology of 3D film and television animation to integrate the scene and characters to integrate virtual reality technology and 3D film and television animation design.

Virtual reality technology is an important direction of simulation technology, which is a combination of simulation technology, computer graphics, human-machine interface technology, multimedia technology, sensing technology, network technology, and other technologies. It is a challenging interdisciplinary and research field with cutting-edge technologies. 3D animation, also known as 3D animation, is not limited by time, space, location, conditions, or objects. It uses various forms of expression to present complex and abstract program content, scientific principles, abstract concepts, etc. in a concentrated, simplified, vivid, and vivid form. Based on the above analysis, it can be concluded that the current research question is how to further improve the effectiveness of 3D film and television animation design, as well as how to integrate virtual reality technology with 3D film and television animation design. Therefore, this paper proposes a fusion method of virtual reality technology and 3D film and television animation design. Through manual modeling, texture mapping, character simulation, and optimization processing, achieve more realistic, interactive, and optimized 3D film and television animation scenes. In order to provide more efficient, realistic, and interactive solutions for 3D film and television animation design, promote the development of virtual reality technology, and bring more diverse and immersive virtual

experiences to entertainment, education, cultural and artistic fields.

II. Comparative Analysis

Compared with similar works mentioned in the introduction, this paper aims to further enhance the design effect of 3D film and television animation, and explore the integration method of virtual reality technology and 3D film and television animation design. Below, a fair comparison will be made between this work and similar previous works, highlighting the advantages of this work:

1) Method innovation: This paper uses 3D Max software from virtual reality technology to manually model 3D film and television animation scenes. Unlike the multi-scale transformation and multi focus image fusion method proposed by Abas et al. [12], this method applies virtual reality technology to the modeling process in a more direct way, thereby improving the realism and detail representation of the scene.

2) Texture mapping processing: In this work, texture mapping processing was applied to the established 3D film and television animation scene. This step makes the scene more realistic and visually appealing. In contrast, Mila et al. [13] did not extensively explore the application of texture mapping processing in the experimental methods of social aspects in game virtual reality.

3) Simulation processing: This paper establishes a 3D film and television animation character model and performs simulation processing on the model. By simulating the actions and behaviors of characters, the authenticity and realism of the characters are enhanced. In contrast, the method proposed by Karagiannis et al. [16] combines simulation and virtual reality technology to establish scene and character models, but does not explicitly mention the simulation processing of character models.

4) Improved optimization algorithm: This paper uses an improved quadratic error measurement algorithm to optimize the established 3D scene and character models. This algorithm helps to improve the accuracy of the scene and the appearance of the characters, thereby enhancing the quality of 3D film and television animation. In contrast, the method proposed by Tastan et al. [14] used handheld user interfaces and direct operations for building modeling, but did not provide a detailed description of the optimization algorithms used.

In summary, compared with previous similar works, this work has innovation in exploring the integration process of virtual reality technology and 3D film and television animation design. By using 3Ds Max software for modeling, texture mapping, simulation, and improved optimization algorithms, this work successfully enhances the realism and visual effects of 3D film and television animation, bringing new possibilities for the integration of virtual reality technology and 3D film and television animation design.

III. Integration Design of 3D Film, Television, and Animation

A. Integrated Technical Architecture

The realization of virtual reality technology is a foundation for 3D film and television animation designs [17]. The essence of 3D animation is a sense of stereo vision generated by the continuous projection of graphics and images. Experiments can transform the imagination in their minds into 3D animation models and realize model creation and visual optimization through the 3D modeling software platform. Virtual reality technology is gradually extended and developed based on three-dimensional animation technology. Its advanced scene reconstruction ability initially needs three-dimensional models to build. The prerequisite for the final realization of virtual reality space scenes is the adaptive transformation ability of three-dimensional images. Experiments can conduct interactive feedback of virtual reality scenes by upgrading 3D animation technology. Therefore, virtual reality technology and 3D animation technology come from a school of thought design [18]. After the development of 3D animation technology, it is called virtual reality technology. However, the interactivity of 3D animation technology stays in the passive information acceptance of the experimenter. The interactivity of virtual reality technology can make the system cooperate and deepen the scene according to the behavior logic of the experimenter and convey the space scene simulated by computer data artistically. 3D animation technology is the fundament of producing virtual reality technology data. Here, the technical architecture for integrating virtual reality technology and 3D film animation has been designed and shown in Fig. 1.

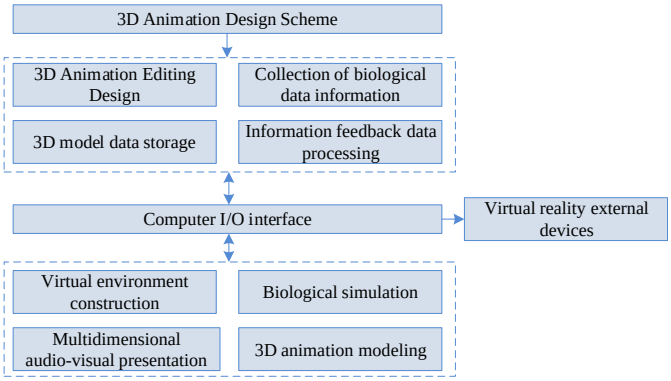


Fig. 1. Architecture of virtual reality technology and 3d film and television animation design integration technology.

The integration technology of virtual reality technology and 3D film and television animation design gets 3D animation pictures of virtual reality through virtual environment construction, multi-dimensional audio-visual presentation, biological simulation, and 3D animation modeling, and then connects 3D animation editing design, biological data information collection, 3D model data storage and information data feedback processing through computer I/O interface, realize the integration of virtual reality technology and 3D film and television animation according to 3D animation design scheme and virtual reality external equipment.

B. Manual Modeling Method of Animation Scene Based on 3Ds Max

The first step of basic 3D animation production is 3D scene modeling, and a 3D model scene is the golden key to open animation production. As a model designer, it is necessary to determine various modeling data of 3D scenes. For some indoor scenes with small space areas, it is relatively easy to obtain basic measurement data. Obtaining more reliable scene data is impossible for virtual scenes with large space and unlimited space. Now, it is essential to collect virtualization information data according to the previously simulated spatial scene data samples [19], reconstruct the scene model with the help of powerful computer data model processing ability, and complete this spatial scene modeling work with the highest efficiency through this virtual scene simulation method. This method enlarges and outputs the scene information through virtual reality, efficiently completes the reconstruction of a 3D model scene, and improves the accuracy of scene model data through virtual reality assistance to create a more realistic virtual world.

1) Construction of 3D film and television animation scene model: This section will introduce the processes of producing complex 3D film and television animation scene models in 3Ds Max [20] and final 3D film and television animation scene models. The technical process of 3D animation 3D Max model construction is illustrated in Fig. 2.

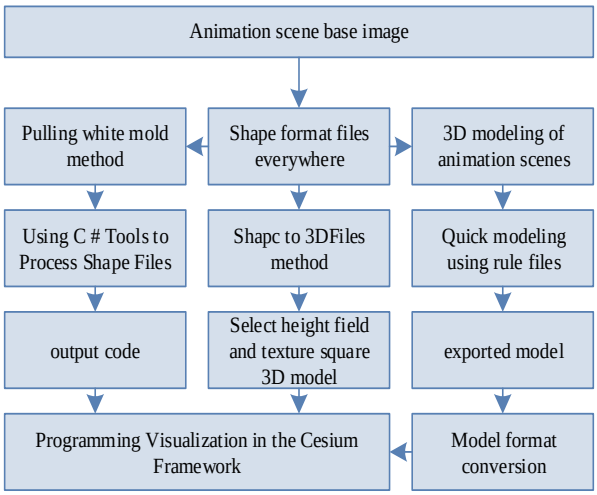


Fig. 2. 3D animation 3D max model construction technology process.

The production of 3D film and television animation scenes is the most difficult part of 3D software. The concept of time dimension is added in the production process, and almost any object or parameter can be animated in 3Ds Max [21]. Import the designed virtual scene base map or real image scene base map into the 3Ds Max software, and export the shape format file through 3Ds Max. Based on the cut file, you can quickly create 3D film and television animation scenes using rule files, export 3D film and television animation scenes and convert the model format, convert the shape format file into 3D Tiles method, select the elevation to generate 3D film and television animation scene, use the pull white mold method, process the shape format file through C # tool and output the code, and then program the code, animation scene model and format converted animation scene model in the Cesium framework to achieve 3D animation 3Ds Max model construction.

2) Realization of 3D film and television animation scene; After the 3Ds Max model is established, 3D Exploration, Wcvt2pov, and other software can convert the 3Ds Max model into the corresponding OpenGL C C++ format file, or the program can directly load the 3Ds Max model. The software format conversion method can preserve the model's color, texture, and other information [22]. However, some limitations are that software support and manual interaction are required for model transformation, and only one model can be transformed at a time. Therefore, the efficiency could be higher. The method of program direct loading can freely control the model to be loaded and promote the efficacy of program operation. After the 3DS model is loaded into the OpenGL program, the corresponding scaling, rotation, movement, and other controls are also required. The detailed process is as follows:

a) Read 3Ds Max models. 3Ds Max files are organized in a block structure, and there is a nesting relationship between blocks. Therefore, the 3D model data in the 3Ds Max file is read from a block in the model file, and the functional information of the block is judged according to the block ID, and then the corresponding processing is performed according to the block ID. A sub-block is read in the block processing process, such as the main editing block; block information is judged, and the corresponding data is extracted and stored.

b) Build model display list. When writing OpenGL programs, if you encounter repeated work, you can create a display list, load the repeated work, and call the display list where necessary. There are generally four steps to using the display list: assigning the display list number, creating the display list, calling the display list, and destroying the display list.

c) Minimum and maximum according to 3Ds Max model x, y the coordinates and the actual required width and height of the 3D scene are used to calculate their scaling ratio and scale the model. The functions used are: $glScalef(GLfloat\ x, GLfloat\ y, GLfloat\ z)$. Multiply the current matrix by a matrix indicating the scaled object. x, y and z refer to the scaling in corresponding directions, respectively.

d) Rotate the model according to the direction angle of the model in the 3D film and television animation scene. The functions used are: $glRotatef(GLfloat\ x, GLfloat\ y, GLfloat\ z)$. Multiply the current matrix by a matrix indicating a rotating object. The object will reach around (0,0,0) reach (x, y, z) the line of rotates counterclockwise, parameter $angle$ represents the rotation angle.

e) Calculate the model according to the position of the 3D movie animation scene model in the 3D scene x, y, and z

translate the 3D movie animation scene model. The functions used are:

$glTranslatef(GLfloat\ x, GLfloat\ y, GLfloat\ z)$ Multi ply the current matrix and a matrix indicating moving objects. These three parameters indicate the displacement values in three coordinates, respectively.

f) Realize the placement of 3Ds Max models in the 3D scene according to the corresponding size, direction, and position, call the display list, and complete the drawing of 3D movie animation scene models.

C. Texture Mapping Processing

After the establishment of 3D film and television animation scene models, to make the scene more realistic, it is necessary to conduct texture mapping processing on it. Before texture mapping, select the contour of a 3D film and television animation scene [23]. Here, a cubic uniform B-spline curve extracts the contour of 3D film and television animation scenes. The expression formula of the curve parameter equation is as follows:

$$C(u,v) = [x(u), y(u), z(u)] \quad (1)$$

In the above formula, $C(u,v)$ represent cubic uniform B-spline curve equation; u and v are parameters; $x(u)$, $y(u)$, $z(u)$ all are cubic B-spline basis functions.

The cubic uniform B-spline curve and expression formula are as follows:

$$r_i(u) = C(u,v)\sum_{j=0}^{3} N_{j,3}(u) \quad (2)$$

In Eq. (2), $r_i(u)$ represents the i th sum of three cubic uniform B-spline curves; $N_{j,3}$ represents the total number of cubic uniform B-spline curves.

After the contour line is drawn, rotate the selected rotational axis to build the model [24]. The right-hand coordinate system is used in texture mapping, x the positive direction of the shaft is to the right and y the positive direction of the axis is upward, so the rotation axis is selected as y shaft. Therefore, the parametric equation of the surface after rotation is:

$$\begin{cases} x = x(u) * sin(v) \\ y = y(u) \\ z = x(u) * cos(v) \end{cases} \quad (3)$$

For the gridding of 3D film and television animation scene models, the unit length of the grid will be intercepted according to the curvature of each point of the drawn contour line, and the grid will be further subdivided where the curvature is large. The curvature solution formula is as follows:

$$k = \frac{r_i(u)|\ddot{r}(s)|}{x^2+y^2+z^2} \quad (4)$$

In the above formula s indicates the arc length parameter; t represents a non-arc length parameter; $\ddot{r}(s)$ means right $r(s)$ perform a second derivative.

After the above steps, the mesh of the 3D movie and television animation scene model is more reasonable, and the resulting mesh quadrilateral tends to be the non-planar area of the original model, which is more conducive to rendering in the texture mapping phase.

Texture mapping technology can simulate the fine and irregular color texture on the surface of people (objects). The texture mapping technology can cover any planar figure (image) on the 3D animation model's surface [25] so that the model surface can produce a more realistic color texture, enhance the authenticity of 3D animation, and facilitate the modeling processes. There are two steps in texture mapping technology. The first step is to determine the texture attribute, determining which part of the person (object) surface parameters must be set as the texture shape. The second step is to create a mapping relation between the texture space and the person (object) space and a mapping relation between the person (object) space and the screen space. Fig. 3 shows the mapping definition.

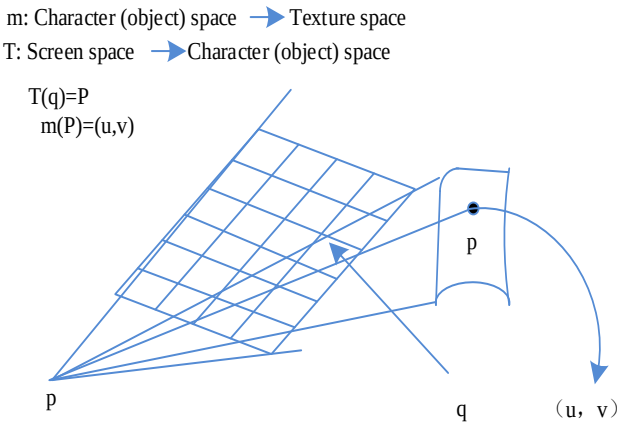


Fig. 3. Mapping definition.

Just basic contour characteristics, i.e., lack of surface texture details, can decrease the realism. It needs to map each module's texture to raise the model's realism.

The general direction of the surface patch specifies which space plane the surface patch projects to and projects to the plane with the smallest angle between the whole direction of the surface patch and the plane. As some parts of the person (object) surface are curved, it is necessary to calculate the overall direction of the curved surface. The computation method can be summarized as follows:

Assume that v_1, v_2, v_3 are three vertices of a triangular patch and cross product $[v_1 - v_2] * [v_2 - v_3]$ is perpendicular to the patch and normalized to get the normal vector of the triangular patch V the average vector sum of the module can be obtained by averaging the sum of all patch normal vectors, and then the general direction of the module can be obtained. The equation is as follows:

$$C = \sum_{i=1}^{PolyNum} \frac{v_i k}{PolyNum} \quad (5)$$

Where, $PolyNum$ indicates the total number of vertices.

To study the mapping relation between the surface patch and its related texture coordinates, the texture coordinates of

the grid points is computed through the perspective projection transformation, and the formula is as follows:

$$V = kX' = CHX \tag{6}$$

In which h_1, h_2, h_3, h_4, h_5 belongs to an unknown parameter, and (u, v) is the texture coordinate, the corresponding homogeneous texture coordinate is X', grid point space homogeneous coordinate is X, the perspective projection matrix of $3 * 4$ is H and the constant coefficient is k. Based on the Eq. (5) and Eq. (6), two independent linear equations exist for each group of corresponding mesh and texture vertices. To transform the matrix H, selecting three sets of feature points is necessary.

Analyzing all texture mapping processes, and setting the texture image size to $m * n$ simplify Eq. (7), and the process is as follows:

$$H = \begin{bmatrix} h_1 & 0 & 0 & h_4 \\ 0 & h_2 & 0 & h_5 \\ 0 & 0 & h_3 & 1 \end{bmatrix} \tag{7}$$

where, s and t is an unknown parameter, (x, y, z) is the surface vertex coordinate.

For Eq. (8), only a set of characteristic points can be used to calculate the unknown parameters. When extracting textures, all textures are made into the smallest bounding box. There are many tangent points at the edge of the image [26], which is very easy to get a group of feature points. Set texture coordinates(u', v') and surface vertex coordinates$(x', y', z')$is a group of characteristic points, which can be obtained by substituting into Eq. (8):

$$H' = \begin{bmatrix} 1/m & 0 & 0 \\ 0 & 1/n & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & s \\ 0 & 1 & 0 & t \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x & y & z \end{bmatrix}^T \tag{8}$$

$$F = H'V \tag{9}$$

After the above steps, the texture mapping of the 3D movie animation scene model is completed.

D. Character Modeling of 3D Film and Television Animation

According to the 3D animation character model and action library, users should just choose their required role prototype, modify the parameters like the height and body proportion of the role prototype, and get the 3D animation target character. The 3D animation target character matches the prototype 3D animation character's skeleton, mask, controller, and original actions in the action library. Fig. 4 is the composition diagram of the 3D animation character.

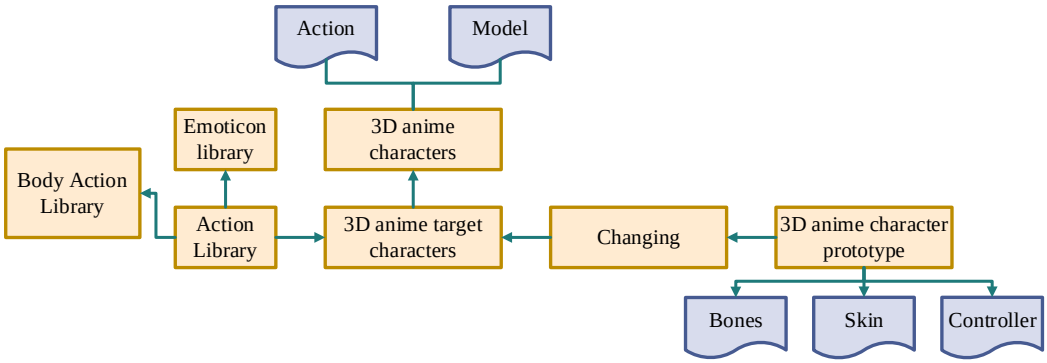


Fig. 4. Composition of 3D anime characters.

Components of a 3D animation character template include a virtual skeleton, skin, and controller, and it is a prototype for 3D animation. A set of basic actions allocated for 3D animation character prototypes is the 3D animation character action library. Users can freely modify the proportions and shapes of 3D animation characters based on the graphical interface, design the 3D animation character they want, or create new 3D animation characters.

E. 3D Animation Character Simulation

Motion capture technology is used to store many motion instances of character motion in the concerned motion library. The attained motion instances are represented in 3D form through virtual artificial synthesis software so that the 3D animation system is visual and convenient to change the action of the character model. The Newtonian Euler motion model is used to verify whether the new action studied is reasonable.

The character action is set as motion (t) by the 3D animation system, and the character's original pose (ti) is changed to gain a new character pose (ti). These operations lead to visual interactive action design. The user window is $O \times G$ in size. By moving the character model with the mouse, the variation in direction f will be Δf and the change in direction g will be Δg. According to Euler's theorem, for $< \alpha, \beta, \gamma >as$ representatives of d, f and g, the relationship of direction rotation after reasoning is as follows:

$$\begin{aligned} \sin\alpha &= aF\Delta f/[O(1-a)\Delta g/G] \\ \sin\beta &= bF\Delta f/[O(1-b)\Delta g/G] \\ \sin\gamma &= cF\Delta f/[O(1-c)\Delta g/G] \end{aligned} \tag{10}$$

In Eq. (10), the influencing factors are a, b and c representing d, f, g in three directions. Δf and Δg are opposite Euler angles and $< \alpha, \beta \; and \; \gamma >$ are the degree of

influence. Calculations achieve the newly pose of the character model, and the 3D animation character model is simulated.

F. 3D Animation Model Optimization Method based on Improved Quadratic Error Measure Algorithm

The edge collapse algorithm, also known as the Quadric Error Metric (QEM) algorithm, is not only fast in running speed, low in memory consumption, and simple in calculation but also has a very high overall similarity between the simplified mesh and the model. This paper improves the original quadratic error measure algorithm (QEM) and uses the improved QEM algorithm to optimize the scene modeling.

The specific steps are as follows:

1) Read in the original mesh data, which contains the vertex coordinates in the mesh and the vertex, sequence and other data information of the connected triangular patch.

2) Find the discrete curvature of each vertex in the mesh $K(h)$ the area of the local area is obtained from the formula $LRA(h)$, and compare the two weighting factors with the matrix $Q(h)$ weighting to obtain the quadratic error matrix of each vertex in the network $Q'(h)$:

$$Q'(h) = \frac{K(L_h)K_H}{\sin\alpha \sin\beta \sin\gamma} \tag{11}$$

where, $K(L_h)$ include vertices for Gaussian curvature of the edge of h,K_H is Gaussian curvature of a vertex h.

The expression of K_H is:

$$K_H = 2\pi - \frac{\sum_n \theta_h}{S_{\sum}(h)} \tag{12}$$

where, θ_h is an adjacency angle of vertex v;n is the number of triangles associated with vertex h; $S_{\sum}(h)$ is a vertex h the sum of adjacent triangular mesh areas, that is the local area.

3) The appropriate threshold through which the input is low $w = 0.9$, determines whether it is a boundary. If it is a boundary, set $\Delta(h) = inf$.

4) Get the quadratic error matrix of each vertex according to step $2Q'(h)$. To calculate the shrinkage cost of each edge ζ, if $(h_1, h_2) \to \tilde{h}$ the shrink cost is put on the stack, the higher the value of ζ, the more backward it is in the stack, and on the contrary, the more forward it is.

5) Take the edge with the lowest shrink cost from the stack to perform the shrink operation, that is, the top of the stack. Simultaneously, update the data structure of the network, calculate the new edge collapse cost, and update the stack sequence.

6) If the ideal simplification requirements are met, the algorithm ends; otherwise, the above process continues until the simplification requirements are met.

After the above steps, the optimization of the 3D animation model is realized, and the resulting film and television animation screen is more fluent, and the structure is more reasonable.

G. Interaction Mode of 3D Film and Television Animation Scene based on Unity3D

Unity3D software is designed with the interaction mode of 3D film and television animation scenes based on the improved A * algorithm. The improved A * algorithm is used to plan the interaction path in the 3D film and television animation walkthrough, obtain the optimal interaction path, and avoid collision in the interactive display process of the 3D film and television animation walkthrough.

The key part of the A * algorithm is the evaluation function $f'(n')$, that is, the source point in the 3D movie animation model p_0 to end p_{end} passing through nodes n' the total cost of $B(n')$ and estimated cost $h'(n')$ add to get.

The collective steps of the A * algorithm to search for the best path in the 3D movie animation model are summarized as follows:

Step 1: Initialize the 3D movie animation model data and add it to the open table of the open list p_0, select in open table $f'$the node with the lowest value is regarded as the current node p_1, if none p_1, then the 3D movie animation model interactive display path search fails, and the path search ends; if $p_1$is p_{end}, then the 3D film and television animation model interactive shows that the path search is successful, and the path search ends; With p_1 by $p_{end}$the path from the source node to the destination in the interactive presentation process of 3D film and television animation model is the search path.

Step 2: In the closed table, add p_1, traversal p_1 all adjacent nodes of p', if close-table exists p', no processing is required; If the close-table does not exist p', you need to solve again the value of $B(n')$, $h'(n')$, set $O_1$by $p'$and add it in the open-table p'; if p'in open-table, comparative analysis p_1 to p' of $B(n')$ is the value below p' of $B(n')$ value, if lower than p' of $B(n')$ value, then no processing is required. If the value exceeds p' of $B(n')$ value, then p_1 to p' of $B(n')$ value is changed to p' of $B(n')$ value with the predecessor node of p_1 by p'.

Step 3: Repeat step 2 to process the remaining adjacent nodes in order until the open table is empty.

To speed up the search efficiency of the interactive display path of 3D film and television animation roaming, a hierarchical strategy is introduced into the A * algorithm, and the 3D film and television animation model is regarded as a complete map divided into several areas of consistent size. For the obstacles within the boundary line, it is necessary to remove the obstacles first, then solve the scale of the obstacles within the boundary line, add the obstacles within the boundary line in the area with the highest scale, and complete the area division of 3D movie and television animation model in this way.

By setting parameters E and R, and designing fixed rules, the search algorithm is obtained and used in each area, and the proportion of obstacles in the 3D film and television animation model, is E which is the ratio of the total number of nodes occupied by obstacles in the current area to the total number of nodes in the area, and the threshold coefficient is R.

Adopt E determine that a certain search algorithm needs to be used in this area and determine that further subdivision is required; The value R has a negative correlation with the area division size, a positive correlation with the number of expansion nodes, and a positive correlation with the path search time. According to the relationship between E and R to establish the following rules:

1) On $E = 0$, it represents no obstacle in the area, and the shortest optimal path can be obtained directly according to the straight line between two points. If p_0 and p_{end} both are in this area, then this path is the optimal path; if p_0 and p_{end} are not in the same area, search for the middle point between the optimal path and the boundary line of the area in the area and regard the middle point as that of the next area p_0.

2) On $E < R$, means there are fewer obstacles in the area, and the hierarchical strategy is still adopted to subdivide the area further.

3) On $E \geq R$, the number of representative obstacles is consistent with the threshold, or even exceeds the threshold. In this case, the A * algorithm is directly used to search the path.

According to the above rules, searching the path in the interactive display process of 3D film and television animation roaming can effectively divide the big map into small maps, which is conducive to processing, selecting the optimal path search method for small maps, speeding up the path search efficiency, and effectively avoiding collision events in the interactive display process of 3D film and television animation roaming.

IV. Experimental Analyses

Taking an animation design scheme as the experimental object, the method in this paper is used to integrate virtual reality technology and 3D film and television animation and present the animation design scheme to validate the practical application impact of the method proposed in the present paper.

Take a scene of the animation design scheme as the experimental object, and present the scene using the method in the current investigation. The results are illustrated in Fig. 5.

It is clear from the analysis of Fig. 5 that the application of the method used in the present paper can effectively establish the 3D film and television animation virtual scene. From the 3D film and television animation virtual scene established in this paper, it can be seen that the color of trees, grass, and flowers is real, and the three-dimensional effect is good. The water body can not only map the scene in the sky but also show a clear effect, and the river bottom can be seen through the water body. These results indicate that the 3D film and television animation virtual scene built by this method has good stereoscopic and real effects.

The texture coordinate calculation results are used to measure the texture mapping effect of this method on 3D film and television animation scenes. With 15 texture coordinate points as experimental objects, this method calculates the texture coordinates. The results are presented in Table I.

Fig. 5. Virtual scene of 3D film and television animation.

TABLE I. Texture Coordinates of 3D Film and Television Animation Scene (cm)

Texture coordinate encoding	Calculated value		Actual value	
	X-direction	*Y-direction*	*X-direction*	*Y-direction*
1	3.24	8.54	3.24	8.54
2	4.06	10.34	4.06	10.35
3	1.58	2.88	1.58	2.88
4	2.79	9.17	2.79	9.17
5	11.07	5.97	11.07	5.97
6	25.13	14.96	25.13	14.96
7	10.25	11.85	10.25	11.85
8	6.98	8.97	6.98	8.97
9	8.94	16.34	8.94	16.34
10	15.27	16.85	15.28	16.85
11	30.17	22.13	30.17	22.13
12	2.89	6.84	2.89	6.84
13	16.63	20.57	16.63	20.57
14	13.79	19.82	13.79	19.82
15	21.11	18.84	21.11	18.84

From the analysis of Table I, it can be seen that the texture coordinates of 3D film and television animation scene texture mapping are calculated using the method in the present research. In the calculation results, only the texture coordinates coded as 10 and 2 have deviations from the actual results in the X and Y directions, but the deviation value is only 0.01cm, which is small. Other texture coordinate values calculated are identical to the actual coordinate values. These results imply that the texture coordinates of 3D film and television animation scene texture mapping calculated by the method in this paper are more accurate, and it has a strong ability for 3D film and television animation scene texture mapping.

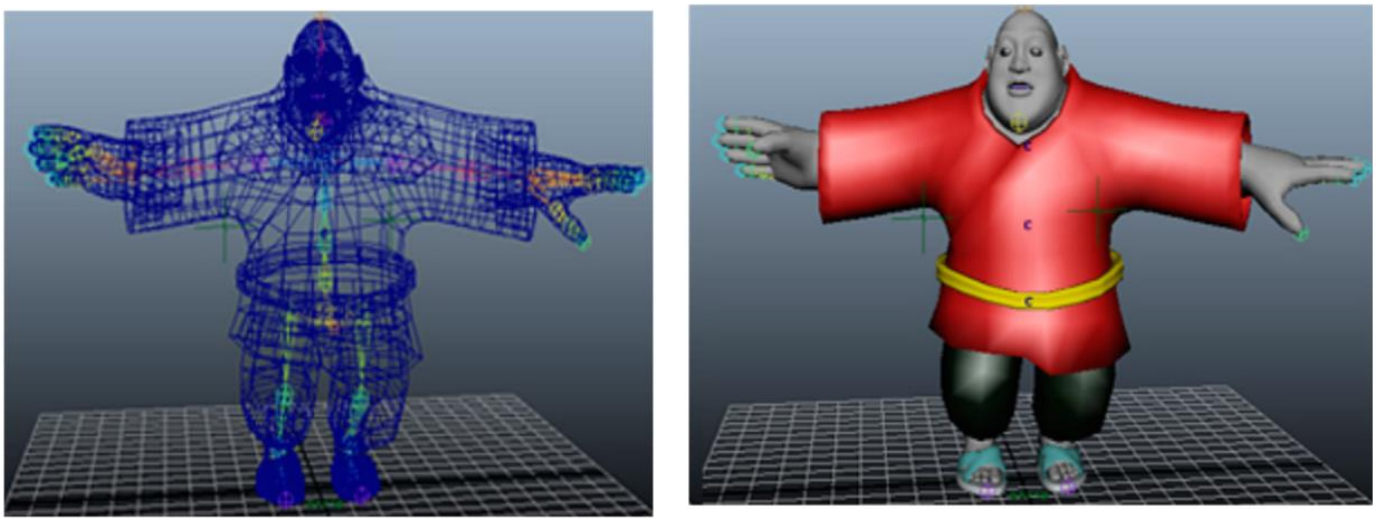
To further verify the texture mapping capability of this method, take a 3D movie animation scene as the experimental object and use this method to conduct texture mapping processing. The texture mapping results are shown in Fig. 6.

(a) Before mapping (b) After mapping

Fig. 6. Texture mapping effect in 3D film and television animation scenes.

According to the analysis of Fig. 6, after applying the method in this paper to texture map the 3D film and television animation scene, the entire 3D film and television animation scene has bright colors, and the texture map position is not prominent or concave, indicating that the method of the current paper has better texture mapping effect on 3D film and television animation scene.

Take an animated character in the animation design scheme as the experimental object, use the method in this paper to establish its virtual character model, and the establishment results are shown in Fig. 7.



(a) Design base drawings (b) Achievement

Fig. 7. Animated character virtual character model.

It is clear from the analysis of Fig. 7 that this method can effectively establish the 3D model of the animated character virtual character based on the base map of the animated character virtual character. The 3D model of the animated character virtual character established has a good three-dimensional effect, and the character's facial expression is more real. To sum up, the method in this paper can better build virtual character models of animated characters.

Taking a 3D movie animation scene as the experimental object, the method in this paper is used to optimize the 3D movie animation scene, and the optimization results are illustrated in Fig. 8.

It is evident from the analysis of Fig. 8 that after the optimization of 3D film and television animation scenes using the method in this paper, the contrast of 3D film and television animation scenes has been improved, and the clarity of the entire 3D film and television animation scenes has also been effectively improved, making the visual effect of 3D film and television animation scenes better. In conclusion, the present research method can effectively optimize 3D film and television animation scenes and has a relatively significant application effect.

Verify the roaming ability of the 3D film and television animation scene established by this method, and test the

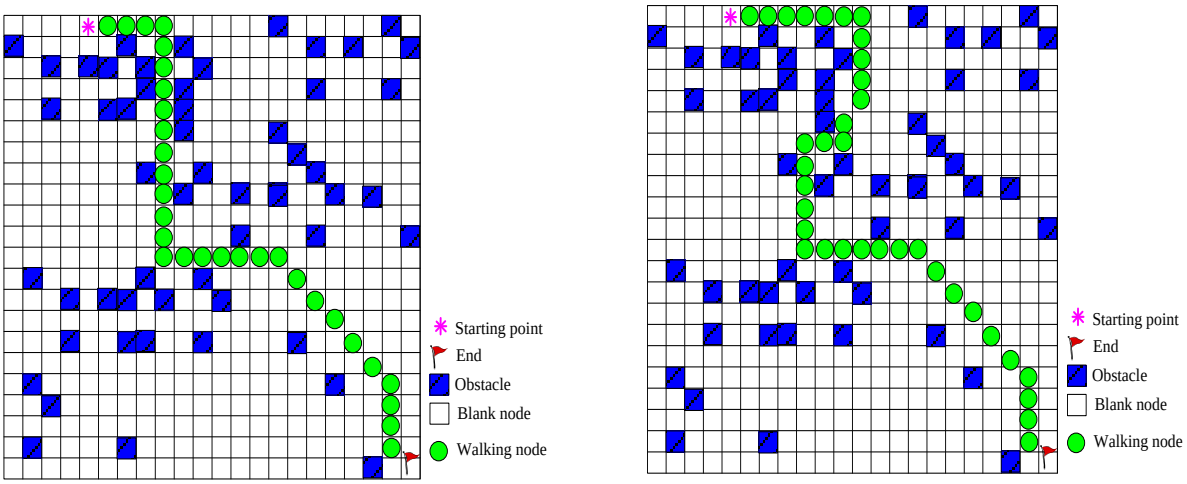
roaming ability of the 3D film and television animation scene established by this method with the roaming path as the measurement index. To make the experimental results more sufficient, the methods utilized in the references [12]- [16] are used to experiment. The experimental results are illustrated in Fig. 9.

It is obvious from the comprehensive analysis of Fig. 9 that when the method in this paper is used to roam and interact with 3D film and television animation scenes, the shortest roaming path selected in this method is only 30 nodes, while the path nodes selected in other reference methods when roaming and interacting with 3D film and television animation scenes are higher than those of the method proposed in the current study. The results show that the method in this paper can effectively combine virtual reality technology with 3D film and television animation design and select the shortest path when roaming 3D film and television animation scenes, which has a strong application effect.

(a) Before optimization

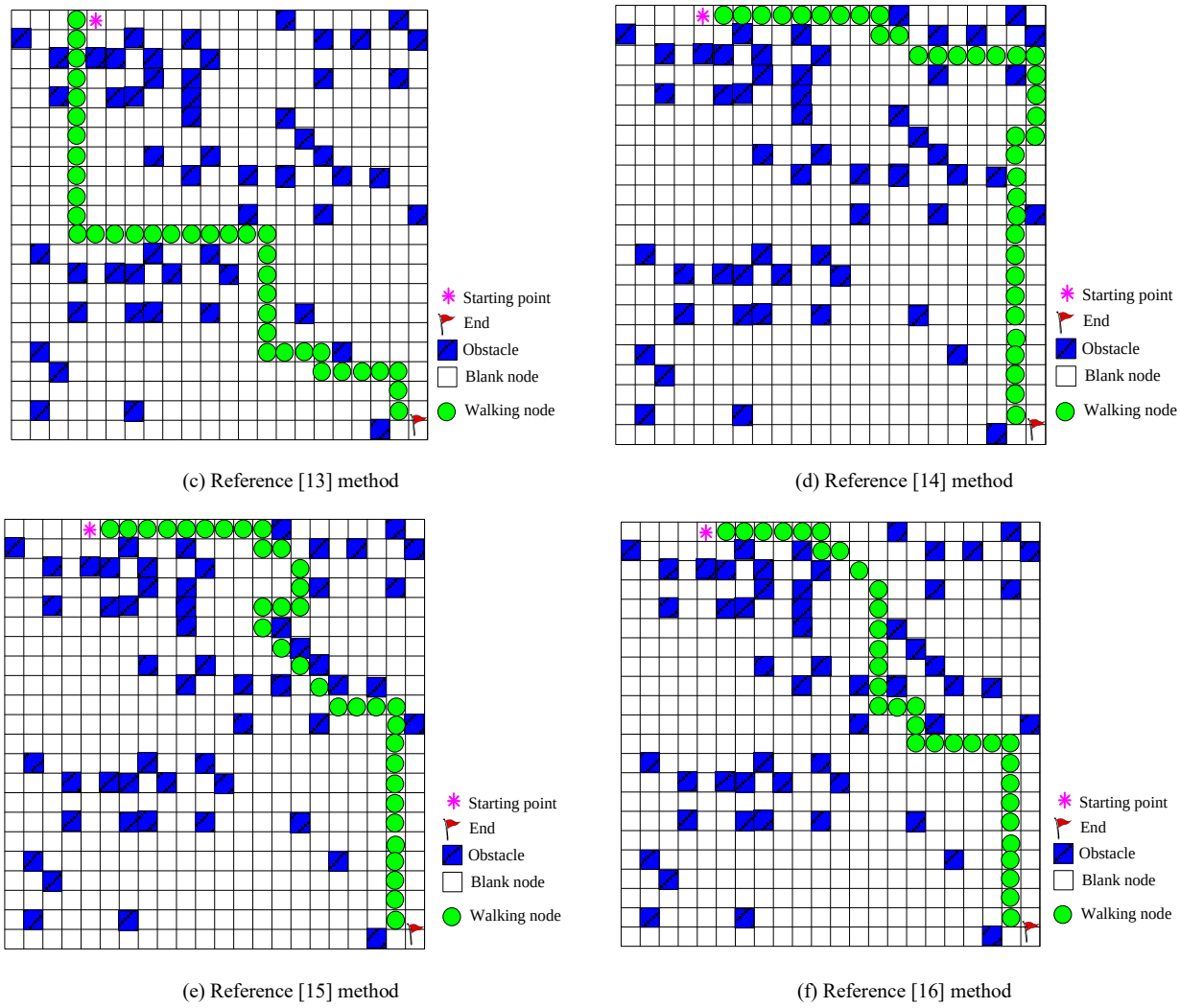
(b) After optimization

Fig. 8. Optimization results of 3d film and television animation scenes.



(a) Proposed method

(b) Reference [12] method



(c) Reference [13] method

(d) Reference [14] method

(e) Reference [15] method

(f) Reference [16] method

Fig. 9. Interactive performance test results of 3d film and television animation scene roaming.

V. Discussion

This paper successfully establishes a realistic and interactive 3D film and television animation virtual scene by studying the integration method of virtual reality technology and 3D film and television animation design. According to the above theories and experiments, it can be concluded that:

By analyzing Fig. 5, it can be seen that the 3D film and television animation scene established in this paper exhibits more realistic colors and three-dimensional effects in areas such as trees, grasslands, and flowers. In addition, water bodies not only reflect scenes in the sky, but also present a clear effect, through which the riverbed can be seen. These results indicate that the 3D film and television animation virtual scene successfully established by the method proposed in this paper has good stereoscopic and realistic effects.

The method used in this paper has high accuracy in calculating texture maps for 3D film and television animation scenes. Most texture coordinates are exactly the same as the actual coordinates, while only a few texture coordinates have slight deviations in the X and Y directions. However, these deviations are only 0.01cm and belong to smaller values. Therefore, the method proposed in this paper demonstrates high accuracy and capability in calculating texture maps for 3D film and television animation scenes.

After applying the method described in this paper to texture map 3D film and television animation scenes, the entire scene presents bright colors and uniform texture map positions, without any protrusions or depressions. This indicates that the method proposed in this paper has shown good performance in texture mapping of 3D film and television animation scenes. Further analysis of Fig. 7 reveals that based on the virtual character base map of animated characters, our method has successfully established a three-dimensional model with good stereoscopic effects and real facial expressions. This indicates that the method presented in this paper demonstrates excellent ability in establishing virtual character models for animated characters.

After optimizing the 3D film and television animation scene using the method described in this paper, the contrast of the scene is improved, and the clarity of the entire scene is also significantly improved. This makes the visual effects of 3D

film and television animation scenes more outstanding. In summary, the method proposed in this paper has significant application effects in optimizing 3D film and television animation scenes. For the roaming interaction of 3D film and television animation scenes, the method selected in this paper has the least number of roaming path nodes, only 30 nodes, which is significantly less than the number of path nodes selected by other literature methods. This indicates that the method proposed in this paper has shown significant advantages in integrating virtual reality technology and 3D film and television animation design, and has chosen the shortest path in the roaming of 3D film and television animation scenes, demonstrating strong application effects.

In summary, through the analysis of various experimental results, it can be concluded that the fusion method of virtual reality technology and 3D film and television animation design proposed in this paper demonstrates excellent capabilities and effects in the establishment of 3D film and television animation virtual scenes, texture mapping, character simulation, optimization processing, and roaming interaction. This is of great significance for promoting the development of 3D film and television animation design, improving user experience, and expanding the application fields of virtual reality technology. Future research can further optimize and expand this method to provide a more satisfactory virtual reality experience.

VI. Conclusion

The continuous integration and complementation of 3D animation technology and virtual reality technology can bring new breakthroughs to the two technical fields and make this virtual 3D animation system exert the best guiding force. The integration of virtual reality and 3D animation technologies makes the expression of 3D animation more diversified. Through the technical support of virtual reality technology, 3D animation can bring more real resonance of literary and artistic ideas to the experience. Therefore, integrating virtual reality technology with 3D animation technology is an inevitable trend, which can bring more opportunities for the future development of both technologies.

Considering that real-time rendering is crucial in virtual reality technology. How to improve rendering speed and efficiency while maintaining high-quality rendering, and ensure smooth operation on virtual reality devices, is an unresolved issue. Therefore, in the future, new real-time rendering algorithms and hardware acceleration technologies will be explored to achieve higher quality graphics rendering and faster rendering speed, providing users with a smoother virtual reality experience.

Data Availability

On Request

Competing of interests

The authors declare no competing of interests.

Authorship Contribution Statement

He Huixuan: Writing-Original draft preparation, Conceptualization, Supervision, Project administration.

Xiang Yuan: Methodology, Software, Validation

Declarations

Not applicable

References

[1] W. Bao, “The application of intelligent algorithms in the animation design of 3D graphics engines,” International Journal of Gaming and Computer-Mediated Simulations (IJGCMS), vol. 13, no. 2, pp. 26–37, 2021.

[2] L. Li and T. Li, “Animation of virtual medical system under the background of virtual reality technology,” Comput Intell, vol. 38, no. 1, pp. 88–105, 2022.

[3] L. Kumarapu and P. Mukherjee, “Animepose: Multi-person 3d pose estimation and animation,” Pattern Recognit Lett, vol. 147, pp. 16–24, 2021.

[4] F. Ying and Z. Bo, “Building virtual scene construction and environmental impact analysis based on image processing,” Sci Program, vol. 2021, pp. 1–14, 2021.

[5] W. J. Dally, S. W. Keckler, and D. B. Kirk, “Evolution of the graphics processing unit (GPU),” IEEE Micro, vol. 41, no. 6, pp. 42–51, 2021.

[6] Y. Wang, Y. Wang, and X. Lang, “Applied research on real-time film and television animation virtual shooting for multiplayer action capture technology based on optical positioning and inertial attitude sensing technology,” J Electron Imaging, vol. 30, no. 3, p. 31207, 2021.

[7] F. Reer, L.-O. Wehden, R. Janzik, W. Y. Tang, and T. Quandt, “Virtual reality technology and game enjoyment: The contributions of natural mapping and need satisfaction,” Comput Human Behav, vol. 132, p. 107242, 2022.

[8] X. Li, J. Ling, Y. Shen, T. Lu, and H. Zhu, “Effect of color temperature of light source in tunnel on driving safety based on virtual reality technology,” Tongji Daxue Xuebao/Journal of Tongji University, vol. 49, no. 2, pp. 204–210, 2021.

[9] S. H. Farhang, O. Rezaifar, M. K. Sharbatdar, and A. Ahmadyfard, “Rapid estimation method for vertical deflection measurement of building components based on image processing techniques,” Journal of Performance of Constructed Facilities, vol. 35, no. 1, p. 04020137, 2021.

[10] D. Li, J. Liu, Y. Zeng, G. Cheng, B. Dong, and Y. F. Chen, “3D model-based scan planning for space frame structures considering site conditions,” Autom Constr, vol. 140, p. 104363, 2022.

[11] P. Dymora, B. Kowal, M. Mazurek, and S. Romana, “The effects of Virtual Reality technology application in the aircraft pilot training process,” in IOP conference series: materials science and engineering, IOP Publishing, 2021, p. 012099.

[12] A. I. Abas and N. A. Baykan, “Multi-focus image fusion with multi-scale transform optimized by metaheuristic algorithms,” Traitement du Signal, 2021.

[13] M. Bujić, A.-L. Macey, S. Järvelä, and J. Hamari, “Playing with embodied social interaction: a thematic review of experiments on social aspects in gameful virtual reality,” Interact Comput, vol. 33, no. 6, pp. 583–595, 2021.

[14] H. Tastan, C. Tuker, and T. Tong, “Using handheld user interface and direct manipulation for architectural modeling in immersive virtual reality: An exploratory study,” Computer Applications in Engineering Education, vol. 30, no. 2, pp. 415–434, 2022.

[15] H. I. Jo and J. Y. Jeon, “Perception of urban soundscape and landscape using different visual environment reproduction methods in virtual reality,” Applied Acoustics, vol. 186, p. 108498, 2022.

[16] P. Karagiannis, T. Togias, G. Michalos, and S. Makris, “Operators training using simulation and VR technology,” Procedia CIRP, vol. 96, pp. 290–294, 2021.

[17] S. Punyani, M. Kahile, and S. Kane, “Can AR, VR, and Gaming Be the Future of Physiotherapy Education and Training?,” ECS Trans, vol. 107, no. 1, p. 16057, 2022.

[18] X. Li and X. Hu, “Current status of ceramic industry and VR technology used in ceramic display and dissemination,” Sci Program, vol. 2021, pp. 1–8, 2021.

[19] Y. Wang and X. Hu, "Three‐dimensional virtual VR technology in environmental art design," International Journal of Communication Systems, vol. 35, no. 5, p. e4736, 2022.

[20] J. Werner, M. Aburaia, A. Raschendorfer, and M. Lackner, "MeshSlicer: A 3D-Printing software for printing 3D-models with a 6-axis industrial robot," Procedia CIRP, vol. 99, pp. 110–115, 2021.

[21] A. Rosinol et al., "Kimera: From SLAM to spatial perception with 3D dynamic scene graphs," Int J Rob Res, vol. 40, no. 12–14, pp. 1510–1546, 2021.

[22] F. Kuisat, F. Rößler, and A. F. Lasagni, "A Process Optimization Strategy for Texturing 3D Surfaces Using Direct Laser Interference Patterning," Adv Eng Mater, vol. 23, no. 5, p. 2001315, 2021.

[23] L. Lian and T. Lei, "Film and Television Animation Sensing and Visual Image by Computer Digital Image Technology," Journal of Mathematics, vol. 2022, pp. 1–8, 2022.

[24] T. Y.-F. Chen, Y.-L. Lo, Z.-H. Lin, and J.-Y. Lin, "Simultaneous extraction of profile and surface roughness of 3D SLM components using fringe projection method," Rapid Prototyp J, vol. 28, no. 4, pp. 789–801, 2022.

[25] L. Peng and Y. Wang, "Optimization of Intelligent Color Matching Algorithm for Animated Drawing Modeling Based on Swarm Intelligence Algorithm," in 2022 2nd International Conference on Networking, Communications and Information Technology (NetCIT), IEEE, 2022, pp. 1–7.

[26] V. M. Kotov, "Two-dimensional image edge enhancement using a spatial frequency filter of two-color radiation," Quantum Elec (Woodbury), vol. 51, no. 4, p. 348, 2021.

Classification Method of Traditional Art Painting Style Based on Color Space Transformation

Xu Zhe
School of Art and Music, Anyang Vocational and Technical College, Anyang, 455000, Henan Province, China

***Abstract*—In order to improve the accuracy and efficiency of traditional art painting style classification, a classification method of traditional art painting style based on color space transformation is proposed. This method preprocesses the traditional artistic painting style, improves the contrast of the image, makes the color and details of the image more vivid, and provides the basis for the subsequent color space conversion. After the traditional artistic painting style is stretched by automatic contrast stretching method, the color space is transformed. The purpose is to transform the image from one color space to another, so as to better extract the features of the image. Based on the traditional artistic painting style image after color space conversion, the traditional artistic painting image is balanced by the adaptive histogram equalization method with limited contrast, and an enhanced traditional artistic painting image is obtained, which further enhances the contrast of the image, makes the details in the image more prominent, and also enhances the overall visual effect of the image. Taking the enhanced traditional art painting images as input, the fuzzy C-means method is used to classify the traditional art painting styles, and the images are effectively divided into different categories according to the characteristics of the images. The experimental results show that this method can effectively enhance the image of traditional art paintings and effectively classify traditional art paintings with different styles, which has strong application effect.**

Keywords*—Color space; traditional art; painting style; classification method; fuzzy c-means

I. INTRODUCTION

In the development of human civilization, culture, and art play a very important role. Painting is an important form of culture and art [1]. It is an important part of a nation's culture. Rooted in the soil of national culture, it reflects a wide range of real-life content through beautiful art forms, thus reflecting the cultural outlook and aesthetic taste of all nations. It is a unique and important way for human beings to observe and express the world. For thousands of years, a large number of paintings have been produced. The study of these paintings is an important means for people to understand the history of human history, culture, art, and the development of science and technology to promote the development of human civilization further. It has been in the ascendant for many years. In reality, painting samples as research materials are often not easy to obtain, which has brought various inconveniences to the research work of art researchers for a long time. With the development and wide application of digital technology, more and more paintings are digitized [2] and saved in various forms such as images, videos, 3D models, multimedia documents, etc. The development of the network makes it possible for anyone to obtain digital works of art through the Internet, which brings great convenience to the majority of art researchers. The acquisition of large quantities of painting images has become a reality, which also makes large-scale art analysis possible. While digital technology has brought researchers a wealth of research materials [3], it has also provided them with many new research topics. For example, after the digitization of the murals in Mogao Grottoes in Dunhuang, art researchers need to process a large amount of data on the flying frescoes in order to study the comparison of different styles of different dynasties and classify them according to the styles of the dynasties before conducting research. Such a huge amount of data, if only classified manually by researchers, obviously requires huge and repeated work, which is undoubtedly feasible and has lost the significance of digitization [4]. In recent years, it has become an important research direction in the field of computer image processing to combine computers with art and use the powerful storage and computing power of computers [5] to realize the processing of large-scale digital painting images [6]. It is one of the research hotspots to study the classification of painting works of art according to the unique artistic style characteristics of painting images.

The artistic style of painting works generally refers to the artistic style, characteristics, style, style, and style shown in the painting works [7], which is a relatively stable and overall artistic feature presented by the interaction between the personality shown by the artist in the creation process and the semantics and context of the art works. The evaluation of the artistic styles and similarities of different paintings [8] is an important means for art researchers to classify paintings. At present, there are also many scholars studying painting image style classification methods, such as Hassanzadeh, T et al. [9], who proposed an evolutionary depth convolution neural network for image classification, input painting images into the evolutionary depth convolution neural network, and output painting image style classification results through the network model. Phasinam, K et al. [10] proposed a framework for real-time image classification. From IoT cameras to mobile applications to machine learning methods, there is everything. Arduino Uno, sensors, and Wi-Fi devices make up the hardware. The computer can "learn" from previous examples and use the machine to detect patterns from noisy or complex data sets, input the painting image into the frame, and then output its style classification results. Arco J et al. [11] proposed a feature space block sparse coding method for image classification. The image in this method is first divided into different tiles, and a dictionary is constructed after PCA is applied to these tiles. Then, the original signal is transformed

into a linear combination of dictionary elements. Then, refactor each component by iteratively activating its associated elements.Finally, the subsequent reconstruction error is used as the feature for classification. Ciga, O et al. [12] proposed the method of learning to use classification labels to segment images, which realizes the style classification of painting images by annotating a few regions of interest. Fernandes, J et al. [13] proposed an end-to-end depth learning method for table detection and table image classification in data table images. This method first learns basic feature representations, such as different types of edges and contours, through the convolution layer of the basic network. Then, using the long-term memory mechanism of the Long Short-Term Memory Network (LSTM), the spatial correlation is modeled in both horizontal and vertical directions. Finally, the spatial correlation features learned are used to construct a classifier for classification.

Although these methods can realize the classification of painting styles, they are affected by the clarity and contrast of painting images. These methods have defects in practical applications. Color space conversion is a method of exchanging different color spaces to obtain texture details and image enhancement in the image [14]. This paper proposes a traditional art painting style classification method based on color space conversion based on color space conversion, so as to improve the level of painting style classification technology. The advantage of this method is to transform the image from one color space to another, so as to better extract the features of the image, further improve the contrast of the image, make the details in the image more prominent, and at the same time enhance the overall visual effect of the image. The technical route of this paper is as follows:

1) This method preprocesses the traditional artistic painting style, improves the contrast of the image, makes the color and details of the image more vivid, and provides the basis for the subsequent color space conversion.

2) After the traditional artistic painting style is stretched by automatic contrast stretching method, the color space is transformed. Based on the traditional artistic painting style image after color space conversion, the traditional artistic painting image is balanced by using the adaptive histogram equalization method with limited contrast, and the enhanced traditional artistic painting image is obtained.

3) Taking the enhanced traditional art painting images as input, the traditional art painting styles are classified by fuzzy C-means method, and the images are effectively classified into different categories according to the characteristics of the images.

II. Classification Method of Traditional Art Painting Style

A. Image Enhancement Processing of Traditional Art Painting based on Color Space Conversion

1) Automatic contrast stretching: Before classifying the style of traditional art painting images [15], it is necessary to enhance them to make their image features more obvious. Automatic contrast stretching is a point operation [16] whose purpose is to change the pixel gray value of the current image so that the pixel value distribution of the resulting image covers all available ranges of pixel gray value. This algorithm maps the current darkest and brightest pixel gray values to the minimum and maximum values in the range of available gray values. Then, it makes the middle gray values linearly distributed. In the classic automatic contrast stretching algorithm, the gray value of each input pixel is calculated using the following formula:

$$f_{ac}(e) = e_{min} + (e + e_{low}) \times \frac{e_{max} - e_{min}}{e_{high} - e_{low}} \quad (1)$$

Where: e_{max}、e_{min} represents the maximum and minimum values of pixel grayscale, e_{low} and e_{high} are the minimum and maximum values of pixel gray values in the current image, and the range of image pixel gray values can be $[e_{max}, e_{min}]$.

The mapping function in Formula (1) is only strongly affected by several extreme values, which may not represent the main content of the image. Therefore, this scheme adopts the modified automatic contrast stretching algorithm, and the formula of the modified automatic contrast stretching algorithm is:

$$f_{mac}(e) = \begin{cases} e_{min} & e \le \hat{e}_{low} \\ e_{min} + (e - \hat{e}_{low}) \times \frac{e_{max} - e_{min}}{\hat{a}_{high} - e_{low}} & \hat{e}_{low} < e < \hat{e}_{high} \\ e_{max} & e \ge \hat{e}_{high} \end{cases} \quad (2)$$

In the above formula, $\hat{e}_{low}$、$\hat{e}_{high}$ represent two thresholds respectively. These two thresholds depend on the content of the image and can be represented by the cumulative histogram of the image $H(i)$ calculated:

$$\hat{e}_{low} = M \times N \times f_{ac}(e) \quad (3)$$

$$\hat{e}_{high} = M \times N \times f_{mac}(e) \quad (4)$$

Where: $M \times N$ is the number of pixels in the image. All pixel values can be derived from the formula $\hat{e}_{low}$ and $\hat{e}_{high}$ values other than (including) are mapped to extreme values respectively e_{min} and e_{max}, intermediate values are linearly mapped to $[e_{max}, e_{min}]$. It can be seen that the mapping function is not only affected by several extreme values but also depends on a group of representative pixel values. In this algorithm, the contrast stretching of gray image Eq. (2) is applied to the R, G, and B color channels of the color image, respectively, to complete the automatic contrast stretching of traditional art painting images.

2) Color space conversion of traditional art painting based on polynomial regression: Based on the traditional art painting image after automatic contrast stretching [17], the color space is converted by polynomial regression.

Polynomial regression belongs to linear regression. In practical application, the independent variable is constructed by selecting the number of polynomial terms x and dependent variable y in the polynomial regression model, the dependent variable y and multiple arguments $x_1, x_2, \cdots, x_M$ with linear relationship, where is the number of independent variables, there are:

$$X_{DF} = (x_{t1}, x_{t2}, \cdots, x_{tN}), t = 1,2, \cdots, N \quad (5)$$

Where, N indicates the number of dependent variables. The relationship between the dependent variable and independent variable can be described as:

$$\begin{aligned} y_1 &= \alpha_0 + \alpha_1 x_{11} + \alpha_2 x_{12} + \cdots + \alpha_M x_{1M} + \varepsilon_1 \\ y_2 &= \alpha_0 + \alpha_1 x_{21} + \alpha_2 x_{22} + \cdots + \alpha_M x_{2M} + \varepsilon_2 \\ y_N &= \alpha_0 + \alpha_1 x_{N1} + \alpha_2 x_{N2} + \cdots + \alpha_M x_{NM} + \varepsilon_N \end{aligned} \quad (6)$$

Among them, $\alpha_0, \alpha_1, \cdots, \alpha_M$ represents the coefficient to be determined $\varepsilon_0, \varepsilon_1, \cdots, \varepsilon_N$ represents an independent random variable.

If the parameter $b_0, b_1, \ldots, b_M$ is estimated by the least squares method β the regression formula obtained by fitting is:

$$\hat{y} = b_0 + b_0 x_1 + \cdots + b_M x_M \quad (7)$$

According to the principle of least squares, the coefficient $b_0, b_1, \ldots, b_M$ all measurements shall be obtained y_t, and regression value $\hat{y}$ minimum sum of residual squares of Q:

$$Q = \sum_{i=1}^{N} (y_t - \hat{y}_t)^2 \quad (8)$$

The polynomial regression method can describe not only linear problems but also nonlinear problems [18]. When describing nonlinear problems, the dependent variable needs to be y and arguments x a selected polynomial model is constructed, which can be correctly applied to the actual color signal processing system. Build the conversion between RGB color space and CMYK color space as fixed K. The polynomial of the conversion between RGB and CMY in the case of value is as follows:

$$C = \sum_{i=0}^{n} \sum_{j=0}^{n} \sum_{p=0}^{n} \alpha_C R^i G^j B^p \quad (9)$$

$$M = \sum_{i=0}^{n} \sum_{j=0}^{n} \sum_{p=0}^{n} \alpha_M R^i G^j B^p \quad (10)$$

$$Y = \sum_{i=0}^{n} \sum_{j=0}^{n} \sum_{p=0}^{n} \alpha_Y R^i G^j B^p \quad (11)$$

where, α is a polynomial coefficient, n is the order of a polynomial, and $i + j + p \leq n$, the polynomial is represented by a matrix as follows:

$$\begin{bmatrix} \boldsymbol{X} & \boldsymbol{Y} & \boldsymbol{Z} \end{bmatrix}^{\boldsymbol{T}} = \boldsymbol{B_{3L}} \times \boldsymbol{\rho_L} \quad (12)$$

where, B_{3L} represents the coefficient matrix, ρ_L represents a polynomial matrix.

After the above steps, the traditional art painting image RGB color space and CMYK color space can be converted.

3) Adaptive histogram equalization with limited contrast: The Histogram equalization (HE) method has the advantages of fast speed and an obvious effect in enhancing image contrast. Histogram equalization can improve the overall image contrast [19]. Still, after processing, the image will appear "too bright or too dark," and local details cannot be processed, resulting in a loss of details and poor effect. Considering this, adaptive histogram equalization (AHE) based on the idea of block processing has been proposed to solve the problem of local highlight or too dark. Still, these two methods also amplify noise while enhancing contrast. Based on the advantages of the AHE algorithm, the concept of limiting contrast is proposed to solve the problem of amplifying noise. CLAHE algorithm not only effectively improves the contrast but also suppresses the generation of noise.

The specific steps of the CLAHE algorithm are as follows:

1) Divide the image into $n \times n$ there are rectangular sub blocks of the same size and non-overlapping each other. As the number of sub blocks increases, the enhancement effect of the image becomes more significant, but more details are lost [20].

2) Calculate the sub-block histogram.

3) Solve restricted values η.

$$\eta = \chi \times \frac{n_x n_y}{\widehat{K}} \quad (13)$$

where:$n_x$represent Subblock x number of direction pixels; n_y express y number of direction pixels; $\widehat{K}$ is the gray level; χ is the limiting factor.

4) Crop the histogram and reassign the pixels. Cut sub-block histogram $h(x)$ restricted value η constraints, the exceeding part of the number of pixels is evenly distributed to other gray levels, and the pixel points are cut and reallocated, as shown in Fig. 1.

5) Sub block histogram equalization.

6) The bilinear difference reconstructs the gray value.

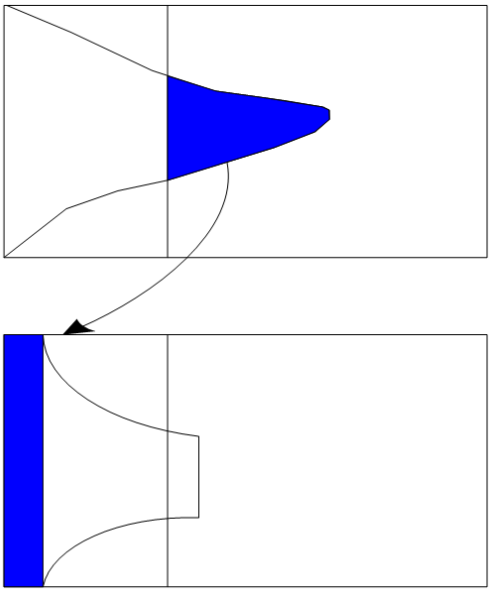


Fig. 1. Pixel cropping and reassigning.

The pixel value obtained only by mapping function transformation will cause the image to be blocky, and the bilinear difference processing for each point of the image can effectively avoid blocky.

The bilinear difference is only for the area surrounded by the center points of four blocks. The center of each sub block is taken as the reference point and recorded as $\theta_{11}(x_1, y_2)$、$\theta_{12}(x_2, y_2)$、$\theta_{21}(x_1, y_1)$、$\theta_{22}(x_2, y_1)$. Points to be calculated: The pixel value of P is determined by four adjacent reference points.

After the above steps P, after pixel reconstruction, traditional art painting image enhancement is realized.

B. Extraction of Salient Information from Traditional Art Painting Images

After the enhancement of traditional art painting images, extract its significant information to prepare for classification [21]. In the traditional art painting image, the image edge is usually used as the image background area, away from the image center of the target and surrounded by the position where the image's prominent target exists. The image background area is used as the super pixel block of the image edge, and the background area is compared with the remaining area from the perspective of color space to obtain the significance information of different areas [22].

The acquisition process of prior significance information of digital painting image background is as follows:

1) Get image manifold sorting: Set the background area of the traditional art painting image as the super pixel block at the image edge, and compare the color space of all the super pixel blocks with that of the background super pixel block to obtain the significance information of all parts of the image [23]. The manifold sorting algorithm is used to calculate the salient information of traditional art painting images. The algorithm obtains the sorting function by setting the correlation between fixed nodes and other nodes. It sets the correlation between nodes and other nodes according to the obtained sorting function measurement [24]. The correlation of each point is obtained by using the sorting method of the internal manifold structure of the data. Set data vector $M = [m_1, m_2, \cdots, m_n]^T$, assign the sorting value to each point in the data vector m_i, and get the final output of the function $g = [g_1, g_2, \cdots, g_n]^T$. Set up $p = [p_1, p_2, \cdots, p_n]^T$ as a marker vector, when $p_i = 1$ as well as $p_i = 0$ respectively m_i set query nodes and non set query nodes for. Set existence diagram $Q = (U, R)$ in the dataset, where U and R represent data node set and image edge set respectively, and use weighted similarity matrix $Q = [q_{ij}]_{n\times n}$ the above parameters can be obtained.

Using an optimization problem to express the optimal ordering of nodes $g*$, the calculation formula is as follows:

$$g* = (B - \alpha Q)^{-1} p \tag{14}$$

where:B represents a diagonal matrix; α represents the Plath coefficient.

2) Significant information calculation: The input digital drawing image is converted to the diagram structure representation according to Formula (14). Select the super pixel segmentation method to represent the image graph structure with multiple super pixel blocks [25], where each super pixel and node $U_i(1 \le i \le |U|)$ correspondingly set the query node as the super pixel node at the edge. The regular graph is used to represent the graph structure of the image [26]. That is, each node in the image is connected to the adjacent node, and the graph is connected to the adjacent node. The weight formula of the image edge is as follows:

$$k_{ij} = exp(-\frac{\sqrt{(L_i - L_j)^2 + (A_i - A_j)^2 + (B_i - B_j)^2}}{\delta^2}) \tag{15}$$

Where: L_i, A_i, B_i represent super pixels in color space i average of LAB value; δ indicates the node edge coefficient. The significance information of each node in the image is represented by the ranking value obtained by Formula (15). The query node is set as the upper boundary, and the ranking vector is obtained by using the significance information of the upper boundary and the significance information of other nodes $g*(i)$, using $\bar{g}*(i)$ indicates the significance information obtained by normalization to $[0,1]$ interval, set $\bar{g}*(i)$ use 1 as a difference to obtain the foreground significance information map.

Obtain saliency information map through image upper boundary $O_t(i)$. The formula is as follows:

$$O_t(i) = 1 - \bar{g}*(i) \tag{16}$$

Repeat the above steps to obtain the significance information map of the left, right, and lower boundaries $O_t(i)$, O_r, O_d means that the four saliency information maps obtained are multiplied to obtain the saliency information map of the final digital painting image, and the formula is as follows:

$$O\ \ (i) = O_t(i) \times O_d(i) \times O_r(i) \times O_l(i) \tag{17}$$

After the above steps, the salient information of traditional art painting images is obtained.

C. Traditional Art Painting Image Style Classification Method based on Improved Fuzzy C-Means Algorithm

The salient information of traditional art painting images obtained in the above sections is used as input, and the improved fuzzy C-means algorithm is used to realize the classification of traditional art painting styles.

The fuzzy C-means algorithm is a local search algorithm that mainly constructs the Lagrange function and iteratively calculates the minimum of the sum of squares of the global weighted distances from each sample to the cluster center to obtain the optimal cluster center. If the sample set is U, the cluster center set is V, the objective function can be expressed as:

$$min\, J\,(U, V) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik}) m (d_{ik})^2 \tag{18}$$

where: n and c are the number of samples to be classified and the number of clusters; m is the fuzzy degree coefficient,

which is responsible for controlling the sharing degree between fuzzy classes; u_{ik} is a sample pair k clustering i and has $0 \le u_{ik} \le 1$, $\sum_{i=1}^{c}(u_{ik}) = 1$; d_{ik} is a sample k to clustering i the Euclidean distance of the center.

Because the fuzzy C-means algorithm is prone to the local optimal solution, slow convergence, and poor classification effect under high-dimensional data, we use the differential evolution algorithm to improve the fuzzy C-means algorithm and use the improved fuzzy C-means algorithm to classify the traditional art painting image style.

The differential evolution algorithm is used to improve the fuzzy C-means algorithm. First, a group of initial centers is randomly generated, and the non-center points are grouped into the center point area one by one according to the principle of minimum distance. Then, the differential evolution algorithm is used to carry out parameter adaptive adjustment, mutation, crossover, and selection operations for the current partition so that the search process changes from an unsupervised state to a dynamic information-guided optimization and performs the second assignment operation on the current part of non-optimal individuals to enhance the small range of the later stage of the algorithm fine search capability. The specific steps are as follows:

Step 1: Initialize the parameters. Determine various control parameters required by the algorithm and make the number of iterations $T = 0$.

Step 2: Initialize the population. The cluster center is used as a population individual to code. First, the initial cluster center is randomly generated, and the individual coding method is as follows:

$$\boldsymbol{X_i} = (\boldsymbol{x_{1,1}}, \boldsymbol{x_{1,2}}, \cdots, \boldsymbol{x_{1,s}}, \boldsymbol{x_{2,1}}, \cdots, \boldsymbol{x_{2,s}}, \cdots, \boldsymbol{x_{c,s}}) \tag{19}$$

where: s is the dimension of the cluster center; c is the number of individuals. Then, the differential evolution algorithm is used to generate a random initial population:

$$x_{i,j} = xi, jmini, jmax_{i,jmin} \tag{20}$$

where: $x_{i,j}$, $x_{i,jmax}$, $x_{i,jmin}$ individual set of population X_i of j components and their upper and lower bounds; r is a random number in the range of [0,1].

Step 3: Formula (18) is taken as the objective function, and the adaptability evaluation function, so the objective function value is the population fitness evaluation result. The smaller the objective function value, the higher the quality of the population of individuals. Calculate the fitness of the current population and determine whether the maximum number of iterations of the algorithm is reached T_{max}.

Step 4: Adaptively adjust the mutation factor F and hybridization factor C. The variance of population fitness can effectively reflect the individual distribution of the contemporary population and dynamically adjust the control parameters in a targeted way, thus having good results. The variance of group fitness σ^2 can be expressed as:

$$\boldsymbol{\sigma^2} = \sum_{\boldsymbol{I=1}}^{\boldsymbol{n}} (\frac{\boldsymbol{\delta_i - \delta_{av}}}{\boldsymbol{\delta_b}})^2 \tag{21}$$

where: N is the population size; δ_i is for i individual fitness; δ_{av} is the average fitness of the population; δ_b is the best fitness of the group.

Based on the above calculation σ^2, the adjustment parameter changes from a fixed value to the following dynamic form:

$$F_k = F_{\min} + (F_{\max} - F_{\min} - \frac{\sigma_k^2 (F_{\max} - F_{\min})}{N} \tag{22}$$

$$\hat{Z}_k = \hat{Z}_{\min} + (\hat{Z}_{\max} - \hat{Z}_{\min}) 1 - \frac{\sigma_k^2 (\hat{Z}_{\max} - \hat{Z}_{\min})}{N} \tag{23}$$

where: F_{max} and F_{min} are the upper bound and the lower bound of the variation factor; $\hat{Z}_{max}$ and $\hat{Z}_{min}$ are the upper bound and lower bound of hybridization factors respectively; σ_k^2 is for k generation population fitness variance.

Step 5: Conduct mutation and crossover operations on the population, generate a trial offspring population, recalculate the fitness, and use the "greed" strategy for selection operations to form a new generation of population.

Step 6: Secondary assignment. Randomly select some individuals from the non-optimal individuals of the current population according to the previously specified probability distribution function for secondary assignment:

$$X_{i,re} = X_i(1 - \delta_{rand}) + O\delta_{rand} \tag{24}$$

$$O = \frac{\lambda_{ad} I X_{optimal} + r X_{r1} - r X_{r2}}{\lambda_{ad} I} \tag{25}$$

where: X_i is the individual value of the current population; δ_{rand} is a binary random decision variable; O is the local search adjustment operator defined in this paper; $X_{optimal}$ is the best individual in the contemporary population; λ_{ad} is an adjustment factor, which is used to adjust the local search sensitivity of the differential evolution algorithm. The larger the value, the stronger the local search ability of the algorithm; I is the number of iterations; X_{r1} and X_{r2} are from a new population of randomly selected individuals and meet $X_{r1} \neq X_{r2}$.

As the number of iterations increases, the optimization scope of the algorithm gradually shrinks, and finally, the optimal clustering division result is obtained, which is the classification result of traditional art painting image style.

III. Experimental Analysis

With 5000 traditional art paintings of different styles as experimental objects, including comics, sketches, ink paintings, watercolors, and other painting styles, this paper uses this method to classify the 5000 traditional art painting styles and analyzes and verify the practical application effect of this method.

Set the parameters needed for the experiment, as shown in Table I.

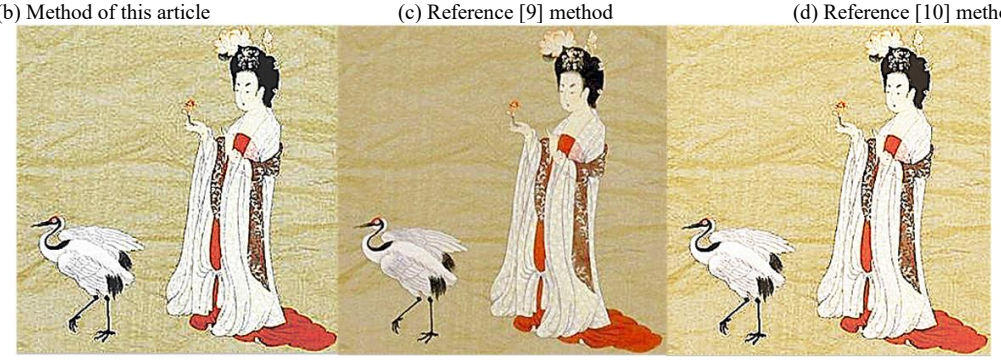
Taking a traditional art painting as the experimental object and contrast stretching as the measurement index, the enhancement ability of this method to the traditional art painting image is tested. To make the experimental results more sufficient, the literature [9] method, literature [10] method, literature [11] method, literature [12] method, and literature [13] method are used at the same time. The test results are shown in Fig. 2. Analysis of Fig. 2 shows that the overall color of the original traditional art painting image is dark, and the internal details of the image are not clear enough. After using six methods to enhance the traditional art painting image, the clarity and contrast of the traditional art painting image enhanced by this method have been significantly improved. Although the contrast of the traditional art painting image enhanced by the literature [9] method, literature [11] method, and literature [13] method is effectively prompted, the contrast is too large, which leads to the disappearance of the internal details of the traditional art painting image, while the traditional art painting image enhanced by the literature [10] method and literature [12] method has a low lightness and darkness. The details in traditional art painting images cannot be clearly presented. To sum up, this method can effectively stretch the traditional art painting image and has a strong traditional art painting image enhancement effect.

TABLE I. CONFIGURATION TABLE OF PARAMETERS REQUIRED FOR EXPERIMENTS

Serial number	Parameter	Content
1	Color space conversion parameter	From RGB color space to HSV color space, it is necessary to convert the coefficients in the formula.
2	Automatic contrast stretching parameter	Relates to parameters such as tensile strength or tensile ratio.
3	Adaptive histogram equalization parameters with finite contrast	Parameters such as contrast limit threshold and the number of histogram partitions are involved.
4	Fuzzy c-means clustering parameters	When using fuzzy C-means method to classify styles, it is necessary to set parameters such as the number of clusters (that is, the number of styles) and fuzzy factors.
5	Relevant parameters of training and testing data sets	Including the ratio of training set to test set, data enhancement parameters (such as rotation angle, cutting size, etc.), batch size, learning rate, etc.

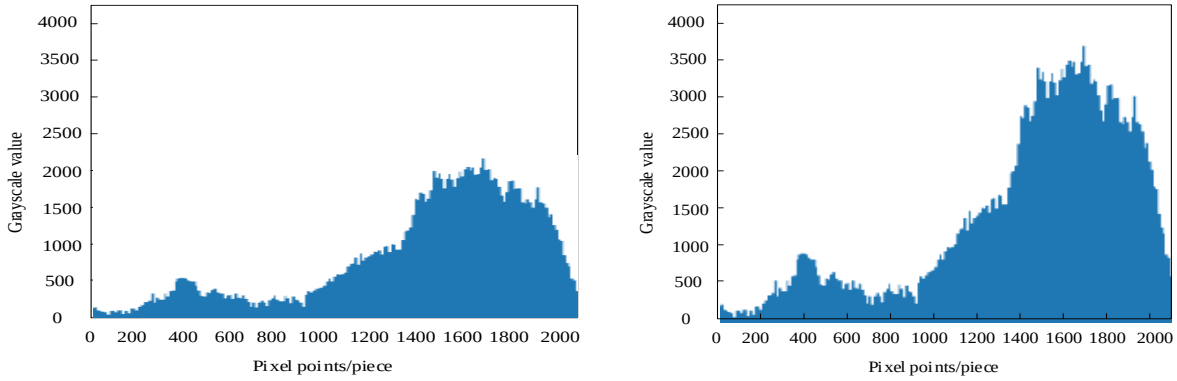
(a) Primitive traditional art painting images

(b) Method of this article (c) Reference [9] method (d) Reference [10] method



(e) Reference [11] method (f) Reference [12] method (g) Reference [13] method

Fig. 2. Test results of image enhancement in traditional art painting.



(a) Before enhancement (b) After enhancement

Fig. 3. Test results of image enhancement in traditional art painting.

To further verify the enhancement effect of this method on traditional art painting images, take a traditional art painting image as the experimental object and use this method to enhance it. Additionally, the histogram is employed to showcase the grayscale values of the traditional art painting image both before and after undergoing enhancement processing. The test results are shown in Fig. 3.

According to the comprehensive analysis of Fig. 3, after applying the method in this paper to enhance the traditional art painting image under the same pixel situation, the pixel gray value can be effectively improved. This result further verifies that the method in this paper has a good enhancement effect on the traditional art painting image.

The color space conversion accuracy is taken as a measure to test the color space conversion ability of the method in this paper for traditional art painting images under different color space lightness and color saturation. The test results are shown in Table II.

According to the analysis of Table II, when using this method to convert the color space of traditional art painting images, the separation accuracy value is higher than 0.9 under the conditions of different color space lightness and color space saturation. This value shows that this method can effectively convert the color space of traditional art painting images and also confirms from the side that this method has a strong ability to classify traditional art painting styles.

Take a large number of traditional art painting images as experimental objects and use the method in this paper to classify their styles. In order to make the verification results more sufficient, the style classification is carried out respectively under the conditions of no rotation, 30-degree rotation, and 70-degree rotation of traditional art painting images. The results are shown in Fig. 4 to Fig. 6, respectively.

TABLE II. PRECISION VALUES OF COLOR SPACE CONVERSION IN TRADITIONAL ART PAINTING IMAGES

Color space brightness	Color space conversion accuracy	Saturation	Color space conversion accuracy
5	0.95	50	0.98
10	0.95	55	0.95
15	0.96	60	0.96
20	0.94	65	0.93
25	0.95	70	0.98
30	0.97	75	0.94
35	0.92	80	0.97
40	0.96	85	0.92
45	0.93	90	0.95
50	0.95	95	0.94
55	0.94	100	0.96

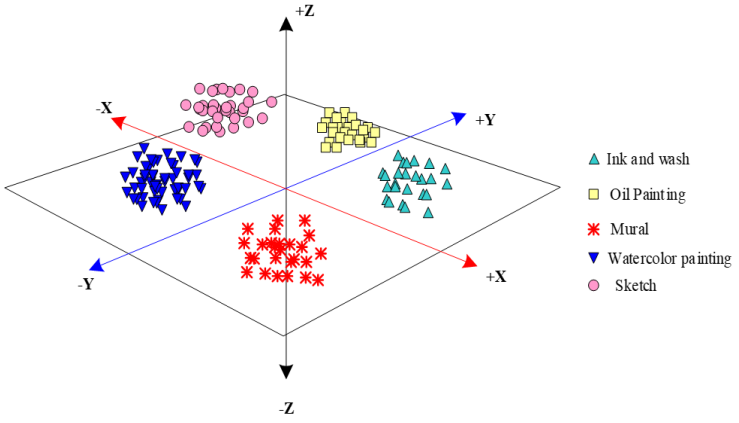


Fig. 4. No rotation.

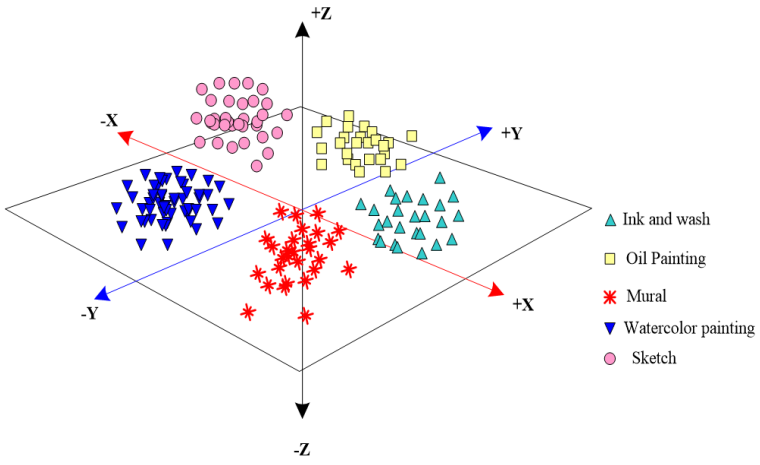


Fig. 5. Rotate 30 degrees.

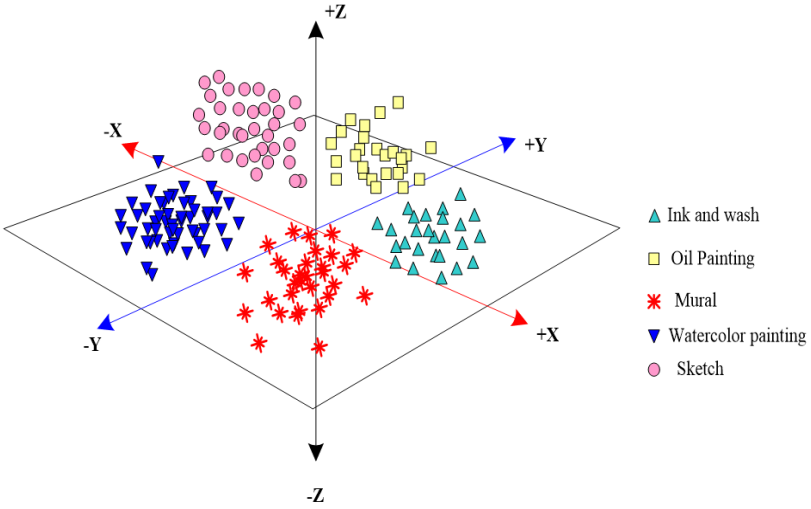


Fig. 6. Rotate 70 degrees.

From the analysis of Fig. 4 to Fig. 6, it can be seen that when the traditional art painting images are not rotated, the distance between different clusters is far when this method classifies their styles, and the distribution of traditional art painting images in the same cluster is relatively dense. When traditional art painting images are rotated 30 degrees and 70 degrees, respectively, the distribution distance of different clusters is shortened when the method in this paper is applied to classify traditional art painting styles, and the distribution of traditional art painting images within the same cluster becomes sparse, but the differences between different clusters are still obvious. The aforementioned findings indicate that the utilization of this approach leads to an efficient classification of traditional art painting styles. When classifying, the density between class clusters is better, and the classification of traditional art painting styles is more accurate, which is not affected by the rotation of traditional art painting images.

The Xie Beni (XB) index is used to measure the classification accuracy of traditional art painting styles. The smaller the value, the higher the classification accuracy of traditional art painting styles. With 15 traditional art painting images as the experimental objects, this paper analyzes the changes in the Xie Beni index when classifying the magnification of traditional art painting images by this method and sets the threshold value of the Xie Beni index to 0.4. The test results are shown in Table III.

TABLE III. CLASSIFICATION OF TRADITIONAL ART PAINTING STYLES BY XIE BENI INDEX

Traditional Painting Image Coding	Magnification		
	1	*1.5*	*2*
1	0.12	0.21	0.32
2	0.11	0.23	0.33
3	0.19	0.25	0.31
4	0.08	0.19	0.28
5	0.07	0.19	0.24
6	0.12	0.25	0.26
7	0.15	0.29	0.27
8	0.09	0.31	0.31
9	0.14	0.24	0.33
10	0.16	0.25	0.35
11	0.08	0.26	0.28
12	0.17	0.22	0.29
13	0.15	0.21	0.31
14	0.16	0.27	0.25
15	0.11	0.18	0.24

Analysis of Table III shows that with the increase of magnification of traditional art painting images, the Xie Beni index when classifying traditional art painting styles in this method shows an upward trend, but the increase is not significant. When the magnification of the traditional art painting image is 2, the maximum value of the Xie Beni index when this method classifies the traditional art painting style is only 0.35, which is lower than the preset threshold value of the Xie Beni index. The above results show that the method in this paper has a high accuracy in classifying traditional art painting styles and has a relatively significant application effect.

IV. DISCUSSION

1) The traditional art painting style classification method based on color space transformation has obviously improved the clarity and contrast of the traditional art painting images, which can effectively stretch the traditional art painting images and has strong traditional art painting image enhancement effect.

2) After the traditional art painting style classification method based on color space transformation is applied to enhance the traditional art painting image, the gray value of each pixel is effectively improved under the same pixel, which further verifies that the method in this paper has a good enhancement effect on the traditional art painting image.

3) When the traditional art painting style classification method based on color space conversion is used to convert the color space of traditional art painting images, the separation accuracy values are all higher than 0.9 under different color space lightness and color space saturation, which shows that the method in this paper can effectively convert the color space of traditional art painting images, and also proves from the side that the method in this paper has strong ability to classify traditional art painting styles.

4) The traditional art painting style classification method based on color space transformation is far away between different clusters, and the traditional art painting images in the same cluster are densely distributed. The application of this method can effectively classify the traditional art painting styles, and the density between clusters is good, and the classification of traditional art painting styles is more accurate, which is not affected by the rotation of traditional art painting images.

5) With the increase of the magnification of traditional art painting images, the Xie-Beni index of the traditional art painting styles classified by this method shows an upward trend, and the classification of traditional art painting styles by this method has a high accuracy and a remarkable application effect.

V. CONCLUSION

The classification of traditional art painting styles is one of the more efficient applications of deep learning methods in the field of traditional art painting. It can effectively enhance the performance of both classification and imitation of traditional art painting styles. This paper studies the classification algorithm of art-style images and proposes a traditional art painting-style classification method based on color space conversion. The salient information contained in traditional art painting images can effectively reflect the image's salient features, improve the accuracy of style classification, and combine the salient information with the fuzzy C-means algorithm. According to the characteristics of traditional art painting images, the salient features contained in different styles of painting images are extracted. Then, the extracted features are passed on to the fuzzy C-means algorithm in order to accomplish the classification of traditional art painting styles. Through the actual verification of the method in this paper, based on the verification results, it is evident that the approach presented in this paper exhibits a high level of effectiveness in categorizing traditional art painting styles, yielding notable practical applications.

Although the traditional art painting style classification method based on color space transformation has certain effectiveness, it also has some limitations.

1) Dependence on color space transformation: One of the core of this method is color space transformation. However, the conversion of color space may be influenced by many factors such as lighting conditions, pigment types and painting techniques. For some special cases, the transformation of color space may not accurately capture the essential characteristics of painting style, which leads to the error of classification results.

2) Limitations of feature extraction: This method mainly relies on image features such as color and contrast for classification. However, the traditional artistic painting style is not limited to color and contrast, but also includes many characteristics such as texture, composition and brush strokes. These methods fail to fully consider other features, which will limit the accuracy and comprehensiveness of classification.

3) Requirements of data set: The method based on color space conversion has higher requirements for training data set. It needs enough diverse and representative samples for training in order to obtain an accurate classification model. If the data set is small or the samples are not rich enough, the generalization ability of the model will be insufficient, and the new paintings cannot be accurately classified.

To sum up, the traditional art painting style classification method based on color space transformation has achieved certain results, but it still faces some limitations. Future research can further combine advanced technologies such as deep learning to improve the accuracy and efficiency of classification.

DATA AVAILABILITY

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation."

CONFLICTS OF INTEREST

The authors declared that they have no conflicts of interest regarding this work."

ACKNOWLEDGMENTS

The work is not supported by any funding."

COMPETING OF INTERESTS

The authors declare no competing of interests.

AUTHORSHIP CONTRIBUTION STATEMENT

Xu Zhe: Writing-Original draft preparation

Conceptualization, Supervision, Project administration.

REFERENCES

[1] X. Xie and B. Lv, "Design of painting art style rendering system based on convolutional neural network," Sci Program, vol. 2021, pp. 1–11, 2021.

[2] Y. Cao, Z. Han, R. Kong, C. Zhang, and Q. Xie, "Technical composition and creation of interactive installation art works under the background of artificial intelligence," Math Probl Eng, vol. 2021, pp. 1–11, 2021.

[3] M. Jang, M. Aavakare, S. Nikou, and S. Kim, "The impact of literacy on intention to use digital technology for learning: A comparative study of Korea and Finland," Telecomm Policy, vol. 45, no. 7, p. 102154, 2021.

[4] Y. Song and X. Liao, "Design of print creation system based on Digitization," in Journal of Physics: Conference Series, IOP Publishing, 2021, p. 032030.

[5] J. Zhan, G. V Merrett, and A. S. Weddell, "Exploring the effect of energy storage sizing on intermittent computing system performance," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 41, no. 3, pp. 492–501, 2021.

[6] N. Wei, "Research on the algorithm of painting image style feature extraction based on intelligent vision," Future Generation Computer Systems, vol. 123, pp. 196–200, 2021.

[7] V. Bihari, A. K. Singh, and R. K. Mallik, "The Artistic Exploration of Balbir Singh Katt: An Introduction," ECS Trans, vol. 107, no. 1, p. 7339, 2022.

[8] C. Li and E. D. Raj, "Feature recognition of abstract art painting multilevel based convolutional ancient recognition neural network method," Journal of Interconnection Networks, vol. 22, no. Supp01, p. 2141003, 2022.

[9] T. Hassanzadeh, D. Essam, and R. Sarker, "EvoDCNN: An evolutionary deep convolutional neural network for image classification," Neurocomputing, vol. 488, pp. 271–283, 2022.

[10] K. Phasinam and T. Kassanuk, "Machine learning and internet of things (IoT) for real-time image classification in smart agriculture," ECS Trans, vol. 107, no. 1, p. 3305, 2022.

[11] J. E. Arco, A. Ortiz, J. Ramírez, Y.-D. Zhang, and J. M. Górriz, "Tiled sparse coding in eigenspaces for image classification," Int J Neural Syst, vol. 32, no. 03, p. 2250007, 2022.

[12] O. Ciga and A. L. Martel, "Learning to segment images with classification labels," Med Image Anal, vol. 68, p. 101912, 2021.

[13] J. Fernandes, M. Simsek, B. Kantarci, and S. Khan, "TableDet: An end-to-end deep learning approach for table detection and table image classification in data sheet images," Neurocomputing, vol. 468, pp. 317–334, 2022.

[14] Z. Su, J. Yang, P. Li, J. Jing, and H. Zhang, "A precise method of color space conversion in the digital printing process based on PSO-DBN," Textile Research Journal, vol. 92, no. 9–10, pp. 1673–1681, 2022.

[15] Y. Sun and E. A. Ball, "Automatic modulation classification using techniques from image classification," IET Communications, vol. 16, no. 11, pp. 1303–1314, 2022.

[16] M. Wang, Y. Shao, and ,H. Liu, "APST-Flow: A Reversible Network-Based Artistic Painting Style Transfer Method," Computers, Materials & Continua, vol. 75, no. 3, pp. 5229-5254, 2023.

[17] S. Liu, X. "Research on the classification method of artistic painting image style based on naive Bayesian," International Journal of Information and Communication Technology, vol. 21, no. 4, pp. 398-411, 2022.

[18] J. Ran, H. Liu, and J. Luo, "The color matching design based on polynomial regression," Textile Research Journal, vol. 92, no. 7–8, pp. 1235–1245, 2022.

[19] X. Du and W. Wang, "Recognizing the Style of Artistic Painting via Information Entropy for Smart City Construction," International Journal of Distributed Systems and Technologies (IJDST), vol. 12, no. 2, pp. 46-54, 2021.

[20] M. Kim and H.-C. Choi, "Uncorrelated feature encoding for faster image style transfer," Neural Networks, vol. 140, pp. 148–157, 2021.

[21] Y. Wu, "Application of improved boosting algorithm for art image classification," Sci Program, vol. 2021, pp. 1–11, 2021.

[22] Q. Ge, F. Ruan, B. Qiao, Q. Zhang, X. Zuo, and L. Dang, "Side-scan sonar image classification based on style transfer and pre-trained convolutional neural networks," Electronics (Basel), vol. 10, no. 15, p. 1823, 2021.

[23] F. Tarsitano, C. Bruderer, K. Schawinski, and W. G. Hartley, "Image feature extraction and galaxy classification: a novel and efficient approach with automated machine learning," Mon Not R Astron Soc, vol. 511, no. 3, pp. 3330–3338, 2022.

[24] H. C. Biscaia, J. Canejo, S. Zhang, and R. Almeida, "Using digital image correlation to evaluate the bond between carbon fibre-reinforced polymers and timber," Struct Health Monit, vol. 21, no. 2, pp. 534–557, 2022.

[25] D. Bo, R. Ma, K. Wang, M. Su, and P. An, "Deep learning-based image inpainting with structure map," J Electron Imaging, vol. 30, no. 3, p. 33028, 2021.

[26] L. Guo, H. Du, and D. Huang, "A quantum image encryption algorithm based on the Feistel structure," Quantum Inf Process, vol. 21, pp. 1–18, 2022.

Research on Image Algorithm for Face Recognition Based on Deep Learning

Qiang Wu
School of Information Technology, Jiangsu Open University, Nanjing, 210036, China

***Abstract*—As people's requirements for applications are getting higher and higher, the recognition of facial features has been paid more and more attention. The current facial feature recognition algorithm not only takes a long time, but also has problems such as large system resource consumption and long running time in practical applications. Based on this, the research proposes a multi-task face recognition algorithm by combining multi-task deep learning on the basis of convolutional neural network, and analyzes its performance in four dimensions of face identity, age, gender, and fatigue state. The experimental results show that the multi-task face recognition algorithm model obtained through layer-by-layer progression takes less time than other models and can complete more tasks in the same training time. At the same time, comparing the best model M44 with other algorithms in four dimensions, it is found that the Mean Absolute Error lowest is 3.53, and the highest Accuracy value is 98.3%. On the whole, the multi-task face recognition algorithm proposed in the study can recognize facial features efficiently and quickly. At the same time, its training time is short, the calculation speed is fast, and the recognition accuracy is much higher than other algorithms. It is applied to intelligent driving behavior. Analysis, intelligent clothing navigation and other aspects have strong practical significance.**

***Keywords*—Multi task deep learning; face recognition; convolution neural network; multi task; dimension**

I. INTRODUCTION

The rapid development of artificial intelligence has made it the core driving force of industrial transformation. It is also widely used in the fields of speech recognition and image recognition, and has gradually replaced human resources [1-3]. As an important pillar in the field of image recognition, face recognition has been widely used in business, identity authentication and other fields, and has gradually received attention. In addition, face recognition is no longer limited to a single task, and the needs of multi-task face recognition. It is also constantly improving [4-5]. In this context, many domestic and foreign scholars have conducted in-depth research on it. Khan AA et al. built a face recognition image model on the basis of neural network and integrated genetic algorithm and principal component analysis, which provided help for face matching in forensic investigation [6]. Srivastava S et al. proposed a new method of biometric authentication face recognition based on artificial neural network, which effectively reduces the error rate of face recognition [7]. Karanwal overcomes the problem of low image recognition rate in local binary patterns by proposing descriptors [8]. However, although the current feature matching algorithms using similarity measures in single task face recognition are simple and fast, they are difficult to robustly determine the threshold size. Although feature matching algorithms using feature subspaces can map intra class differences to subspaces for compression, their noise will also be mapped to subspaces and easily amplified. Although feature matching algorithms using statistical models have good robustness and identification, they are prone to data overfitting, and many deep learning methods also have shortcomings such as low accuracy and high hardware requirements. The methods in multitasking facial recognition have drawbacks such as multiple parameters, slow running speed, and high resource consumption. Based on this, the research proposes a multi-task deep learning algorithm by integrating deep learning on the basis of Convolutional Neural Network (CNN). The purpose is to effectively improve the accuracy of face recognition through a new face recognition algorithm. At the same time, it reduces its recognition time and provides effective suggestions for artificial intelligence face recognition.

The research is divided into six sections. Section I is the introduction, Section II is about related work. Research on Face Recognition image algorithm is mentioned in Section III, results and discussion in Section IV and Section VI concludes the paper.

II. RELATED WORK

With the development of artificial intelligence, face recognition technology is gradually applied to all walks of life, and it is also a very important link in biometrics. Compared with fingerprint recognition, iris recognition and other recognition technologies, it can better meet the needs of users, and the recognition accuracy and recognition speed are very fast [9]. The current face recognition technology has many defects in practical application. How to improve face recognition has become the focus of current research. Based on this, scholars at home and abroad of Everbright have conducted in-depth research on it. Chen et al. proposed a new Collaborative Representation Based Fuzzy Discriminant Analysis (CRFDA) algorithm based on collaborative representation, and effectively extracted the relevant features of the image through dimension reduction, which effectively improved the feature extraction standard and improved the face recognition accuracy [10]. Alami et al. proposed a new model of quaternion discrete orthogonal matrix neural network, thereby reducing the time consumption of the model in face recognition training, and further effectively improving the accuracy of color face recognition [11]. Tripathi et al. effectively improved the accuracy of face age recognition by using descriptors in local gradient relationship patterns [12]. Singh et al. put forward a new and robust description of color texture, thus emphasizing the advantages of related color

models in identifying color components, so as to help improve the accuracy of color face recognition [13]. Soni et al. proposed a hybrid optimization algorithm based on bird search to remove image noise and improve face recognition accuracy [14]. Sharma et al. designed a face recognition system based on field programmable gate array to improve the security of face recognition [15], aiming at the problem of imperfect user data protection at present. Long et al. have effectively improved the recognition performance of face recognition by using singular value decomposition (SVD) to solve the problem of poor face recognition technology [16].

In addition, Sharma et al. used the depth learning method to achieve multi-modal determination of personal recognition and improve the robustness of face identification [17]. On the basis of deep learning, Han et al. proposed a new personalized convolution method, which significantly improved the efficiency of face recognition [18]. Srivastava et al. used deep learning to build a hybrid model to improve the accuracy of personal face recognition involving violence [19]. Zhao et al. aimed at the problem that face recognition takes too long; they reduced the time consumption in face recognition by building an unsupervised deep learning network [20]. Vedantham has constructed a classifier by using depth learning to improve the accuracy of facial expression recognition [21]. Deshmukh et al., aiming at the low efficiency of biometric feature extraction in face recognition, built a multi-mode biological learning system using deep learning networks to improve the efficiency of biometric feature extraction [22]. Silva et al. have effectively improved the accuracy of individual identification of wild Asian elephants by using deep learning methods, thus providing help for the protection of wild Asian elephants [23].

From the research of scholars at home and abroad, the current face recognition technology has problems such as long model training time, high resource consumption in practical applications, and deep learning has a good effect on collaborative operation. Therefore, it is of great significance to study the algorithm proposed by combining the depth learning algorithm with human face recognition in practical application. In addition, the face recognition algorithm proposed in the study creatively uses the relevance of multi task reflection to achieve the multi task recognition requirements of the algorithm, which effectively changes the simplification of traditional face recognition.

III. Research on Face Recognition Image Algorithm Based on Multi-Task Deep Learning

A. Analysis of Convolutional Neural Networks in Multi-Task Deep Learning

Aiming at the problems of slow training time and large system resource consumption for the current face attribute recognition task, the research conducted a related analysis on the face recognition image algorithm on the basis of multi-task deep learning. Deep learning includes two parts: deep and learning. Specifically, deep learning is a deep network model and has a suitable training learning method. The current fast-developing deep learning method is CNN, and for multi-task depth in terms of learning, convolutional neural networks are the basis for face recognition [24]. CNN uses a large amount of data for learning, and performs multiple multi-level convolution and non-linear mapping on it, so as to achieve the purpose of feature extraction or classification. It has the advantages of high accuracy and strong robustness. In other words, the working principle of CNN is to simulate the recognition process of the human brain [25]. Strictly speaking, CNN imitates the human brain to process information in more detail, and its principle structure is shown in Fig. 1.

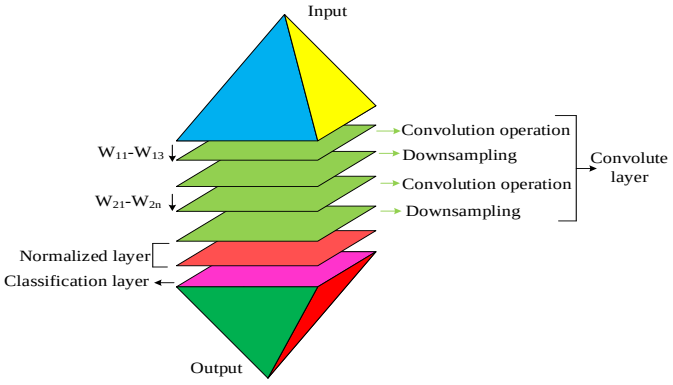


Fig. 1. Principle and structure of CNN.

It can be clearly found from Fig. 1 that the principle structure of CNN includes five levels of input, convolution, normalization, classification and output, which are progressive until the desired information is output. Specifically, after inputting the initial training data, CNN uses the reverse transfer algorithm to perform repeated learning, revise the parameters of different neurons, and finally obtain a convergent neural network model. On this basis, by inputting the image into the trained CNN model, a large number of abstract expressions ranging from low-level intuition to high-level can be obtained, and then through linear or nonlinear combination, accurate representation can be obtained to complete classification and feature extraction. Therefore, the basic structure of CNN can be divided into two parts, namely the feature extraction part and the fully connected part, and the more important ones are the convolution layer, the activation function layer and the fully connected layer (Softmax). At the signal level, the image of the convolution layer can be regarded as a visual signal, and its convolution operation can be understood as a process of filtering the signal, and its related calculation expression is shown in Eq. (1).

$$I(x,y'')*W(x,y'')=\sum_{s=-w}^{w}\sum_{t=-h}^{h}(s,t)I\left(x-s,y''-t\right) \quad (1)$$

In Eq. (1), I represents the pixel value; x represents the row; y'' represents the column; represents the convolution kernel; W the t parameter w represents the width of the convolution kernel; h represents s the height of the convolution kernel; At the matrix level, the convolution operation of the image is to use the convolution core to perform continuous convolution on the perceptual region of the image, and finally use it as the feature value of the region to obtain the feature map. The specific operation process is shown in Fig. 2.

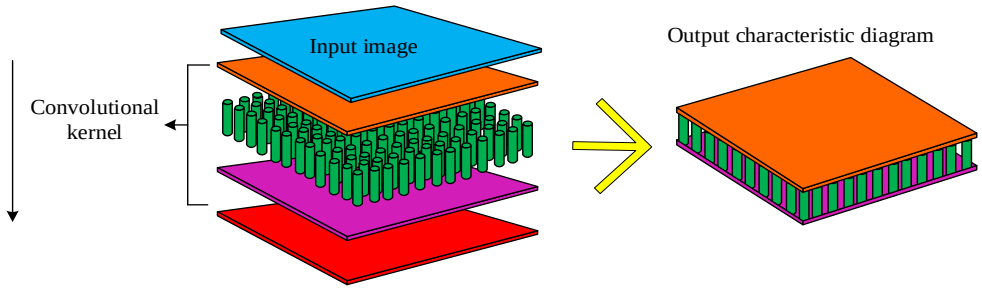


Fig. 2. Operation process of convolution.

It can be seen from Fig. 2 that the operation process of convolution is relatively simple. First, input the corresponding image into the convolution kernel, and then output the feature map after the operation. For the convolution layer, the convolution layer is usually composed of multiple convolution cores to better extract the image features. In CNN, the convolution layer uses a variety of convolution to check the local receiving area of the image, and obtains the feature maps of different convolution cores through the sliding convolution operation. It can be seen that CNN can effectively detect the features of the face in face recognition by recognizing the local feature map, thus improving the accuracy of face recognition. In this process, the corresponding calculation formula is shown in Eq. (2).

$$h_{ij}^{k} = f\left(\left(W_k * x'\right) + b_k\right) \tag{2}$$

In expression (2), x represents the convolution result, where i and j represents the position in the matrix, specifically represented as rows and columns, k represents the number of convolution kernels; f represents the activation function; x' represents the input image; b represents the convolution The corresponding bias of the kernel. The activation function layer is also known as the normalization layer. Its purpose is to use the relevant activation function to add corresponding nonlinear correlation factors, so as to ensure that the features are unique and invariant, thereby improving the overall expression ability of the model. The linear activation function is the Linear Rectification Function (RELU) [26]. Its related expression formula is shown in Eq. (3).

$$f\left(z\right) = \max\left(z, 0\right) \tag{3}$$

In Eq. (3), it $\max$ represents a comparison function that takes a larger value; z it represents the size of the input. However, in the actual training process of CNN, if the backpropagation presents a large gradient, if a RELU neuron flows through it, the neuron will always be in a "dead" state, which will as a result, the parameters of CNN are no longer continuously updated, thus consuming too many resources and related operations. The activation function selected for the study can be understood as a layer of network in the neural network, namely the Maxout activation function, which is relative to the sparse feature of the RELU activation function. It is more compact, and feature selection and dimensionality reduction can also be performed at the same time, and the calculation formula of its related neurons is shown in formula (4).

$$h_i\left(x\right) = \max_{j\in[1,k']} z'_{ij} \tag{4}$$

In Eq. (4), it $h_i\left(x\right)$ represents i the output of the neuron of the Maxout layer; it represents k' the number of input layers connected to the Maxout layer; j it represents the j input layer; z' it represents the input layer connected to the neurons of the Maxout layer. value, and its evaluation formula is shown in Eq. (5).

$$z'_{ij} = W_{\cdots ij} * x^{T} + b_{ij} \tag{5}$$

In Eq. (5), it $W_{\cdots ij}$ represents a certain convolution kernel among multiple convolution kernels; x^{T} it represents the input. For the Maxout layer, the dimensions of each CNN layer input must be consistent. On this basis, the maximum value is selected as the value of the same coordinate of the output layer after each input of the same coordinate is compared. This can be seen. The Maxout function is a piecewise function, and its input gradient formula is shown in Eq. (6).

$$\frac{\partial hi\left(x\right)}{\partial z'ij} = \begin{cases} 1 & if\ zij > ziq \\ When\, q \neq j\ and\ q \in [1,k] & \\ 0 & otherwise \end{cases} \tag{6}$$

In Eq. (6), q represents the input value. It can be clearly seen from the Eq. (6) that the slope exhibited by each segment of the Maxout function is determined by the output value, so it presents a dynamic change and does not completely fix a certain value. In addition, the Softmax layer is essentially a nonlinear classifier. In deep learning, it is often used as the output layer of CNN to output a probability vector. The output probability correlation calculation formula of the Softmax layer is shown in Eq. (7).

$$p\left(y' = i \middle| x', \theta\right) = \frac{e^{\theta j^{T} x}}{\sum_{j=1}^{k} e^{\theta j^{T} x}} \tag{7}$$

In the output probability calculation expression (7), θ represents the bias term; y' represents the output category, T represents transposition. On the whole, the Softmax layer is to map the multiple outputs after the original convolution to

the values 0 to 1 with the Sofmax function, and the accumulation of these values is 1, which satisfies the characteristics of probability. On this basis, it selects the final output. As a result, the largest prediction result can be selected to achieve multi-classification tasks, and in actual face recognition, multi-task face recognition can be effectively achieved. In face recognition technology, there are three very important technologies, which are the recognition of face age, gender, and fatigue status, which together lay the foundation for face recognition technology. Specifically, face recognition technology includes face-related feature extraction and comparison technology.

B. *Research on Multi-Task Face Recognition Algorithm based on Convolutional Neural Network*

CNN is a relatively common deep learning method, which is also the basis of multi-task face recognition algorithm. To understand multi-task face recognition algorithm, you need to understand the principle of multi-task learning. The current deep learning methods for multitasking mainly have the following two characteristics: one is structurally speaking, shallow parameter sharing between tasks; the other is that, in terms of impact, the common data characteristics hidden between different tasks are excavated. This requires that a model can handle multiple tasks at the same time, and also requires multiple tasks to work together to enhance the generalization ability. The overall structure is shown in Fig. 3.

It can be seen from the overall structure of multi task learning in Fig. 3 that its specific content includes model input, shared parameters and relevant parameters of specific tasks. By inputting relevant values in model input, shared parameters are obtained, and relevant parameters of specific tasks are output in subsequent output. An important premise of multi task learning is that there is a certain correlation between multiple tasks, so learning efficiency can be improved through the correlation between tasks. Although there is no clear definition, when looking for related tasks, a basic assumption is that the characteristics concerned by each task are interrelated, or the learning process is beneficial in different tasks. In the research, there is some correlation between facial features and features. Through the association training of multiple tasks, the accuracy of multiple tasks can be effectively improved. In the actual face recognition, the task is no single, and the improvement of the accuracy of multi task recognition can also effectively improve the overall recognition accuracy.

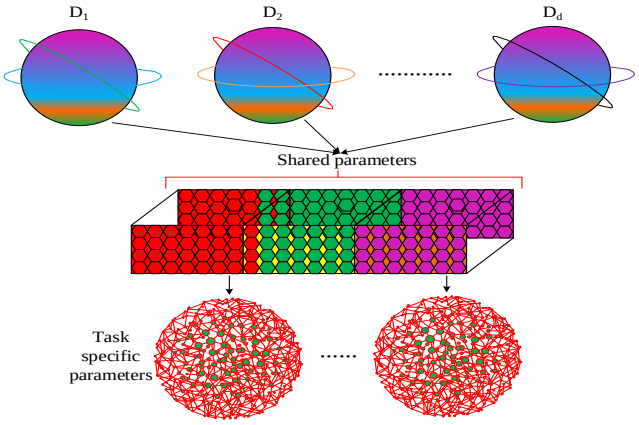


Fig. 3. The overall structure of multi task deep learning.

The traditional face recognition algorithm has the problem of low recognition task accuracy. Aiming at this problem, a multi-task deep learning face recognition algorithm is proposed. This algorithm uses the relatively strong correlation of face identity, age, and gender and fatigue recognition. To carry out research, in order to build a multi-task face recognition model in a short time, and to complete the relevant recognition tasks under the premise of maintaining high precision and high speed. High accuracy depends on the relevance of multiple learning tasks, and high-speed operation requires optimization of model parameters. Therefore, the research integrates CNN to construct a CNN network for multi-task face recognition. Its structure is shown in Fig. 4.

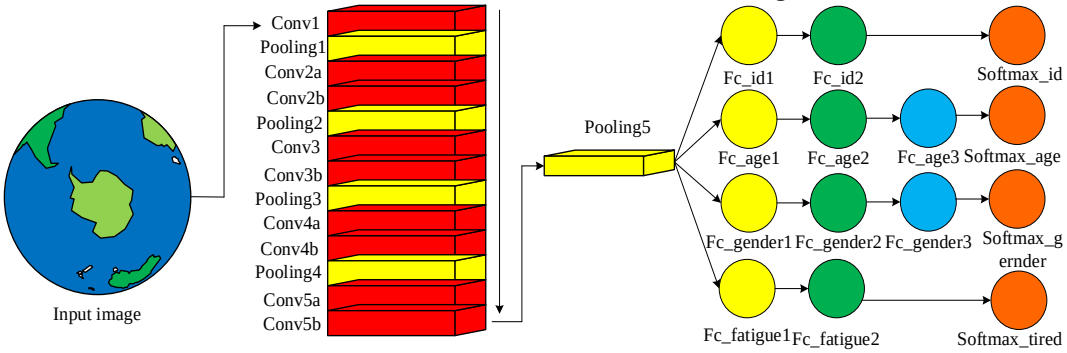


Fig. 4. Multi task face recognition network structure.

As can be seen from Fig. 4, the overall network structure of CNN has a total of nine convolutional layers, the pooling layer uses the maximum pooling function, the activation function selects the Maxout function, and all tasks in the training select the Softmax function to perform related classification training tasks for face recognition. It is worth noting that before constructing the actual network training, it is necessary to set the objective function and the training algorithm. This is because the work related to face recognition is closely related to the data, including identity, gender, age, etc. It is very difficult to generate a multi-label data set with information such as fatigue, fatigue, etc. Therefore, after setting the two in detail, the multi-label data set can be trained synchronously, and the multi-task training on the basis can be

obtained. Same result. In the proposed multi-task deep learning face recognition algorithm, a brief description of the relevant forward propagation expression is shown in Eq. (8).

$$output_{k'} = f(W_{k'}X + b_{k'}) \tag{8}$$

In the expression Eq. (8), it k' represents the first k' face recognition task; $W_{k'}$ it represents the convolution related parameters. In contrast, the formula expression of backpropagation is shown in Eq. (9).

$$W_{k'} = W_{k'} - \eta_{k'} \frac{\partial L_{k'}}{\partial W_{k'}} \tag{9}$$

In the Eq. (9) of backpropagation, η it represents the learning rate; L it represents the loss of identifying characters. It is worth noting that the overall back-propagation formula of the recognition task can be divided into two parts: the independent parameter part of a single task, whose back-propagation formula is consistent with Equation (9), and the back-propagation formula of the common parameter segment of multiple tasks is as in Eq. (10).

$$W_{share} = W_{share} - \sum_{k'=1}^{M} \eta_{k'} \frac{T_k \partial L_{k'}}{\partial W_{share}} \tag{10}$$

In Eq. (10), it W_{share} represents the shared parameters between multiple tasks; M it represents the total number of face recognition tasks; and $T_{k'}$ it represents the weight of the task. It can be seen from Eq. (10) that the training strategy given by the study is consistent with the training of multiple tasks at the same time, except that there is a sequence in the training time of each task. Therefore, for different recognition tasks, the specific training strategy not only grasps the three important foundations of face recognition, but also introduces identity recognition. In face recognition, the forward propagation calculation formula of the deep learning model CNN structure is shown in Eq. (11).

$$o = g\left(f_n\left(W_n \cdots f_2\left(f_1\left(W_1 X_0 + b_1\right) + b_2\right) \cdots + b_n\right)\right) \tag{11}$$

In Eq. (11), o represents the forward propagation training value; W_n represents the n weight of b_n the first n layer; represents the displacement bias of the first layer; and $g()$ represents the final classification function. The ultimate goal of network training is to obtain the minimized loss function value loss, so the Softmax-loss function is selected to calculate the loss rate, and its specific formula is shown in Eq. (12).

$$Loss\left(y, o_y\right) = -\log\left(o_y\right) = -\log\left(\frac{e^{z_y}}{\sum_{j=1}^{m} e^{z_j}}\right) \tag{12}$$

In Eq. (12), it y represents the real label of the face image; o_y it represents the probability value of the position in the output probability vector when Softmax outputs y. In addition, in face age recognition, a unique loss function is selected to speed up the training process, and its specific expression is shown in Eq. (13).

$$G\_Loss\left(y, o_y\right) = \mu G(y, o_y) + \left(1-\mu\right) Loss\left(y, o_y\right) \tag{13}$$

In Eq. (13), it μ represents the weight of loss; $G(y, o_y)$ it represents the special loss function selected by the research, and its calculation formula is shown in Eq. (14).

$$G(y, o_y) = \left(1 - \exp\left(-\frac{o_y - y}{2\delta^2}\right)\right) \tag{14}$$

In Eq. (14), it δ represents the standard deviation of the age distribution in the training data set, and its calculation formula is shown in Eq. (15).

$$\delta = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu'\right)^2} \tag{15}$$

In Eq. (15), it N represents the total number of pictures of human faces; x_i represents i the age of the th face photo; μ' represents the mean value.

IV. Application and Performance Analysis of Face Recognition Image Algorithm based on Multi-Task Deep Learning

In order to verify the performance of the multi task deep learning face recognition algorithm proposed by the research, the research selects the face recognition training data set (CASIA Webface), human age recognition data set (IMDB-WIKI 500k+), face gender recognition data set (Celeba) and face fatigue recognition data set (bus driver face fatigue data set) before the experiment, and the test sets are respectively IJB-A, FG-NET, Celeba and human eye closure data set (CEW). It is worth noting that the data sets selected for the study are all found on the Internet, part of which are internet data (such as Wikipedia) and part of which are intranet data. In addition, in the experiment, the deep learner Aki chose Caffe, and the processor chose Intel (R) Core (TM) 7-7700 CPU @ 3.6GHz, with 8-core 8GB of memory; Select 64 bit Ubuntu 16 04 for the operating system; Choose NVIDIA GeForce GTX1070 graphics card with 8G memory; Select the Python 27 environment and Anaconda related scientific computing library for the development environment and tools. At the same time, before the experiment, the dataset was preprocessed for possible missing faces or imbalanced data, mainly through three steps: affine transformation, random cropping, and data balancing. The specific content of the dataset is shown in Table I.

It can be clearly seen from Table I that in the training data set selected for the experiment, the CASIA-Webface data set contains 494,414 identity photos of 10,575 people, and the training data set IJB-A data set contains the face pictures of 500 people, including 5,396 still images, 20,412 frames of video images; IMDB-WIKI 500k+ dataset contains 524,230 face images of 82,612 people, of which 460,723 are from the Internet Movie Database (IMDB) and 62,328 are from WiKi ,

its test set contains 1,002 photos of 1,297,028 people; the Celeba data set divides 202,599 face images of 10,177 people into training and test sets in a ratio of 9:1; the bus driver's face fatigue data set contains There are 47,500 photos of 10,100 individuals, including 10,000 photos of fatigued state, 30,000 photos of normal state, and 7,500 photos of standing wearing sunglasses. The test set CEW contains 2,423 photos of 40 individuals, including 1,192 photos of closed eyes, and 1,231 photos of closed eyes. Open-eye photos, closed-eye photos are obtained from a web crawler, and open-eye photos are obtained from the face database (Labeled Faces in the Wild, LFW). According to the four parts of the dataset, the research first analyzes the face recognition task, and its experimental results on the training set LFW are shown in Fig. 5.

TABLE I. SPECIFIC CONTENTS OF TRAINING DATA SET AND TEST DATA SET SELECTED BY THE EXPERIMENT

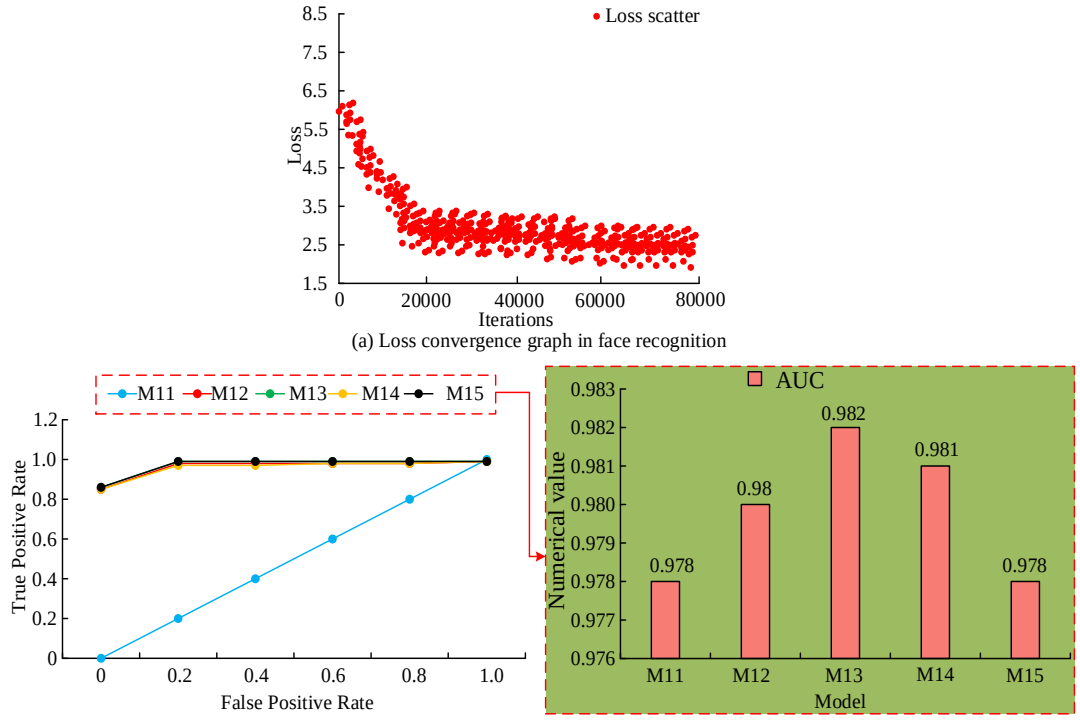
Training Set	Number of people	Number of photos			Test Set	Number of people	Number of photos	
CASIA-Webface	10575	494414			IJB-A	500	5396 (static state)	20412 (Video image)
IMDB-WIKI 500k+	82612	460723 (IMDB)	62328 (Wiki)		FG-NET	1297028	1002	
Celeba	10177	182339			Celeba	10177	20260	
Bus drivers face fatigue	10100	10000 (Fatigue state)	30000 (Normal state)	7500 (Wear sunglasses)	CEW	40	1192 (Closed eye photograph)	1231 (Eye opening photo)

In Fig. 5, Fig. 5(a) is the loss function curve obtained by using the loss function. Using this loss function curve, the model to be verified for face identification is divided into five models, which are respectively used M_{11} to M_{15} represent, and through the analysis of the five models The model's rate of change (Rate of Change, ROC) and the area under it (Area Under Curve, AUC) are compared, and Fig. 5(b) is obtained. It can be seen from Fig. 4 that the AUC value is the highest M_{13}, and the AUC value is 0.982 at this time. Therefore, the study uses it as the subsequent face age recognition pre-training model, and the relevant formula is used to obtain the threshold value of 0.35, and the accuracy rate is as high as 92 %. The experimental results show that the preprocessing model can effectively extract the identity features of the face. In the face gender experiment, the research uses the face identification with the highest AUC value M_{13}, trained for four hours in the same equipment environment, and obtained the corresponding loss curve, and also obtained five models according to the obtained loss function curve, respectively. Used M_{21} to M_{25} represent, and the corresponding face gender experiment is on the model with the best performance in the face age recognition task, after 10 minutes of training, the obtained 5 models are respectively used M_{31} to M_{35} represent, the two are respectively in LG- The experimental results on the NET and Celeba training sets are shown in Fig. 6.

Fig. 6 (a) is the loss convergence scatter and Mean Absolute Error (MAE) obtained in the face age recognition training process, and Fig. 6(b) is the loss obtained in the face gender recognition training process Convergence curve and face recognition accuracy value (Accuracy, Acc). It can be seen from Fig. 6(a) that the MAE minimum value appears at the model M_{22}, and the algorithm has the highest accuracy at this time, so it can be used as the pre-training model for face gender recognition in Fig. 6(b). It can be seen from Fig. 6(b) that the Acc value M_{34} is the highest at the model, which is 0.983. The experimental results show that with the continuous optimization of the model accuracy, the training time for the algorithm to obtain loss convergence is decreasing, indicating that the error of the algorithm is decreasing, and it is continuously improving the accuracy of the model until it is optimal. Finally, the loss convergence curve was obtained after only two minutes of training in human fatigue state recognition. The experimental results on the CEW dataset and the total convergence time of the four identification experiments are shown in Fig. 7.

It can be clearly seen from Fig. 7 that the recognition accuracy of face fatigue M_{44} appears on the model, which is 0.983. In addition, in the introduction of three deep learning models (Facenet), Face Recognition Based on Visual Geometry Group Network (Vggface) and Levi G's four best models in the research itself From the comparison, it can be found that on the basis of the algorithm proposed by the research, the four tasks can complete the continuous training of the optimal mode within 10 hours, which is much shorter than the training time of deep CNN networks such as Facenet and Vggface; Compared with the face feature recognition method proposed by Levi [27] et al., this method can only recognize the age and gender of the face, but the algorithm proposed in the study can complete the face recognition identity within the same training time. On the basis of gender, it can also realize the recognition of face age and fatigue state. The research results show that the multi-task training method can effectively reduce the completion time of training tasks.

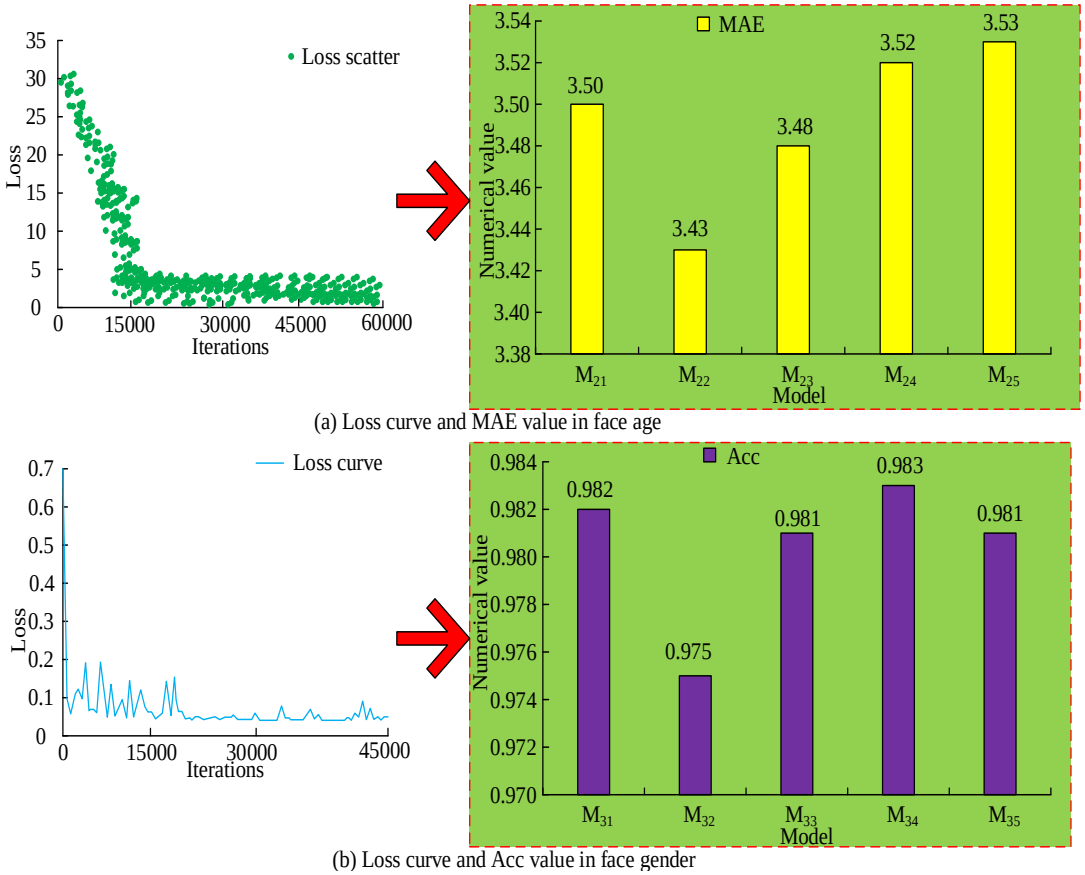
In order to further verify the effectiveness of the proposed algorithm, the research M_{44} evaluates the performance of the final multi-task face recognition model. Chen algorithm, Webface algorithm and block cipher algorithm (Gosudarstvennyi Standard, GOST) algorithm are introduced again in the experiment. The experimental results are shown in Table II.



(a) Loss convergence graph in face recognition

(b) ROC value change and AUC value of the model in 5

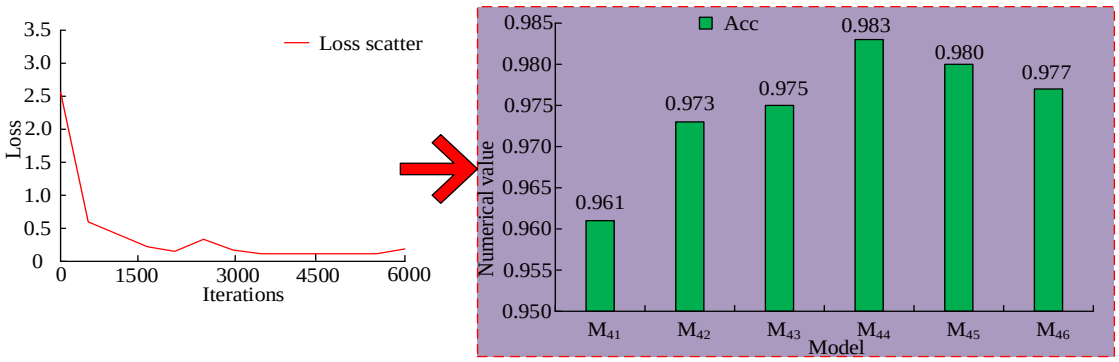
Fig. 5. Training process of face recognition.



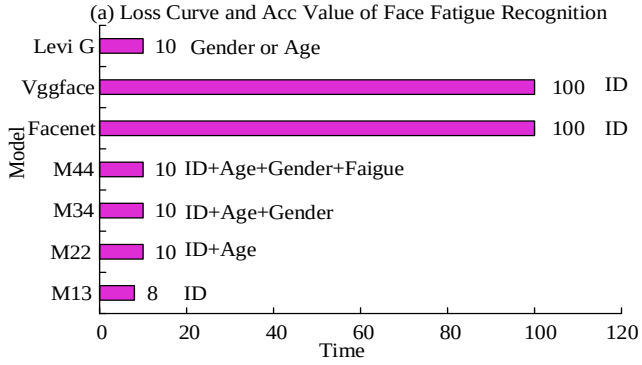
(a) Loss curve and MAE value in face age

(b) Loss curve and Acc value in face gender

Fig. 6. Face age and gender training process.



(a) Loss Curve and Acc Value of Face Fatigue Recognition



(b) Convergence time of four kinds of identification to reach the best model and time of partial depth algorithm

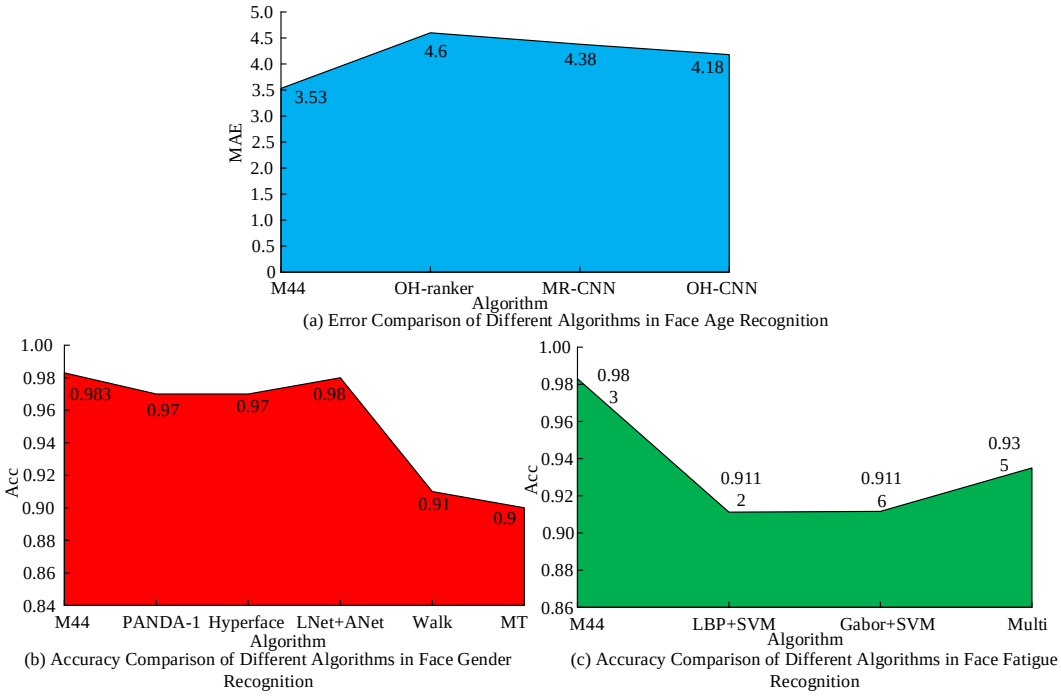
Fig. 7. Experimental results on the CEW dataset and the time spent on four face recognition experiments and three depth algorithms.

TABLE II. PERFORMANCE COMPARISON OF MULTIPLE ALGORITHMS

Model	Feature extraction time	Number of parameters	Enter image dimensions	Accuracy (%)	FAR=0.1	FAR=0.01	FAR=0.001
$M44$	41ms	5555K	64×64×1	97.88 0	0.951	0.844	0.790
Chen	67ms	5004K	100×100×1	97.71 0	0.966	0.837	-
Webface	-	5011K	31×36×3	97.49 0	-	-	-
Facent	>558ms	140690K	224×224×3	99.62 0	0.950	0.834	0.750
Vggface	>528ms	134249K	224×224×3	97.27 0	0.936	0.804	0.603
GOST	-	-	-	-	0.620	0.400	0.190

As can be seen from Table II, in the LJB-A dataset with more complex actual scenes, M_{44} the model performance of the model is better than other models, and the accuracy reaches 97.880%. In addition, when the False Accept Rate (FAR) value is 0.01, the True Accept Rate (TAR) value at this time is 0.844, indicating that the M_{44} model has higher robustness in actual complex situations, which can distinguish face identities well. At the same time, the research compares the performance and accuracy of different algorithms (methods) in the recognition of face age, gender, and fatigue state. Among them, multi region convolution neural network (MR-CNN) and two ranking based methods are introduced into face age recognition, namely OH ranker and OR-CNN; panda recognition (PANDA-1), hyperspectral face recognition (Hyperspectral Face Recognition, Hyperface), LNet+ANet, based on deep learning are introduced in face gender recognition Face detection and alignment based on depth learning (MT) and walk recognition (Walk); in face fatigue recognition, Support Vector Machines (SVM) are added, and linear inverse Projection algorithm (Local Binary Patterns, LBP), Gabor and multimedia algorithm (Multi), the specific experimental results are shown in Fig. 8.

It can be clearly seen from Fig. 8 that M_{44} the value of the model MAE is 3.53, which is much lower than other algorithms; in addition, its ACC value is 98.3% in face gender and fatigue recognition, which is also higher than other algorithms. The experimental results are obvious. Compared with other face recognition algorithms and depth algorithms, the performance of the proposed algorithm has higher recognition performance, and it also speeds up the training time and improves the recognition accuracy.



(a) Error Comparison of Different Algorithms in Face Age Recognition

(b) Accuracy Comparison of Different Algorithms in Face Gender Recognition

(c) Accuracy Comparison of Different Algorithms in Face Fatigue Recognition

Fig. 8. Performance comparison of different algorithms in face recognition.

V. Results and Discussion

The demand for multitasking facial recognition has significantly improved in many current fields. In order to achieve multitasking facial recognition, existing technologies generally use multiple models in parallel to complete multiple different recognition tasks, which may have serious problems in some cases. Firstly, multiple models have slow training time and long application cycles. Secondly, in practical use, there is a problem of high system resource consumption, which also means high hardware requirements and increased usage costs. Based on this, a multi task face recognition algorithm was proposed by combining CNN with multi task deep learning.

The experimental results show that the Acc value is the highest at model M34, at which point it is 0.983. As the model accuracy is continuously optimized, the training time for the algorithm to achieve loss convergence is continuously decreasing. When the FAR value is 0.01, the correctly accepted proportional TAR value is 0.844, indicating that the M44 model has higher robustness in practical complex situations and can distinguish facial identities well, which is better than the results of Basil et al. [28]. In addition, the MAE value of the M44 model is 3.53, which is much lower than other algorithms. Its ACC value in facial gender and fatigue recognition is 98.3%, which is also higher than other algorithms. This result is also superior to the results of Vu et al. and Mohammed Ali et al. [29-30].

In the dataset LFW, the highest AUC value appears in M13, at which point the AUC value is 0.982; In the dataset LG-NET and Celeb, the minimum MAE value appears at model M22, the Acc value is the highest at model M34, and the accuracy of facial fatigue recognition appears at model M44, which is 0.983. This result shows that the research algorithm has high performance on different datasets, and the final optimal model has the best performance, indicating the generalization of the algorithm.

Overall, the algorithm proposed in the study not only consumes less training time, but also has higher accuracy and lower error values, indicating good performance.

VI. Conclusion

In order to solve the problems of long training time and high consumption of face recognition, the research proposes a multi-task face recognition algorithm based on CNN and multi-task deep learning, and conducts experimental analysis on its performance and application. Four dimensions of face identity, age, gender and fatigue state are used to verify the performance of the proposed algorithm. The experimental results show that the AUC value of the best model for face identity recognition is 0.982, indicating that the extracted face identity features are very effective; in face age recognition, the absolute error value of the best model is 3.43, while the in the gender experiment, the highest ACC value was 0.983, the average accuracy rate was 98.04%, and the average accuracy rate of fatigue state recognition also reached 97.5%. In addition, in the experiment of the same four dimensions, when comparing the performance and accuracy of it with other algorithms, the time-consuming feature of the proposed algorithm for face identification is only 41ms, which is much lower than other algorithms; face age recognition The MAE value of Acc is 3.53, the Acc value of gender recognition is 0.983, and the Acc value of fatigue state recognition is also 0.983. The three actually show far lower errors than other algorithms and far higher accuracy than other algorithms. On the whole, the proposed algorithm not only consumes less training time, but also has higher accuracy and lower error value, and has better performance. However, although the

study utilized the Maxout function to extract features, there is still room for expanding the inter class spacing between different individuals in facial identity recognition tasks. Later work can incorporate the triplet loss function for further research.

FUNDINGS

The research is supported by: 2017 Jiangsu Education Informatization Project: Research on Building a New Lifelong Learning Platform in the Context of Internet+(20172075); The 13th Five Year Plan for Educational Science in Jiangsu Province: Research on Building Jiangsu Intelligent Lifelong Learning Service Platform in the "Internet+" Era (B-b/2018/01/33).

REFERENCES

[1] Liu Q, Wang Y, Ye B, Ding M. Recognition Confidence of Welding Seam Defects in TOFD Images Based on Artificial Intelligence. Automatic Control and Computer Sciences, 2022, 56(2): 180-188.

[2] Nasir N, Kansal A, Barneih F, Al-Shaltone O, Bonny T, Al-Shabi M, Al Shammaa A. Multi-modal image classification of COVID-19 cases using computed tomography and X-rays scans. Intelligent Systems with Applications, 2023, 17: 200160.

[3] Tahoun N, Awad A, Bonny T. Smart assistant for blind and visually impaired people//Proceedings of the 3rd International Conference on Advances in Artificial Intelligence. 2019: 227-231.

[4] Bonny T, Haq A. Emulation of high-performance correlation-based quantum clustering algorithm for two-dimensional data on FPGA. Quantum Information Processing, 2020, 19: 1-21.

[5] Tamilselvi M, Karthikeyan S. An ingenious face recognition system based on HRPSM_CNN under unrestrained environmental condition. Alexandria Engineering Journal, 2022, 61(6): 4307-4321.

[6] Khan AA, Shaikh AA, Shaikh ZA, Laghari AA, Karim S. IPM-Model: AI and metaheuristic-enabled face recognition using image partial matching for multimedia forensics investigation with genetic algorithm. Multimedia Tools and Applications, 2022, 81(17): 23533-23549.

[7] Srivastava S, Kumar A, Singh A, Prakash S, Kumar A. An improved approach towards biometric face recognition using artificial neural network. Multimedia Tools and Applications, 2022, 81(6): 8471-8497.

[8] Karanwal S. Robust local binary pattern for face recognition in different challenges. Multimedia Tools and Applications, 2022, 81(20): 29405-29421.

[9] Marmolejo-Ramos F, Murata A, Sasaki K, Yamada Y, Ospina R. Your Face and Moves Seem Happier When I Smile: Facial Action Influences the Perception of Emotional Faces and Biological Motion Stimuli. Experimental Psychology, 2020, 67(1): 14-22.

[10] Chen C, Zhou X. Collaborative representation-based fuzzy discriminant analysis for Face recognition. The Visual Computer, 2022, 38(4): 1383-1393.

[11] Alami AE, Berrahou N, Lakhili Z, Mesbah A, Berrahou A, Qjidaa H. Efficient color face recognition based on quaternion discrete orthogonal moments neural networks. Multimedia Tools and Applications, 2022, 81(6): 7685-7710.

[12] Tripathi RK, Jalal A SA robust approach based on local feature extraction for age invariant face recognition. Multimedia Tools and Applications, 2022, 81(15): 21223-21240.

[13] Singh C, Majeed S. Novel and robust color texture descriptors for color face recognition. Multimedia Tools and Applications, 2022, 81(15): 21313-21347.

[14] Soni N, Sharma EK, Kapoor A. Deep neural network and 3D model for face recognition with multiple disturbing environments. Multimedia Tools and Applications, 2022, 81(18): 25319-25343.

[15] Sharma V, Joshi A M. VLSI Implementation of Reliable and Secure Face Recognition System. Wireless Personal Communications, 2022, 122(4): 3485-3497.

[16] Long X, Zhang Z, Li YA singular value decomposition representation based approach for robust face recognition. Multimedia Tools and Applications, 2022, 81(6): 8283-8308.

[17] Sharma A, Jindal N, Thakur A, Rana PS, Garg B, Mehta R. Multimodal Biometric for Person Identification Using Deep Learning Approach. Wireless Personal Communications, 2022, 125(1): 399-419.

[18] Han C, Shan S, Kan M, Wu S, Chen X. Personalized Convolution for Face Recognition. International Journal of Computer Vision, 2022, 130(2): 344-362.

[19] Srivastava A, Badal T, Saxena P, Vidyarthi A, Singh R. UAV surveillance for violence detection and individual identification. Automated Software Engineering, 2022, 29(1): 1-28.

[20] Zhao R, Shi F. I2DKPCN: an unsupervised deep learning network. Applied Intelligence, 2022, 52(9): 9938-9951.

[21] Vedantham R. Adaptive increasing-margin adversarial neural iterative system based on facial expression recognition feature models. Multimedia Tools and Applications, 2022, 81(3): 3793-3830.

[22] Deshmukh S, Abhyankar A, Kelkar S. DCCA and DMCCA framework for multimodal biometric system. Multimedia Tools and Applications, 2022, 81(17): 24477-24491.

[23] Silva ED, Kumarasinghe P, Indrajith K, Pushpakumara TV, Vimukthi RDY, Zoysa K De, Gunawardana K, Silva S De. Feasibility of using convolutional neural networks for individual-identification of wild Asian elephants. Mammalian Biology, 2022, 102 (3): 909-919.

[24] Mageshkumar N, Lakshmanan L. An improved secure file deduplication avoidance using CKHO based deep learning model in a cloud environment. The Journal of Supercomputing, 2022, 78(13): 14892-14918.

[25] Kuo YL, Tang S C. Deep regression of convolutional neural network applied to resolved acceleration control for a robot manipulator: Transactions of the Institute of Measurement and Control, 2022, 44(4): 784-798.

[26] Liang C, Zhang Z, Wu Q, Li X. Barrier Lyapunov function-based robot control with an augmented neural network approximator. Industrial Robot: the international journal of robotics research and application, 2022, 49(2): 359- 367.

[27] Levi G, Hassner T. Age and gender classification using convolutional neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2015: 34-42.

[28] Basil N, Raad M, Wazzan A N, Marhoon H M. Face recognition with real-time framing based on multi task convolutional neural network: a case study. Int. J. Mech. Eng, 2022, 7(2): 3170-3178.

[29] Vu H N, Nguyen M H, Pham C. Masked face recognition with convolutional neural networks and local binary patterns. Applied Intelligence, 2022, 52(5): 5497-5512.

[30] Mohammed Ali F A, Al-Tamimi M S H. Face mask detection methods and techniques: A review. International Journal of Nonlinear Analysis and Applications, 2022, 13(1): 3811-3823.

A Model for Analyzing Employee Turnover in Enterprises Based on Improved XGBoost Algorithm

Linzhi Nan[1], Han Zhang[2*]
School of Economic and Trade, Haojing College of Shaanxi University of Science and Technology, Xi'an, 710000, China[1]
Shaanxi Hantong Consulting Services Co., LTD, Xianyang, 712000, China[2]

***Abstract*—To accurately predict the possibility of employee turnover during enterprise operation and improve the benefits created by talents in the enterprise, research based on the limit gradient enhancement algorithm has received widespread attention. However, with the exponential growth of various types of resignation reasons, this algorithm is not comprehensive enough when dealing with complex character psychology. To solve this problem, this study uses the limit gradient enhancement algorithm to predict employee turnover in the Company dataset, and uses differential automatic regression moving average variable optimization to generate a fusion algorithm. The research first involves stepwise regression processing of the training data, expanding the objective function to a second-order Taylor expansion; Then variance coding is added to the square integrable linear white noise, and the step cooling curve is smoothed by changing the temperature control constant; Then to calculate the root mean square error of Newton's law of cooling, and obtain its derivative loss variable. Linear white noise is the chaotic data produced by the improved extreme gradient lifting algorithm in forecasting the original data of enterprise employees, which will affect the results of data preprocessing in the loss analysis. In order to reduce the operation error of the algorithm, the step cooling curves are drawn according to the cooling law, and then their root mean square errors are calculated. Finally, the fusion algorithm studied was applied to the Company dataset and the prediction accuracy of the particle swarm optimization algorithm was tested and compared with the fusion algorithm. A total of 400 experiments were conducted, and the fusion algorithm achieved a prediction accuracy of 398 times, with an accuracy rate of 99.5%; The accuracy of particle swarm optimization algorithm is close to that of fusion algorithm, at 83.2%. The experimental results indicate that the algorithm model proposed in the study can accurately predict the possibility of employee turnover in enterprises, and the company will also receive timely information to make the next budget step.**

***Keywords*—Data preprocessing; linear white noise; root mean square error; newton's law of cooling; step cooling curve**

I. INTRODUCTION

In the rapidly advancing society of internet technology, as competition between companies intensifies, the number of employees in enterprises is gradually receiving widespread attention, and the requirements for algorithms are also increasing [1-2]. In the operation of the enterprise, talents rely on interpersonal relationships and key technologies to continuously create profits for the enterprise. Such employees will be widely valued within the company, but leaving is also inevitable. The resignation of employees is not conducive to the development of the enterprise, and may even lead to technical gaps that can lead to internal management problems. So, corporate executives are gradually paying attention to the reasons for employee turnover. In this context, algorithms for analyzing employee psychology have received widespread attention. In recent years, the Limit Gradient Lifting Algorithm (XGBoost) has attracted the attention of many scholars due to its powerful self-checking ability [3]. However, XGBoost is only suitable for analyzing differentiable loss variables. For non-differentiable variables, the algorithm marks them as isolated points, which affects the accuracy of the detection data. To address this issue, this study is based on Differential Autoregressive Moving Average Variable (DAMAV) to optimize XGBoost and generate a fusion algorithm (DV-XGBoost) pioneering. This algorithm calculates the root mean square error of non-differentiable variables and can convert non differentiable variables into differentiable loss variables. For the experimental design of the analysis model of employee turnover in enterprises, the research first collects the human resources data of enterprises, and determines the completeness of employee turnover prediction according to the relevant experience of employee turnover in the past; Then clean the collected data to ensure the availability of the data, and extract the characteristics of the data according to the business needs of the enterprise. For the key characteristics of employee turnover, DV-XGBoost algorithm is used to judge it, and the data is divided into training set and test set. The evaluation indexes include accuracy, accuracy and F_1 value. The innovation of this study is mainly reflected in the following two points. First, aiming at the limitations of XGBoost algorithm, this study proposed a fusion algorithm based on DAMAV to optimize XGBoost (DV-XGBoost). This fusion algorithm shows its unique innovation when dealing with underivable variables. The second innovation is that for the non-derivable variable, the traditional XGBoost algorithm will mark it as an isolated point, while the DV-XGBoost algorithm converts it into a derivable loss variable by calculating the root-mean-square error of the non-derivable variable. This method is innovative in solving the problem of non-derivable variables. The contribution of this research is mainly reflected in the following three points. First, DV-XGBoost algorithm improves the accuracy of detection data by converting non-derivable variables into derivable loss variables. It is of great practical significance for enterprises to analyze the reasons of employee dimission and then make corresponding strategies. The second is to broaden the application range of XGBoost, a powerful algorithm that is widely used in all kinds of data contests and real-world

problem solving. However, it is only suitable for analyzing derivable loss variables; DV-XGBoost algorithm further broadens the application range of XGBoost by solving the problem of non-derivable variables. Finally, it can promote the progress of talent management, and enterprises can better understand the reasons for employees' dimission, so as to improve the company's talent management strategy, reduce the employee dimission rate, improve employee satisfaction, and promote the stable development of enterprises.

The research is mainly divided into six sections. Section I mainly analyzes and summarizes the application and effectiveness of the current XGBoost model. Section II introduces the factors that affect the employee turnover model and constructs the DV-XGBoost model. Section III is the experimental study on enterprise characters. Result and discussion is mentioned in Section V and Section VI concluded the paper.

II. Related Works

A very important branch of human psychological prediction technology, namely employee turnover prediction in enterprises, plays a very important role in computer learning and the rational use of human resources in enterprises [4]. Lu et al. designed real driving tasks to extract data and proposed a stress monitoring model based on driving behavior. Driving is described by the acceleration of the vehicle, and the driving environment is quantified using an extended residual network model. According to the distribution range of driver ambiguity, the video image is segmented into sub regions. They constructed an Extreme Gradient Enhancement (XGBoost) model to monitor stress, and compared it with other models, the XGBoost model outperformed mainstream learning algorithms. It can also surpass most traditional models without using psychological data [5]. Deng et al. combined XGBoost and multi-objective optimization genetic algorithm for cancer classification. They first sort genes based on XGBoost ensemble selection, effectively removing unrelated genes and generating the most relevant genome for this class; then, their fusion algorithm searches for the optimal subset based on gene groups. They conducted comprehensive experiments using learning classifiers on publicly available microarray datasets to compare state-of-the-art feature selection methods. Their experiments have shown that the algorithm outperforms existing algorithms in multiple evaluation indicators such as accuracy, precision, and recall [6]. Li et al. developed a reliable prediction method that estimates the sink area of the road surface based on the main feature of road surface temperature. They proposed chaotic particle swarm optimization and segmented regression strategy to optimize the XGBoost model. Compared with the classical learning algorithm, their experimental results show that the root mean square error and absolute error of XGBoost algorithm are increased by 5.80 and 1.59 respectively. This algorithm has obvious advantages in dealing with nonlinear problems, and can also reduce the frequency of deflection without affecting its estimation accuracy, promoting rapid evaluation of road conditions [7]. Tao et al. proposed a robust method for diagnosing turbine blade icing. They extracted features of short-term icing effects based on icing physics to establish stacked XGBoost models for blade icing diagnosis. They evaluated the methods proposed in the wind farm and further compared them with models based on a single algorithm. The results indicated that their mixed features enhance the similarity between different datasets, and their model has higher accuracy and better generalization power compared to models based on a single algorithm [8].

Li et al. proposed an orthopedic classification prediction model based on the XGBoost algorithm. After building the XGBoost model, they also built the same model based on the random forest algorithm, and made a comparative analysis of them. Compared with random forest model, XGBoost algorithm prediction model has higher accuracy and is more suitable for orthopedic clinical data. The XGBoost algorithm can handle diverse medical data and better meet the requirements of diagnostic timeliness and accuracy [9]. Gu et al. proposed a new LC decision model that enables automated vehicles to make human decisions. Their method combines a deep encoder network with the XGBoost algorithm, using time series from multiple sensors to establish a robust multivariate reconstruction model; then, they reconstructed the error using normal data for training data extraction. They adopted the XGBoost algorithm with Bayesian optimization for the multi-parameter problem of autonomous LC decision-making process. This model can accurately identify the LC behavior of vehicles [10]. Li et al. cross matched the bass with the spectral database of the Sloan Digital Telescope to obtain the spectral categories of known samples. Then, the samples were cross matched with the ALLWISE database, and they constructed different classifiers using the XGBoost algorithm based on the optical and infrared information of the samples. Finally, all selected items in the bass directory are classified by these classifiers. When the prediction results of binary classification are the same as those of multi class classification, the prediction results of light sources without infrared information can be used as a reference. Their classification results have great reference value for future research [11]. Osman et al. developed a model for predicting groundwater levels. They tested three machine learning models: Xgboost, artificial neural network, and support vector regression. The experiment shows that if the combination of rainfall data with a delay of 3 days is used as input, the performance of the model is the worst; For all input combinations, their proposed Xgboost model outperforms the other two models. When using groundwater level with a 1-day delay as input, the performance of the Xgboost model is significantly improved. Their research results provide application prospects for the Xgboost algorithm to predict future groundwater levels [12].

The research of multiple scholars mentioned above has found that the predictive ability of Xgboost algorithm is very popular internationally, but there is still little research on fusion algorithms. This study pioneered the introduction of differential autoregressive moving average variables, taking into account the impact of employee personal factors and internal control within the enterprise, and generated a fusion algorithm (DV-XGBoost).

III. Construction of Employee Turnover Information Model based on XGBoost

With the development of the human resources industry, in order to rationalize the distribution of employees in enterprises, research on various information has become increasingly popular, and the probability of employee transfer is the most important part [13]. However, objective factors such as the variety of resignation data and the complexity of employee psychology have increased the workload of predicting employee turnover information devices. This study combines the XGBoost algorithm with DAMAV, first introducing a model built on XGBoost, and then describing the fusion method of the two.

A. Establishment of XGBoost Model for Enterprise Employee Resignation

Before analyzing data on employee turnover in enterprises using models, it is first necessary to preprocess the data. Preprocessing is the most complex part of the experiment, and the results of this part not only occupy time, but also determine the predictive ability of the data. Real data often contains a large amount of data noise and data redundancy [14]. These artificially generated abnormal data are very detrimental to the results, so these data need to be cleaned. The preprocessing work for the data is Fig.1.

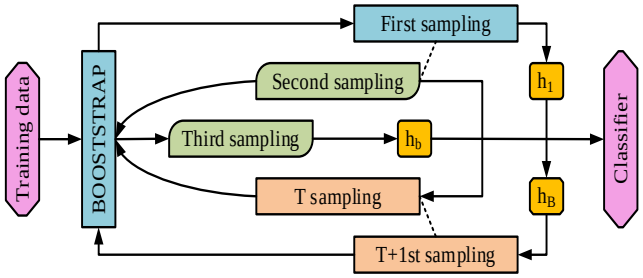


Fig. 1. Diagram of data preprocessing process.

From Fig. 1, the cleaning of training data is first run in the boot program, improving missing values and noise during each sampling process; Then oversampling and under sampling the data, and then transforming the data into a whole smooth curve to get the final output results. The variables selected through Fig. 1 can be used for stepwise regression, and the complex collinear data that meets the requirements is called biased estimation. The data that conforms to the characteristics of continuity can be obtained through a penalty function, as Formula (1).

$$p_\lambda\left(B_j\right)=\begin{cases}\lambda\left(B_j\right) & \left|B_j\right|\le\lambda \\ -\dfrac{\left(\left(B_j\right)^2-2a\lambda\left(B_j\right)+\lambda^2\right)}{2(a-1)} & \lambda\le\left|B_j\right|\le a\lambda \\ \dfrac{(a+1)\lambda^2}{2} & a\lambda\le\left|B_j\right|\end{cases} \tag{1}$$

In Formula (1), a is a constant with a value of, and the adjusted parameter is denoted as λ, with a range of non negative values. B_j is called effective data set, which is often based on natural logarithm [15]. At this point, the unit of random entropy is recorded as bits and is only related to B_j. As the value of B_j increases, the range of random entropy variables becomes larger. The relationship between them is Formula (2).

$$H\left(B_j\right)=-\sum_{i=1}^{n}p_i\log p_i \tag{2}$$

In Formula (2) above, p_i is a random variable with a value range over $[0,+\infty]$. The improvement of XGBoost algorithm relative to decision trees lies in gradient correlation, which is a running model that enhances decision-making ability. The range of regularization terms determines the complexity of the algorithm, and the objective function determines its optimal solution, as shown in formula (3).

$$\underline{y}=\sum_{t=1}^{T}f_t\left(x_i\right),\ f_t\in F \tag{3}$$

In Formula (3), the number of XGBoost decision trees is denoted as T, the decision forest is represented as F, and the specific XGBoost decision tree is denoted as f_t. If the previous prediction for round t is denoted as $\underline{y}$, then the objective function can be expanded using second-order Taylor expansion. In order to measure the quality of the decision tree, the scoring function of XGBoost is introduced, as listed in Formula (4).

$$Mark=0.5\left[\frac{G_L^2}{\alpha_L+\lambda}+\frac{G_R^2}{\alpha_R+\lambda}-\frac{\left(G_L+G_R\right)}{\alpha_L+\alpha_R+\lambda}\right]-\gamma \tag{4}$$

In Formula (4), G_L, G_R represent the segmentation points on the left and right sides of the mean, respectively; The mean values of multiple segmentation points are denoted as α_L, α_R; The difference in the parameters of the decision tree is called γ. The higher the value of $Mark$, the better the quality of the XGBoost decision model. The decision tree is continuously segmented according to this method, and the existing nodes in the time series can be predicted according to the past segmentation points and linear white noise, as expressed in Formula (5).

$$Y_t=\varphi_1 y_{t-1}+\varphi_2 y_{t-2}+\ldots+\varphi_p y_{t-p}+u_t \tag{5}$$

In Formula (5), the predicted value of the existing nodes is Y_t, and the autoregressive coefficients between the segmentation points are represented by φ_p, with a range of positive integers of $(1,p)$. Linear white noise is recorded as u_t, and white noise sequence in regional time can be calculated by Formula (6) [16].

$$W_f\left(\phi,\chi\right)=\chi^{-0.5}\int_{-\infty}^{+\infty}\delta(t)\eta^*\left(\frac{t-\phi}{\chi}\right)dt=\left\langle\delta,\kappa_{(a,b)}(t)\right\rangle \tag{6}$$

In Formula (6), the shift factor and displacement factor are denoted as ϕ,χ, and their value ranges are non negative.

$\delta(t)$ represents a square integrable signal, complex conjugation is denoted as $*$, and $\kappa_{(a,b)}(t)$ is used to represent the cluster function [17]. When XGBoost calculates the importance of features, it studies the method of selecting and calculating the obtained values. By traversing all intermediate nodes, it maps the number of feature times of the branch. The trained tree model in XGBoost regression tree is exhibited in Fig. 2.

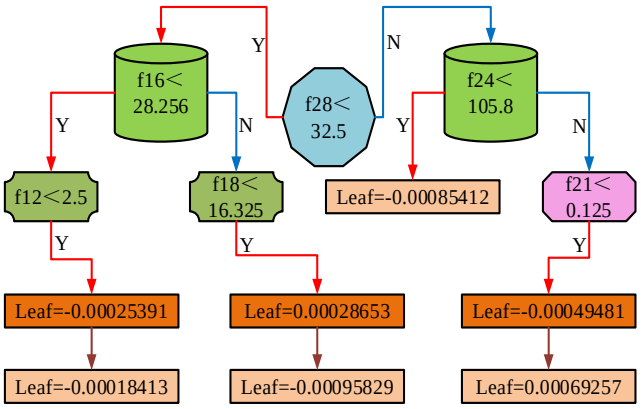


Fig. 2. XGBoost regression tree model trained in.

There are seven leaf nodes in Fig. 2, represented by rectangles, with the remaining shapes serving as attribute partitioning conditions. From Fig. 2, the complete XGBoost regression tree model has a total of four layers, including 13 nodes. The importance of customer data in the features can be intuitively seen in the figure, which provides guidance for the selection of later features. Dimension explosion is prone to occur during encoding extraction, leading to an increase in the working time and isolated points of the model. To avoid this issue, the study used variance encoding to change qualitative features by labeling samples, and the formula for calculating sample labels is Formula (7).

$$P\left(y=Y\middle|k=K\right)=\frac{N_Y^K}{N^K} \tag{7}$$

In Formula (7), $P\left(y=Y\middle|k=K\right)$ is called the posterior probability of sample labels, y is the selected number of target classes, the total number is recorded as Y, k is the value of qualitative characteristics, and the number of samples under this condition is recorded as E. When the feature value is K, the sample size is converted to N^K. In order to balance the mean effect between prior probability and posterior probability, conditional parameters with multiple repetitions and variability are introduced. The calculation method is Formula (8).

$$Q\left(y=Y\middle|k=K\right)=\phi*\bar{O}\left(y=Y\right)+\left(1-\phi\right)*P\left(y=Y\middle|k=K\right) \tag{8}$$

In Formula (8), the probability of mean encoding is denoted as $Q\left(y=Y\middle|k=K\right)$, and $\bar{O}\left(y=Y\right)$ can control the slope of the conditional parameter, that is, as the value of $\bar{O}\left(y=Y\right)$ increases, the rate of change of ϕ with y,k becomes slower.

B. Fusion Algorithm based on Differential Automatic Regression Moving Average Variables

DAMAV is suitable for predicting time series and is widely used in linear regression, as expressed in Formula (9).

$$h_t=\mu_0+\mu_1 h_{t-1}-\nu_t-\nu_1\varpi_{t-1} \tag{9}$$

In Formula (9), the order of autoregression and moving average is recorded as h_t,μ_0, ν_t represents the error in stochastic process, and the number of differences made in regional time is expressed as ϖ_{t-1} [18]. On a practical basis, importing the formed dataset into the constructed model can complete autonomous training and calculate the results. The newly created dataset can also keep the model fresh, update its automation capabilities, and effectively reduce human resources, as demonstrated in Fig. 3.

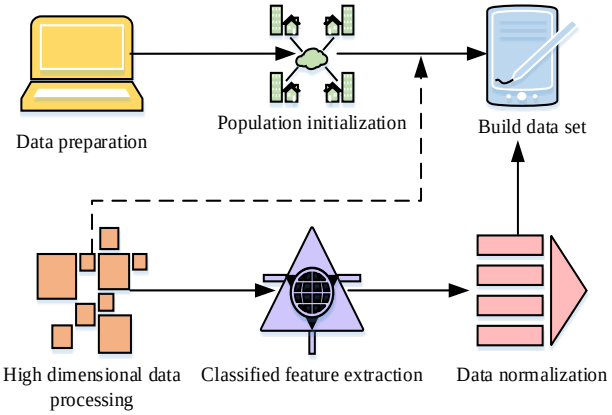


Fig. 3. XGBoost regression tree autonomous learning flow chart.

Fig. 3 shows the self-learning process of XGBoost regression tree. Firstly, it is necessary to collect a complete and comprehensive dataset, which can be formed on the network or experimentally measured; Then, set the data format for the algorithm, which is the target variable or feature value; Next, the outliers in the data set are screened, such as outlier and noise; Finally, the formatted data is run in the algorithm to extract the corrected values of the parameters [19]. Among them, to determine the indicators of predictive performance, this study selected Root Mean Square Error (RMSE) to consider it, as displayed in Formula (10).

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}\left(\theta^i-\overline{\overline{\theta}}^{=i}\right)^2}{n}} \tag{10}$$

In Formula (10), the true value is denoted as θ, and the predicted value is represented by $\overline{\overline{\theta}}$. In employee turnover in enterprises, temperature values are used instead of turnover values. The lower the temperature, the more severe the employee turnover in the enterprise is. In the iterative process of the algorithm, the training results will make the weight larger, so that the model can be steadily improved on the basis of the mean value, which can be described by the Newton's

Cooling Law (CL). The formula for CL is Eq. (11).

$$T'(t) = -\vartheta\left(T(t) - H\right) \tag{11}$$

In Formula (11) above, the temperature of the object is denoted as T, t represents the operating time, and the cooling rate is the derivative of T changing with t, called $T'(t)$. Room temperature is represented by H, where ϑ is a constant with a value range of $[0.1,1.0]$ to control the cooling rate. When the value of ϑ is fixed, the larger the temperature difference in the environment, the faster the cooling rate. In the employee turnover model, when compared to the average salary of all enterprises, the lower the salary, the faster the employee turnover rate. To describe the degree to which employees attach importance to the company, sample weights were introduced to simulate the degree to which employees pay attention to nearby companies, and Formula (12) was established.

$$w_k = w_0 e^{-\sigma\varsigma_k} \tag{12}$$

In Formula (12), the popularity of company k in the training set is denoted as w_k. The average treatment of all companies is represented by w_0. The heat loss coefficient of the model is denoted as σ. The company's decay rate over time is represented by ς_k. The XGBoost model is a foundational learner suitable for various types, which can transform linear problems into regression problems and is suitable for most work environments. XGBoost's loss function is composed of regularization terms, and its expansion is Formula (13).

$$\Omega = \omega\xi + 0.5\psi\sum_{j=1}^{\psi}\zeta_j^2 \tag{13}$$

In Formula (13), ξ represents the richness of the decision tree, and ω is the weight of the decision tree. The work done by the leaves in unit time is recorded as ζ, and the output regularization weight is expressed as ψ. When the XGBoost model predicts employee turnover, missing values in features can be ignored and assigned to leaves, thereby improving the overall training speed. In the working process of the decision tree, out of pocket errors are obtained from out of pocket data. The performance used to calculate the decision tree is Formula (14).

$$\text{Im } por\tan ce\left(FeatureX\right) = \left(\sum_{i=1}^{N} errOOB_2 - errOOB_1\right)/N \tag{14}$$

In Formula (14), the importance of the feature is denoted as $\text{Im } por\tan ce\left(FeatureX\right)$, and the bag contains a total of N data, where $errOOB$ represents the out of bag error value [20]. To train all samples in the dataset, a random variable is randomly extracted from the type features, and then converted into a numerical value based on the label of the previous sample. The weight coefficient of the priority is adjusted as Formula (15).

$$E_k^i = \frac{\left(\sum_{j=1}^{p-1}\left(E_{\Delta j,k} = E_{op,k}\right)Y_{\sigma j}\right) + ap}{\left(\sum_{j=1}^{p-1}E_{\Delta j,k} = E_{op,k}\right) + a} \tag{15}$$

In Formula (15), $E_{\Delta j,k}$ represents the training set in the input algorithm model, and the differentiable loss variable is denoted as $E_{op,k}$. a, p is a constant, and its value size reflects the priority of the sample. This study is based on the improved XGBoost regression tree algorithm model of DAMAV (DV-XGBoost), as listed in Fig. 4.

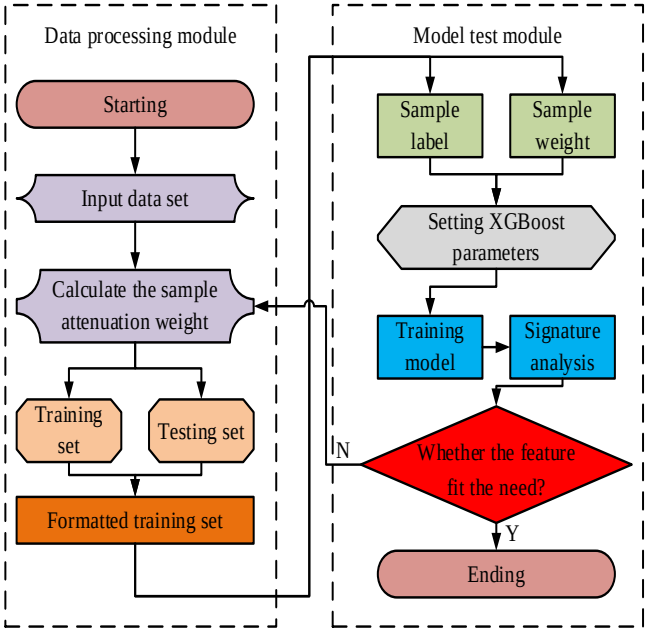


Fig. 4. Flow chart of improved DV-XGBoost regression tree algorithm model.

The DV-XGBoost model in Fig. 4 can be divided into two modules, totaling four parts. Firstly, to calculate the attenuation weight of the input dataset and divide it into a test set and a training set; Then, format the training set in the data preparation module and input it into the model test set; Next is to iterate the labels and weights of the samples to analyze the parameters of the DV-XGBoost model; Finally, the feature is judged. If it meets the requirements, the final value is output. Otherwise, the data preparation process is returned and the attenuation weight is recalculated.

IV. Experimental Study on Enterprise Character Loss Information Based on XGBoost

To verify the effectiveness of the DV-XGBoost algorithm in practical applications, iteration and accuracy verification were conducted. Finally, the DV-XGBoost model was applied to the Company dataset for simulation experiments.

A. DV-XGBoost System Development Environment and Model Parameter Determination

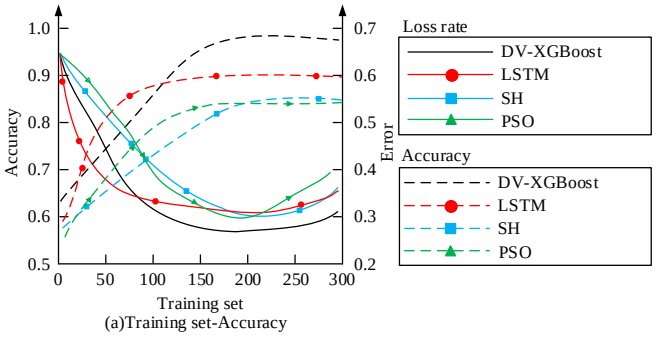
This study selected a self-collected Company dataset, including four types of companies: fine organic, electronic

engineering, aerospace, and automation, with a total of 6523 enterprises. Considering the limited types of data, the dataset will be divided into a training set and a testing set in a ratio of 2:3. The specific equipment and software used in the experiment are Table I. In the experiment of DV-XGBoost, the research first sets the necessary experimental conditions. There are some key parameters to be set in DV-XGBoost algorithm, including iteration times and acceleration factor because the problem of employee turnover in enterprises is difficult to predict and the computing resources are limited. Therefore, the optimization objective function is studied to solve the optimal solution of DV-XGBoost. The objective function has a clear optimization goal, and there are constraints on the range of the number of brain drain for each iteration of DV-XGBoost algorithm, the position, velocity and fitness data of each particle are collected and used to calculate the convergence performance of DV-XGBoost algorithm.

Experimental Parameters

Data set	Development language	Number of cores	Internal storage
Company	Python 8.5	8	1024 G
Operating system	Display card	Database	Processor
127Ubnutu 22.01.21	37.0 GHz	Mysqppl 5.30.2023	Intel Core i9
Web development framework	Language	Operator	Model
Django2.22.3	Easy Chinese	Electric, orangic...	F2.9LII-US M

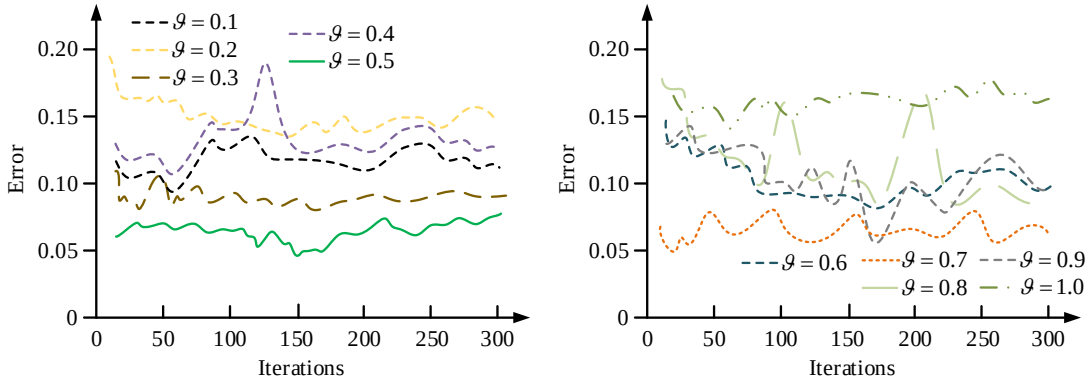
The collected dataset needs further processing to enable the studied algorithm to learn. The dataset was processed using DV-XGBoost for iterative optimization. To verify its accuracy, traditional Spotted Hyena Algorithm (SH), Long Short Term Neural Network (LSTM), and Particle Swarm Optimization Algorithm (PSO) will be compared with it. The accuracy and error rate results in the training set are Fig. 5.



(a)Training set-Accuracy

Fig. 5. Comparison of accuracy-training set image and error rate-training set image.

From Fig. 5, before 100 training sessions, the accuracy of DV-XGBoost algorithm is slightly lower than that of SH and PSO, and the error rate is higher. However, when the iterations reaches 100 or more, the accuracy of DV-XGBoost is higher than both algorithms, and tends to stabilize at 190 iterations, which is higher than the other three algorithms. Although increasing the iterations may reduce the operational efficiency of the model, after comprehensive consideration, the accuracy weight of the model is higher. Therefore, the DV-XGBoost algorithm proposed in the study has better performance. After the learning of the DV-XGBoost algorithm is completed, it is also necessary to consider the parameter determination during testing, as demonstrated in Fig. 6.



(a) The rate control constant of the step cooling curve is 0.1-0.5

(b) The rate control constant of the step cooling curve is 0.6-1.0

Fig. 6. Error-training times image of regularization term.

The parameter of this study is the rate control constant $\vartheta \in [0.1,1.0]$ of the cooling curve between the enterprise and employees. From Fig. 6, when the rate control constant of the step cooling curve is 0.5 and the number of iterations is 150, the error rate is the lowest, which is 0.042. Therefore, the final number of iterations was determined to be 150 and the rate control constant was taken as 0.5.

B. Experimental Verification of Employee Turnover Prediction in Enterprises based on DV-XGBoost

To verify the accuracy of the DV-XGBoost model in predicting employee turnover in enterprises, simulation experiments were conducted. It evaluates the practicality of the DV-XGBoost algorithm by observing whether an employee has resigned. First, the DV-XGBoost algorithm is initialized, and then the employee enterprise information flow is entered in the data preparation module. Finally, the rate control constant of the step cooling curve is set to 0.5, and the

employee dynamics and resigned employee information within 60 days are collected. The image drawn after calculating the error is Fig. 7.

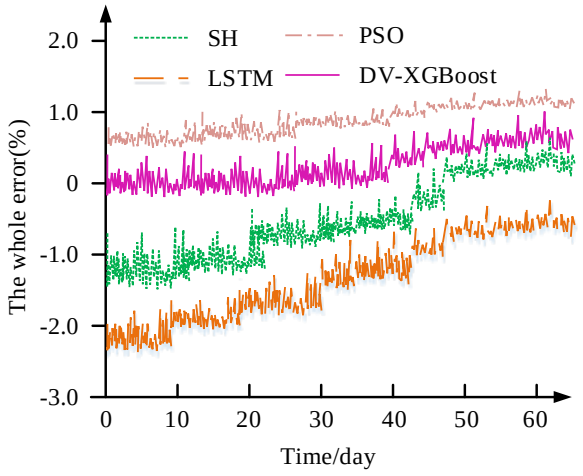


Fig. 7. Total error-time image of four algorithms.

Fig. 7 shows the error comparison of four algorithms in the experiment. In Fig. 7, after 38 days, the error of DV-XGBoost in determining employee turnover has approached zero, while the other three algorithms have a wide range of error fluctuations. Especially for the PSO algorithm, on the third day, the highest error value of the four algorithms was -2.94%. The total error range of DV-XGBoost, SH, LSTM, and PSO is significantly different, making it easy to compare algorithm performance. However, relying solely on the analysis of total error is not objective enough, so the study analyzed the four model analysis errors caused by individuals or enterprises, as shown in Fig. 8.

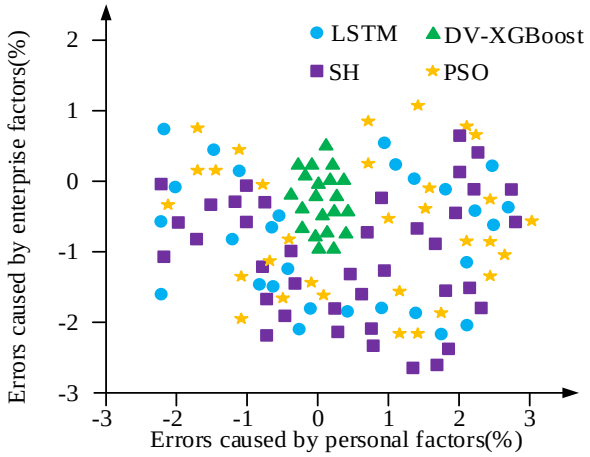


Fig. 8. The types and genres errors of the four algorithms.

From Fig. 8, it can be clearly seen that the experimental results of the DV-XGBoost model are concentrated in the range of total error of 0. The error range caused by personal factors is [-0.6%, 0.8%], while the error range caused by corporate factors is [-1.0%, 1.1%]. The error distribution of the remaining three algorithms is wide, and the distribution of larger errors is sparse. To more intuitively distinguish the ability of the four algorithms to correct errors, 400 experimental data records were conducted and the images displayed in Fig. 9 were plotted.

From Fig. 9, in 400 error testing experiments, LSTM has the largest range of error variation, with the highest frequency of errors recorded as [-0.75%, 0.62%]; Next is the SH algorithm, which is between [-0.18%, 0.23%]. The error variation range of PSO is close to DV-XGBoost, with values above [-0.12%, -0.04%]; the error curve of DV-XGBoost fluctuates between -0.03% and 0.02%, with the smallest fluctuation range. Excluding the LSTM and SH algorithms with the highest error ranking, only comparing the experimental results of DV-XGBoost algorithm and PSO for correct prediction, and drawing the error matrix. The resulting image is Fig. 10.

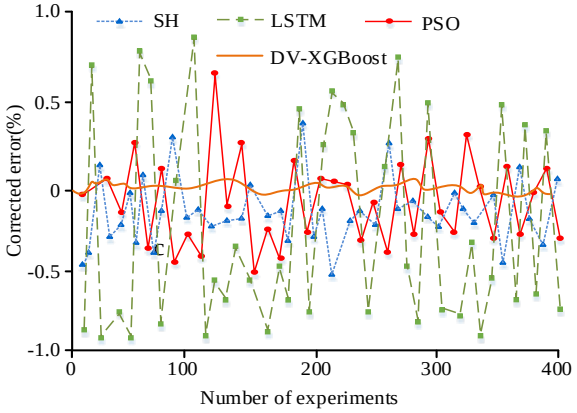


Fig. 9. Error changes of four algorithms in four hundred calibration experiments.

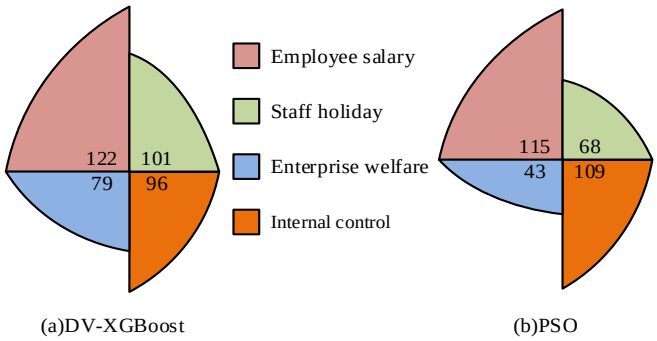


Fig. 10. Error matrix of DV-XGBoost algorithm and PSO algorithm.

Fig. 10 predicts four types of employee turnover in companies based on four conditions: employee salary, employee leave, and enterprise characteristics, as well as internal control, based on employee differences. The experimental results accurately predicted by DV-XGBoost and PSO are presented. The prediction accuracy of DV-XGBoost reached 398 times, with an accuracy rate of 99.5%, and the accuracy rate of PSO was 83.2%. To observe the experimental results of DV-XGBoost and PSO more intuitively, a linear fitting graph based on matrix drawing of two algorithms and Golden Sine algorithm (GS) was studied. The predicted values of the two were compared with the true values, as expressed in Fig. 11.

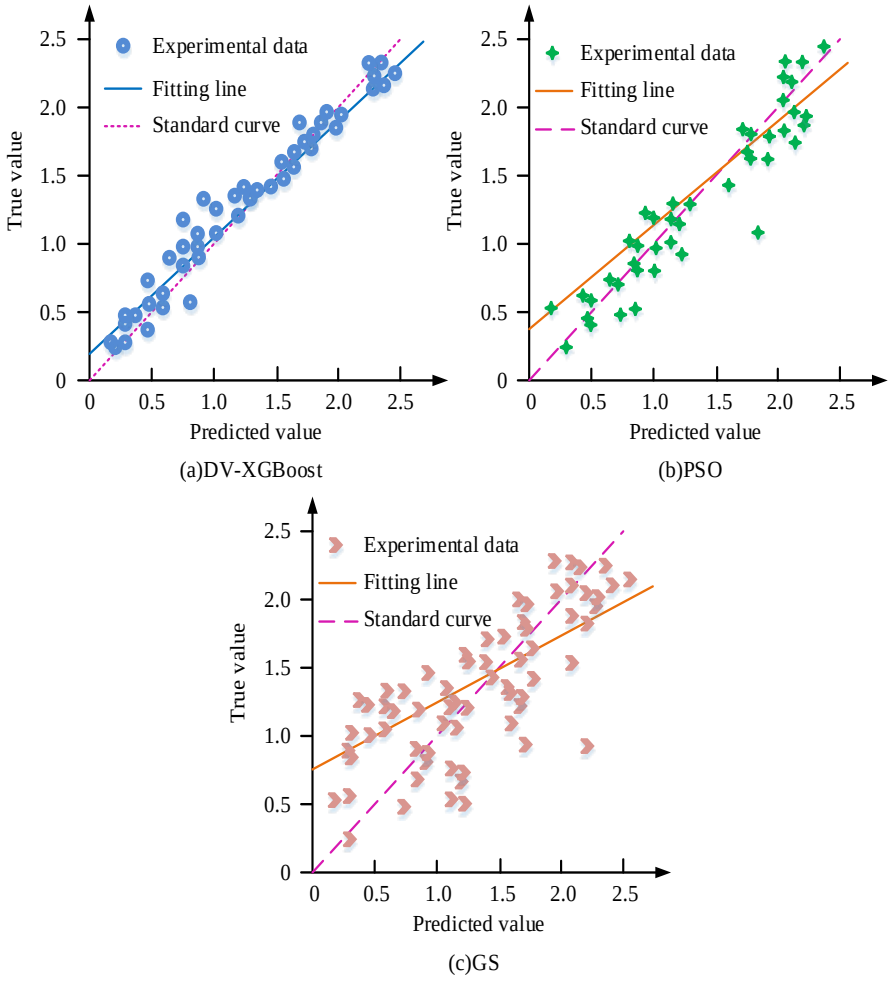


Fig. 11. Linear fitting diagram of DV-XGBoost PSO and GS.

Fig. 11 shows the comparison of three algorithms on predicted and true values. From Fig. 11, the linear fit (R^2) of the DV-XGBoost algorithm is 0.9914, the R^2 of PSO is 0.9547, and the R^2 of GS is 0.8825, indicating that there is no underfitting in the model. In summary, it can be concluded that the DV-XGBoost algorithm model can accurately predict the situation of employee turnover in enterprises, provide internal control warnings, or notify the human resources department (HR) of the company to release recruitment information.

V. Results And Discussion

With the increase of employee turnover rate in enterprises, the utilization rate of human resources is in a downward trend. In order to predict this kind of problem, and then make relative measures as early as possible, this study uses the improved limit gradient lifting algorithm to draw a step cooling curve for the collected employee's psychological data according to the cooling law, so as to predict their turnover possibility. In order to reduce the chaotic data during the running of the algorithm, the experiment was carried out in the Company data set. In order to analyze the proposed DV-XGBoost algorithm extensively, the experimental results are compared with PSO, SH and GS, and finally the number of iterations is 150, and the rate constant is 0.5. In the error analysis experiment, the error curve of DV-XGBoost fluctuates between-0.02% and 0.04%, with the smallest fluctuation range. The prediction accuracy of DV-XGBoost is 99.5%, and that of PSO is 83.2%. The linear fitting of DV-XGBoost is 0.9914, the linear fitting is excellent, and the linear fitting of PSO is 0.9547. The experimental results show that the DV-XGBoost model proposed in this study has strong robustness in predicting employees' psychology and is suitable for improving the utilization rate of human resources for the company. However, the algorithm is only applicable to companies, and the research on the employment factors of school employees is still insufficient, which will be gradually improved in future research.

VI. Conclusion

With the growth of internet technology, predicting the turnover psychology of enterprise employees is becoming increasingly important, such as increasing work efficiency for the administrative department of the company and providing early warning for the talent gap period of the enterprise. This study is based on XGBoost and DAMAV to generate the fusion algorithm DV-XGBoost. The experiment took into account both personal and corporate factors during resignation,

and conducted simulation experiments on the Company dataset, comparing with three algorithms such as LSTM. 40% of the Company dataset was extracted and trained on the DV-XGBoost model. Through controlling the rate constant of the cooling curve, the final iteration number was determined to be 150 times, with a rate constant value of 0.5. In the error analysis experiment, a total of 400 experiments were conducted, and the errors of the SH and LSTM algorithms fluctuated within the range of [-2.3%,1.4%] and [-2.2%,1.9%], respectively. The error curve of DV-XGBoost fluctuates between -0.02% and 0.04%, with the smallest fluctuation range; The variation range of PSO is close to DV-XGBoost, between [-0.13%, -0.03%]. Draw an error matrix for the experimental results of DV-XGBoost algorithm and PSO. In 400 experiments, the prediction accuracy of DV-XGBoost is 99.5%, and the accuracy of PSO is 83.2%. This study drew linear fitting graphs for two algorithms based on matrices. The R^2 of DV-XGBoost was 0.9914, indicating excellent linear fitting, while the R^2 of PSO was 0.9547. In summary, the DV-XGBoost model can accurately predict employee turnover in enterprises, improve the efficiency of human resources departments, and enable enterprises to cope with talent shortages. However, the DV-XGBoost model is only suitable for analyzing companies with employees from the same location. For comprehensive companies from multiple regions, it is difficult to analyze the emotional changes caused by local customs among employees, and the model will label them as noise. This is because human psychology is complex and diverse, and the dataset for research and analysis contains fewer types. With the increase of volunteers, it is believed that future research can be improved.

ACKNOWLEDGMENT

The research is supported by Shaanxi Federation of Social Sciences, Research on the Characteristic Development and Innovation Path of County Economy under the Background of Transcendence in Shaanxi New Era, (No.2021ND0035); Rural Revitalization Bureau of Xingping City, Shaanxi Province, Rural Revitalization Plan of Xingping City, (No. GH [2021]00026); Development and Reform Bureau of Xingping City, The High-tech Industrial Park's Industrial Development Plan of Xingping City, (No. HX2022001).

REFERENCES

[1] Guo Y, Mustafaoglu Z, & Koundal D. Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms. Journal of Computational and Cognitive Engineering, 2022, 2(1), 5–9.

[2] Chen J, Zhao F, Sun Y, Y Yin. Improved XGBoost model based on genetic algorithm. International Journal of Computer Applications in Technology, 2020, 62(3): 240-245.

[3] Calanna P, Lauriola M, Saggino A, M Tommasi, S Furlan. Using a supervised machine learning algorithm for detecting faking good in a personality self-report. International Journal of Selection and Assessment, 2020, 28(2): 176-185.

[4] Zhao W P, Li J, Zhao J, D Zhao, J Lu, X Wang. Xgb model: research on evaporation duct height prediction based on xgboost algorithm. Radioengineering, 2020, 29(1): 81-93.

[5] Lu Y, Fu X, Guo E, Tang. XGBoost algorithm-based monitoring model for urban driving stress: Combining driving behaviour, driving environment, and route familiarity. IEEE Access, 2021, 9: 21921-21938.

[6] Deng X, Li M, Deng S, L Wang. Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification. Medical & Biological Engineering & Computing, 2022, 60(3): 663-681.

[7] Li Z X, Shi X L, Cao J D, XD Wang,W Huang. CPSO-XGBoost segmented regression model for asphalt pavement deflection basin area prediction. Science China Technological Sciences, 2022, 65(7): 1470-1481.

[8] Tao T, Liu Y, Qiao Y, LG B, JL A, CZ A, WA Yu. Wind turbine blade icing diagnosis using hybrid features and Stacked-XGBoost algorithm. Renewable Energy, 2021, 180(Dec.): 1004-1013.

[9] Li S, Zhang X. Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm. Neural Computing and Applications, 2020, 32: 1971-1979.

[10] Gu X, Han Y, Yu J. A novel lane-changing decision model for autonomous vehicles based on deep autoencoder network and XGBoost. IEEE Access, 2020, 8(99): 9846-9863.

[11] Li C, Zhang Y, Cui C, D Fan, Y Zhao, XB Wu, B He, Y Xu, S Li, J Han. Identification of BASS DR3 sources as stars, galaxies, and quasars by XGBoost. Monthly Notices of the Royal Astronomical Society, 2021, 506(2): 1651-1664.

[12] Osman A I A, Ahmed A N, Chow M F, YF Huang, A El-Shafie. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal, 2021, 12(2): 1545-1556.

[13] Song P, Liu Y. An XGBoost algorithm for predicting purchasing behaviour on E-commerce platforms. Tehnički vjesnik, 2020, 27(5): 1467-1471.

[14] Gao S, Li S. Bloody Mahjong playing strategy based on the integration of deep learning and XGBoost. CAAI Transactions on Intelligence Technology, 2022, 7(1): 95-106.

[15] Ünver M, Olgun M, Türkarslan E. Cosine and cotangent similarity measures based on Choquet integral for Spherical fuzzy sets and applications to pattern recognition. Journal of Computational and Cognitive Engineering, 2022, 1(1): 21-31.

[16] Wang H, Yue W, Wen S, X Xu, HD Haasis, M Su, P Liu, S Zhang, P Du. An improved bearing fault detection strategy based on artificial bee colony algorithm. CAAI Transactions on Intelligence Technology, 2022, 7(4): 570-581.

[17] Shahbazi Z, Byun Y. Product Recommendation Based on Content-based Filtering Using XGBoost Classifier. International Journal of Advanced Science and Technology, 2020, 29(4):6979-6988.

[18] Osman A I A, Ahmed A N, Chow M F, YF Huang, A El-Shafie. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal, 2021, 12(2): 1545-1556.

[19] Oslund S, Washington C, So A, Chen, T, &Ji, H. Multiview Robust Adversarial Stickers for Arbitrary Objects in the Physical World. Journal of Computational and Cognitive Engineering, 2022, 1(4): 152-158.

[20] Li C, Zhang Y, Cui C, D Fan, Y Zhao, XB Wu, B He, Y Xu, S Li, J Han. Identification of BASS DR3 sources as stars, galaxies, and quasars by XGBoost. Monthly Notices of the Royal Astronomical Society, 2021, 506(2): 1651-1664.

Intelligent Design of Ethnic Patterns in Clothing using Improved DCGAN for Real-Time Style Transfer

Yingjun Liu, Ming Wu*
School of Art, Guangxi Minzu Normal University, Chongzuo 532200, China

***Abstract*—In view of the problems that traditional real-time style transmission technology requires a large number of sample map training, low image quality, lack of realism and detail, this study combines the improved generative adversarial network (GANs) with real-time style transfer technology, and enhances the real-time style transfer calculation with adaptive instance normalization. As a result, a novel intelligent clothing ethnic pattern design model is developed. Experimental results show that the model reduces physical memory usage by 45.7%, with only 453MB, and utilizes only 26% of CPU resources in terms of CPU usage. The training time is approximately 20 minutes and 48 seconds. This model performance is obviously higher than other models. The designed intelligent clothing ethnic pattern design model in this study demonstrates higher clarity and shorter processing time, and has potential applications in the field of image generation.**

Keywords*—Computer vision; improved DCGAN; style transfer; adaptive instance normalization; intelligent design of patterns

I. Introduction

In today's era, clothing design plays a crucial role in personalization and innovation. With the growing demand for unique styles and cultural diversity, ethnic patterns have become a popular design element. However, manually drawing and applying these patterns require a significant amount of time and labor, limiting their application in large-scale production. To address this issue, computer vision and deep learning technologies have been widely applied in the field of fashion design [1]. In particular, style transfer methods based on Generative Adversarial Networks (GANs) have made significant progress. Among them, Deep Convolutional GAN (DCGAN) is a powerful generative model that can learn from input data and generate realistic images [2-3]. Style transfer is a technique of applying the style of one image to another image, which is widely used in artistic creation and design. AdaIN is an instance normalization technique that is adaptive to adjust the style of an image, making the result of style transfer more natural and realistic.However, there are some defects in the traditional intelligent design technology. On the one hand, traditional techniques often require a large number of sample images for training, which can be very difficult and time-consuming for specific ethnic patterns. On the other hand, traditional techniques may lead to a low quality of the generated images and a lack of realism and detail. Moreover, traditional techniques may not very well preserve the style and character of the original pattern. For the above problems, this study proposed a series of improvement measures for standard DCGAN, introduced in the decoder network, the generated false image more close to the real image, adopted the method of multi-scale feature extraction and fusion in the generator and discriminator network, makes the generated false image more real, by stacking the convolution layer and deconvolution layer to improve the quality of the generated image, and an improved algorithm of deep convolution generation against network (IDCGAN) is developed. In addition, the adaptive instance normalization (AdaIN) is also applied to the real-time style transfer technology to design an intelligent model for the clothing ethnic pattern design. The study aims to achieve image generation through adversarial learning, allowing DCGAN to learn specific ethnic patterns from a small number of samples and generate high-quality, realistic images. The article consists of four main parts. The second part provides a comprehensive review of the current research status of intelligent clothing design and style transfer systems. The third part establishes an intelligent model for ethnic pattern design in clothing based on improved DCGAN for real-time style transfer. The fourth part includes comparative experiments and efficiency verification to evaluate the optimization effects of the model.

II. Related Works

As people's pursuit of personalization and unique styles continues to grow, traditional clothing design no longer meets the demands of consumers. In order to meet this demand, researchers have been exploring intelligent design technologies that are adaptable to these needs. Ding et al. addressed the issues of low accuracy and stability in traditional manual crown design by developing an automatic crown design strategy with DCGAN, and the outcomes indicates that it had the smallest morphological differences compared to natural teeth [4]. Abd Al et al. proposed a DCGAN based algorithm and a novel Capsule Network to assist semiconductor manufacturers in identifying defect patterns in wafers, and the experimental results showed that the method achieved a training accuracy of 99.59% and a validation accuracy of 97.53% [5]. Bian et al. developed a compound screening model based on DCGAN to screen and design novel compounds with target-specificity for cannabinoid receptors, and the experimental results showed that the model had the highest accuracy compared to other models [6]. Cheng et al. proposed a Data Enhancement Communication Behavior Recognition (DECBR) scheme to

address the limitations of traditional communication behavior recognition techniques in accurately analyzing communication behaviors, and the DECBR scheme significantly improved the accuracy and efficiency of behavior recognition under small sample conditions [7]. Li et al. designed an image classification model that combines DCGAN and AlexNet for rapid differentiation of multiple forms of glioblastoma images, and the experiment outcomes indicates that the model reached 0.920 accuracy and an AUC of 0.947 for distinguishing PsP and TTP after 10-fold cross-validation [8]. Ni et al. addressed the issue of large deviations in carrot quality identification using traditional visual inspection methods by designing a carrot quality identification model that combines DCGAN and Squeeze-and-Excitation Deep Networks, and the experiment outcomes showed that the model reached 98.36% accuracy [9].

Jing et al. proposed a new normalization module called Dynamic Instance Normalization (DIN) to address the deployment challenges of style transfer systems in resource-constrained environments. DIN allows for flexible and more efficient transfer of arbitrary styles. The experimental results showed that this approach reduced computational costs by more than 20% compared to existing methods [10]. Reimann et al. addressed the issue of one-shot stylization in existing style transfer computations, which mostly limit the style elements interactive adjusting. They designed a fast style transfer network which is stroke-adjustable. It can simultaneously control stroke intensity and size. The experimental results showed that the model make users achieving resolutions exceeding 20 million pixels and good output fidelity [11]. Hollandi et al. developed a deep learning-based cell nucleus segmentation framework that utilizes image style transfer to automatically generate cell nucleus segmentation masks. This framework aims to find a method for locating 2D cell nuclei in different regions. The experimental results showed that the model effectively identifies cell nuclei in different experiments without the need for expert annotations [12]. Huang et al. addressed the lack of diversity in traditional style transfer by designing a style transfer model that combines region semantics with multi-style transfer. The experimental results showed that the model seamlessly combines multiple styles together, and, with the assist of semantic matching, assigns corresponding styles to content regions [13]. Xu et al. tackled the problem of indistinguishable details between different types of objects caused by single-band imaging. They designed a target detection-oriented style transfer network for panchromatic remote sensing images. After style transfer, the target detection accuracy on panchromatic remote sensing images significantly improved [14]. Zhou et al. proposed a new approach that combines attention mechanism with style transfer models to enhance the flexibility of style transfer tasks. The experimental results showed that this approach is effective and produces high-quality images [15].

In summary, DECBR and style transfer have a solid theoretical and implementation foundation in the field of intelligent clothing design. However, there is limited research that combines the two for ethnic clothing design. Therefore, the study aims to improve DECBR and combine it with style transfer computations to develop an intelligent clothing design model, in order to further advance the clothing design industry.

III. Improvement of DCGAN Algorithm and Establishment of Intelligent Pattern Design Model

This chapter contains two sections. The first gives an introduction on the standard Deep Convolutional Generative Adversarial Network (DCGAN) and proposes some improvement strategies to address its limitations. The second section focuses on improving traditional real-time style transfer networks and establishing an intelligent design model based on real-time style transfer networks.

A. Improvement of DCGAN Algorithm Design

DCGAN is a neural network model used for generating realistic images. It combines the ideas of generative models and adversarial training, mainly containing two components: the generator and the discriminator [16-17]. The former is a network that uses random noise vectors to do input and attempts to bring up images similar to the training data. The generator gradually constructs the image through multiple convolution, deconvolution, and activation function operations. The discriminator is also a convolutional neural network whose goal is distinguishing between the images generated and real ones. The discriminator extracts image features through operations such as convolution, pooling, and activation functions, and outputs a probability value between 0 and 1, indicating the likelihood that the input image is a real image [18]. In DCGAN, the two components are alternately trained. The generator manufacture images that is real enough for deceiving the discriminator, while the later strives to differentiate between the images generated by the generator and real images [19]. This process stimulates the generator to continuously enhance the generated images' quality and makes the discriminator more accurate, as shown in Fig. 1.

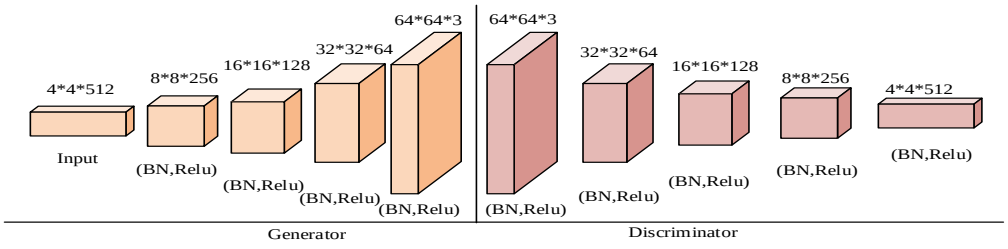


Fig. 1. Schematic diagram of DCGAN structure.

Although DCGAN is a widely used entity in the studying areas of image synthesis, editing, and super-resolution reconstruction, there are some limitations when applying DCGAN to real-time style transfer of ethnic clothing patterns' intelligence design. Firstly, DCGAN has relatively weak feature extraction capabilities. Despite using convolutional neural networks (CNN) to learn image features, it may not fully capture the complex features of ethnic patterns due to network structure and training data limitations. This can result in generated clothing that is not realistic enough and deviates significantly from the target style. Secondly, there is a lack of overall style transfer constraints. DCGAN primarily focuses on generating realistic images during the training process, with less emphasis on maintaining consistency and layout of local patterns. In clothing design, maintaining pattern consistency is crucial, but DCGAN may not fully consider the layout and details of patterns in different parts of the garment, resulting in clothing that does not resemble a normal garment and lacks coherence and integrity. To address these limitations, a modified DCGAN approach is proposed, and the network structure during the training phase is shown in Fig. 2.

Fig. 2 illustrates the network structure during the training phase of IDCGAN. At the beginning of training, random noise is input into the generator. The generator processes the noise through decoding and encoding operations to generate fake images. These fake images gradually approach real images through the generator. At the same time, real and fake images are simultaneously input into the discriminator. It is another network responsible for classifying the input images and outputting the feature space Z. This feature space represents the representation of the images in the discriminator. Next, the classification loss and the real/fake loss are calculated by comparing the classification results of real and fake images. The classification loss measures the accuracy of the discriminator in distinguishing real and fake images, while the real/fake loss reflects the adversarial training process within the two components. Throughout the training process, the two components engage in a competitive dynamic, continuously optimizing their parameters. The generator's objective is to produce increasingly realistic fake images to deceive the discriminator, while the discriminator aims to distinguish between real and fake images, improving its accuracy. The algorithm flow is shown in Fig. 3.

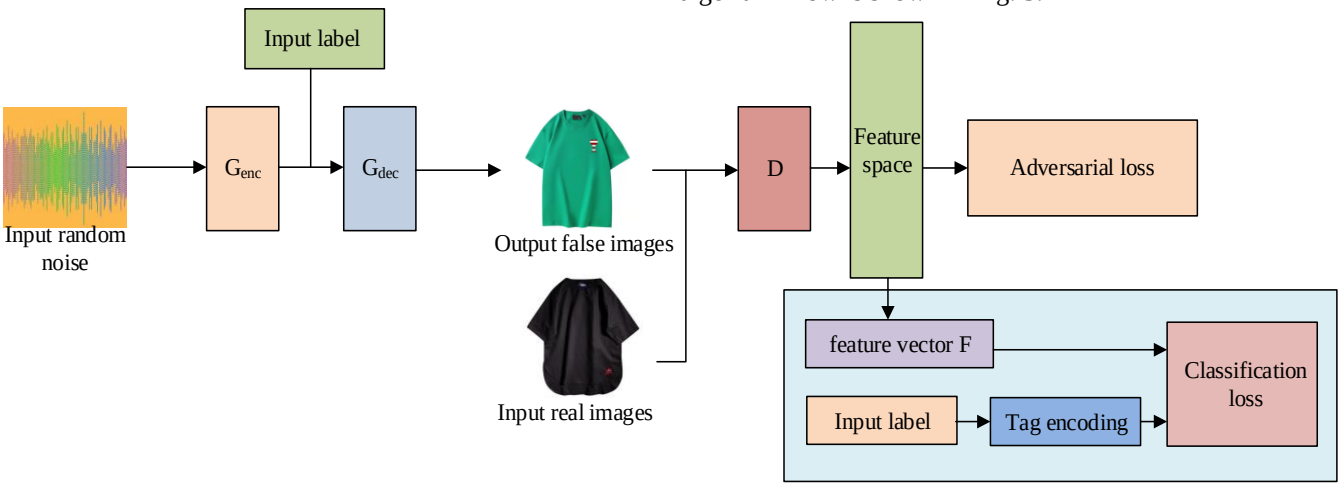


Fig. 2. Improved DCGAN network training stage structure diagram.

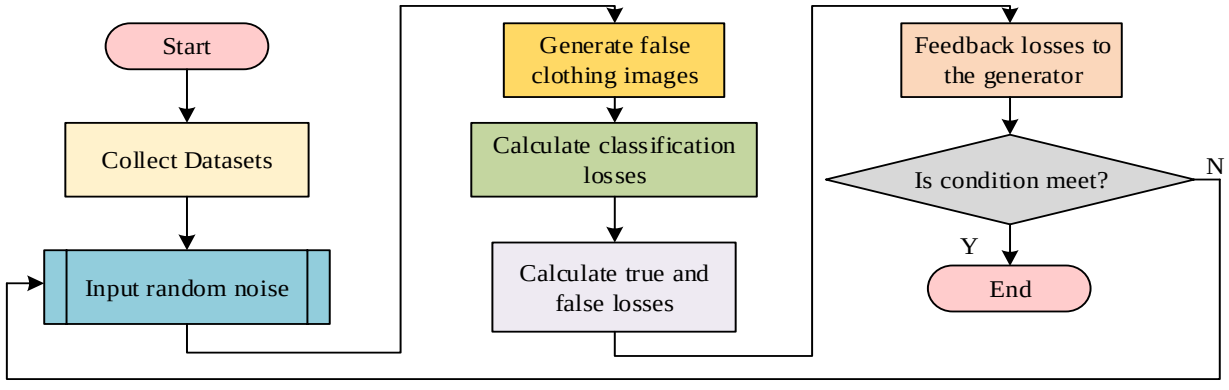


Fig. 3. IDCGN algorithm flowchart.

Fig. 3 shows the flowchart of the IDCGAN algorithm. Firstly, a dataset is collected for model training. Next, random noise is input into the encoding network of the generator to extract feature tensors. The third step is to input the extracted feature tensors and the specified style to be transformed into the decoding network of the generator, generating fake clothing images. Then, the discriminator is trained to improve its discriminative ability to distinguish the true and fake images. At the same time, the classification loss is calculated based on the feature space Z to measure the performance of the generator in generating different categories of clothing. The sixth step is to calculate the real/fake loss based on the feature space Z, which helps the generator generate more realistic clothing images. Next, the loss is fed back to the generator to adjust its strategy for generating images. Finally, steps two to seven are repeated in a loop until the total loss of the network converges, achieving the desired training effect. Through this iterative process, the IDCGAN algorithm continuously optimizes the balance between the generator and the discriminator, achieving better quality in generating fake images. The loss function used in the training process is the conditional contrastive loss. To further explain this loss function, it is necessary to first explain the NT-Xent loss function, which is expressed as Eq. (1).

$$A=\{x_1,T(x_1),...,x_m,T(x_m)\}=\{a_1,a_2,...,a_{2m}\} \quad (1)$$

In equation (1), $T(x_m)$ represents the random data augmentation for this loss function. After some transformations, the expression of the NT-Xent loss function is shown in Eq. (2).

$$\delta(a_i,a_j,t)=-\log(\frac{\exp(l(a_i)^T l(a_j)/t)}{\sum_{k=1}^{2m}\exp(l(a_i)^T l(a_j)/t)}) \quad (2)$$

In Eq. (2), t is the temperature that controls the push and pull forces. By incorporating the embedding equation into Eq. (2) and (3) is obtained.

$$\delta(x_i,y_j,t)=-\log(\frac{\exp(l(x_i)^T l(y_j)/t)}{\exp(l(x_i)^T e(y_j)/t)+\sum_{k=1}^{m}\exp(l(x_i)^T l(x_k)/t)}) \quad (3)$$

In equation (3), $e(y)$ represents the embedding equation. By adding the cosine similarity of negative samples in equation (3), the final loss function is shown in Eq. (4).

$$\delta(x_i,y_i,t)=-\log(\frac{\exp(l(x_i)^T e(y_i)/t)+\sum_{k=1}^{m}y_k=y_i\cdot\exp(l(x_i)^T l(x_k)/t)}{\exp(l(x_i)^T e(y_i)/t)+\sum_{k=1}^{m}k\neq i\cdot\exp(l(x_i)^T l(x_k)/t)}) \quad (4)$$

B. Intelligent Design Model Based on Real-Time Style Transfer Network

Real-time style transfer network is a computer vision technique used to transfer the style of an input image to another target style while preserving the content of the input image. Typically, this network combines CNN with methods for image stylization. The goal of real-time style transfer network is to perform style transformation on an image in a short period of time, making it appear as if it was drawn or rendered using the target style. By minimizing a loss function, the real-time style transfer network can generate an output image with the desired target style. The network structure is shown in Fig. 4.

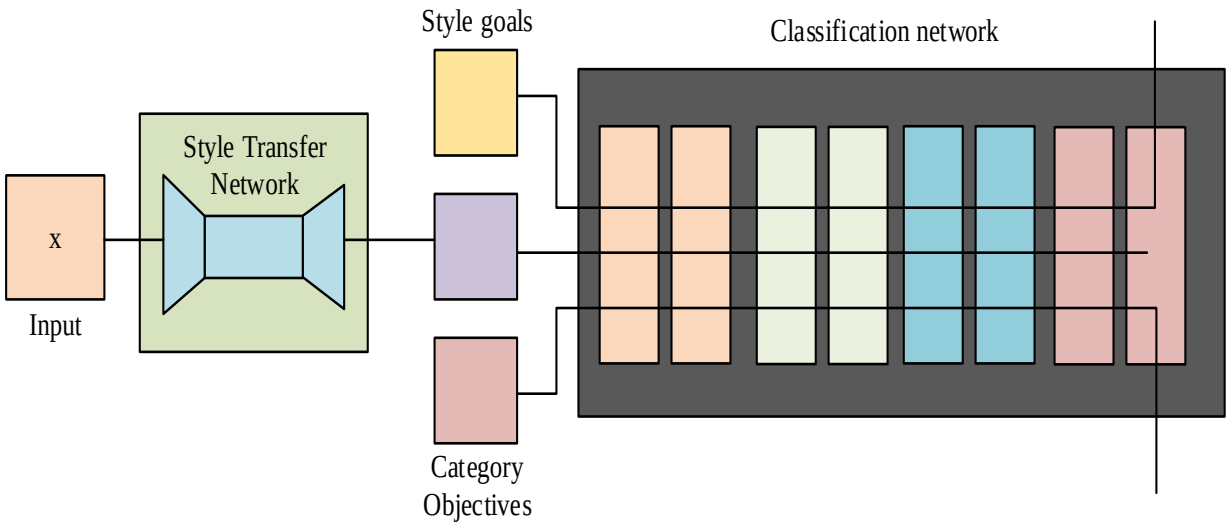


Fig. 4. Network structure of real-time style transfer network.

Fig. 4 illustrates the network structure of the real-time style transfer network. The left half represents the image transformation network, which consists of a series of CNN layers and deconvolution layers. These layers are used to gradually transform the input image into an output image with the target style. Each CNN layer can extract different features

from the input image, while the deconvolution layers are used to synthesize these features into the final output image. The right half represents the loss network, which is used to calculate the content loss and style loss. The content loss ensures that the output image preserves the content information of the input image by comparing the feature representations of the input and generated images. The style loss captures the target style features by comparing the feature statistics of the input image, generated image, and target style image. The loss function is shown in Eq. (5).

$$L = \gamma_1 L_1 + \gamma_2 L_2 \quad (5)$$

In Eq. (5), L_1 represents the content loss function, and L_2 represents the style loss function. The formula for the content loss function is shown in Eq. (6).

$$L_1^{\varphi,j}(\hat{y}, y) = \frac{1}{C_j H_j W_j} \| \phi_j(\hat{y}) - \phi_j(y) \|_2^2 \quad (6)$$

In Eq. (6), y represents the original image, and $\hat{y}$ represents the generated image. The formula for the style loss function is shown in Eq. (7).

$$L_2^{\varphi,j}(\hat{y}, y) = \| G_j^{\varphi}(\hat{y}) - G_j^{\varphi}(y) \|_F^2 \quad (7)$$

In Eq. (7), G_j^{φ} represents the Gram matrix, which is a matrix that describes the correlations between features by taking the inner product between different channels in the feature map of the j th layer. The formula for the Gram matrix is shown in Eq. (8).

$$G_j^{\varphi}(x)c, c' = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'} \quad (8)$$

In Eq. (8), ϕ represents the pre-trained network model, and $\phi_j(x)_{h,w,c}$ represents the values of the feature map in the image network with a height of h, width of w, and c channels. The detailed structure size of the generator network is shown in Fig. 5.

Fig. 5 shows a detailed schematic diagram of the network structure size for real-time style transfer. The network includes one reflection padding layer, six convolutional layers, and five residual blocks. With this network, style transfer between images can be achieved, and the time required to generate images is significantly reduced. However, although this network has achieved certain results, there is still room for improvement. For example, there are still areas that can be optimized in terms of image quality, detail preservation, and style restoration. In addition, the current network structure needs to be trained for each specific style, which limits its applicability. To address these issues, some improvements have been made to the network, as shown in Fig. 6.

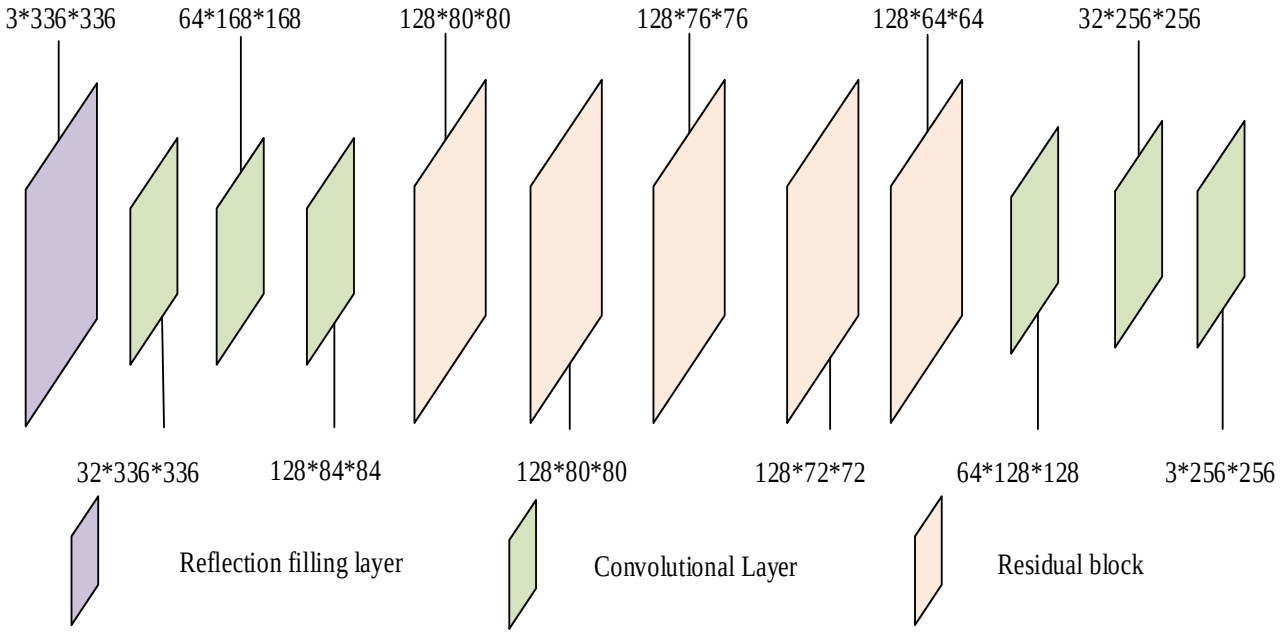


Fig. 5. Generate a detailed schematic diagram of the network structure size.

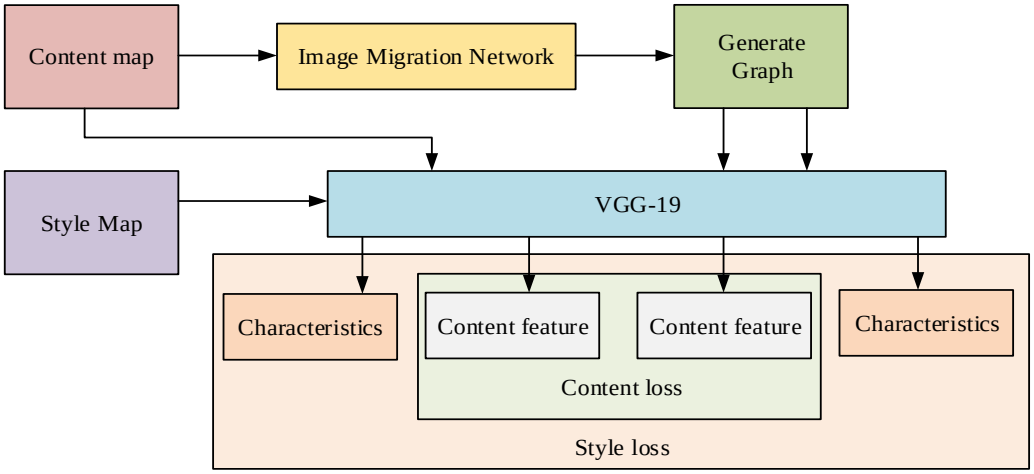


Fig. 6. Improved real-time style migration network structure diagram.

Fig. 6 shows the structure diagram of the improved real-time style transfer network. The network consists of two main components: the upper part and the lower part. In the upper part, the input content image is passed through an image encoder network to generate the generated image. This network can be a CNN that encodes the content image into an initial version of the generated image. The generated image and the style image are then passed through the VGG-19 model for feature extraction [20]. In the lower part, the content loss and style loss are propagated back through the process of backpropagation, and the pixel values of the generated image are updated using the gradient descent optimization algorithm. The optimization objective is to minimize the content loss and style loss, thereby preserving the content and matching the target style in the generated image. Additionally, the research addresses the issue of traditional normalization methods struggling to learn highly nonlinear features by introducing Adaptive Instance Normalization (AdaIN), which is formulated as Eq. (9).

$$Z_{bcwh} = \frac{x_{bcwh}}{\sum_{w=1}^{W}\sum_{h=1}^{H} x_{bcwh}} \quad (9)$$

In Eq. (9), $x \in R^{B\times C\times W\times H}$. W and H represents the width and height of the image, respectively. To allow Eq. (9) to be fitted by the ReLU activation function, a transformation is applied to the equation, as shown in Eq. (10).

$$Z_{bcwh} = \frac{x_{bcwh} - \eta}{\sqrt{\sigma^2 + \varepsilon}} \quad (10)$$

In equation (10), b represents the image index in the batch, σ^2 represents the variance, and ε represents the mean. The calculation of the variance is approximated as shown in Eq. (11).

$$\sigma_{bc}^2 = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(x_{bcwh} - h\eta_{bc})^2 \quad (11)$$

In Eq. (11), c represents the number of channels in the image. Compared to traditional instance normalization (IN), AdaIN only requires one forward pass, as shown in Eq. (12).

$$Z = \sqrt{\sigma^2 + \varepsilon}\,\frac{x - \eta_x}{\sqrt{\sigma_x^2 + \varepsilon}} + \eta_y \quad (12)$$

In Eq. (12), y represents the input values for style, and x represents the input values for content. In addition, the style loss needs to be optimized by removing the Gram matrix from the loss function of AdaIN, as shown in Eq. (13).

$$L_2^{\varphi,j}(\hat{y}, y) = \| \eta\phi_j^{\varphi}(\hat{y}) - \eta\phi_j^{\varphi}(y) \|_2 + \| \lambda\phi_j^{\varphi}(\hat{y}) - \lambda\phi_j^{\varphi}(y) \|_2 \quad (13)$$

In Eq. (13), $\|.\|_2$ represents the L2 norm.

IV. Performance Testing and Application Analysis of Pattern Intelligent Design System

This chapter is divided into two sections. The first section mainly verifies the improvement effect of the IDCGAN algorithm by comparing it with the standard DCGAN algorithm. The second section focuses on the application analysis of the pattern intelligent design system and its application in practical clothing design.

C. Comparative Experiment of IDCGAN Algorithm

In order to address some limitations of the standard DCGAN algorithm in the field of clothing design, the study made a series of improvements and finally formed the IDCGAN algorithm. To verify whether this improved algorithm is superior compared with standard DCGAN algorithm, the study used PyTorch 1.4 software on Ubuntu64-bit platform, the learning rate starting value can be set to 0.0002, batch size using batch size such as 16,32 or 64, noise dimension set to 100, loss function as binary cross-entropy loss function, IDCGAN and DCGAN using the CV-PTON dataset for 40,000 iterations. The PR curves were used as an assessment criterion. The results are shown in Fig. 7.

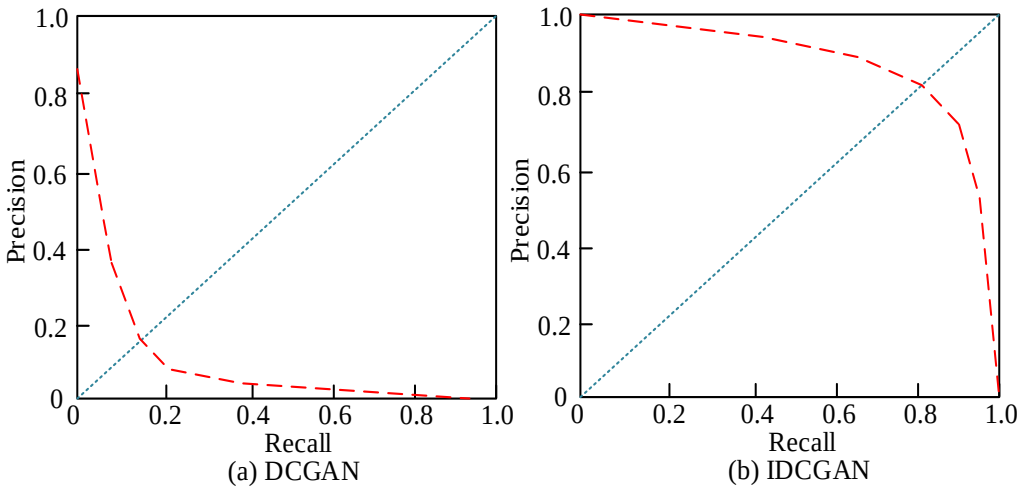


Fig. 7. Comparison of PR curves before and after DCGAN improvement.

Fig. 7 shows the changes in the PR curves before and after the improvement of the DCGAN algorithm. From Fig. 7(a), after the model convergence, the recall rate of the DCGAN algorithm has improved, but the precision has significantly decreased. This indicates that the original DCGAN algorithm may have some noise or errors in generating samples. From Fig. 7(b), it can be seen that by using a deep feature extraction network to improve the algorithm, the feature extraction capability of DCGAN is significantly enhanced. This allows the improved algorithm to generate samples with better style transfer effects while maintaining high recall rate and precision. To verify the improvement effect of the loss function comparison in this study, a similarity heatmap of the IDCGAN during the experimental process was visualized, as shown in Fig. 8.

Fig. 8 shows the similarity heatmap of IDCGAN. By observing the results in the figure, it can be seen that the contrastive loss used in this algorithm is very effective in distinguishing between input and generated patterns of the same category, with a similarity score of 1. This means that the improved algorithm can accurately identify and generate samples that are similar to the input pattern. Additionally, for different categories of style patterns, the similarity score is close to 0. This indicates that the improved algorithm can differentiate between samples of different categories and will not mistake them for similar patterns. Furthermore, a series of experiments were conducted to evaluate the resource consumption of IDCGAN and DCGAN, and the results are shown in Fig. 9.

Based on the results in Fig. 9, it can be observed that IDCGAN and DCGAN differ in terms of CPU resources, memory resources, and training time. By comparing Fig. 9(a) and Fig. 9(b), IDCGAN reduces the physical memory usage by 45.7% compared to DCGAN, with only 453MB compared to DCGAN's 834MB. Additionally, in terms of CPU resource utilization, IDCGAN only utilizes 26% of the CPU resources, indicating its relatively low computational demand. This is advantageous for devices or environments with limited resources. Furthermore, it is worth noting that the training time of IDCGAN is approximately 20 minutes and 48 seconds, while DCGAN takes about 34 minutes and 49 seconds. This comparison shows that the training time of IDCGAN is reduced by approximately 70%. This means that IDCGAN is more efficient in terms of training speed and can complete training tasks faster.

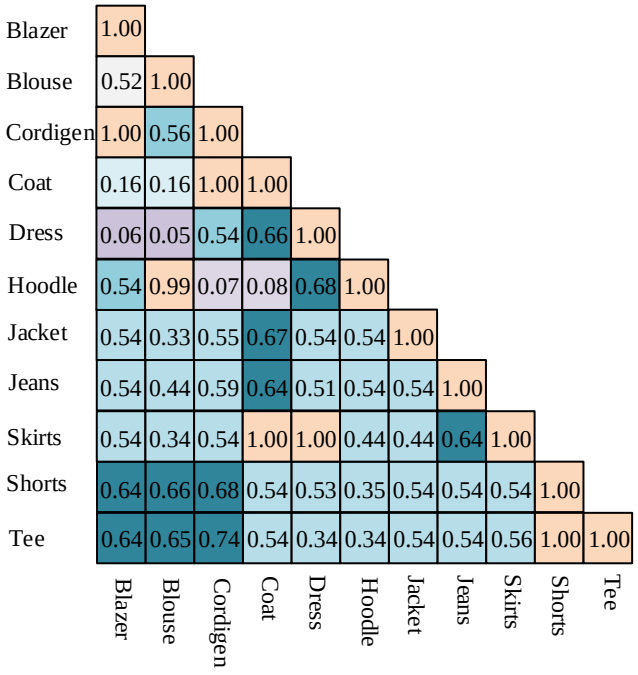


Fig. 8. Improving DCGAN similarity thermal map.

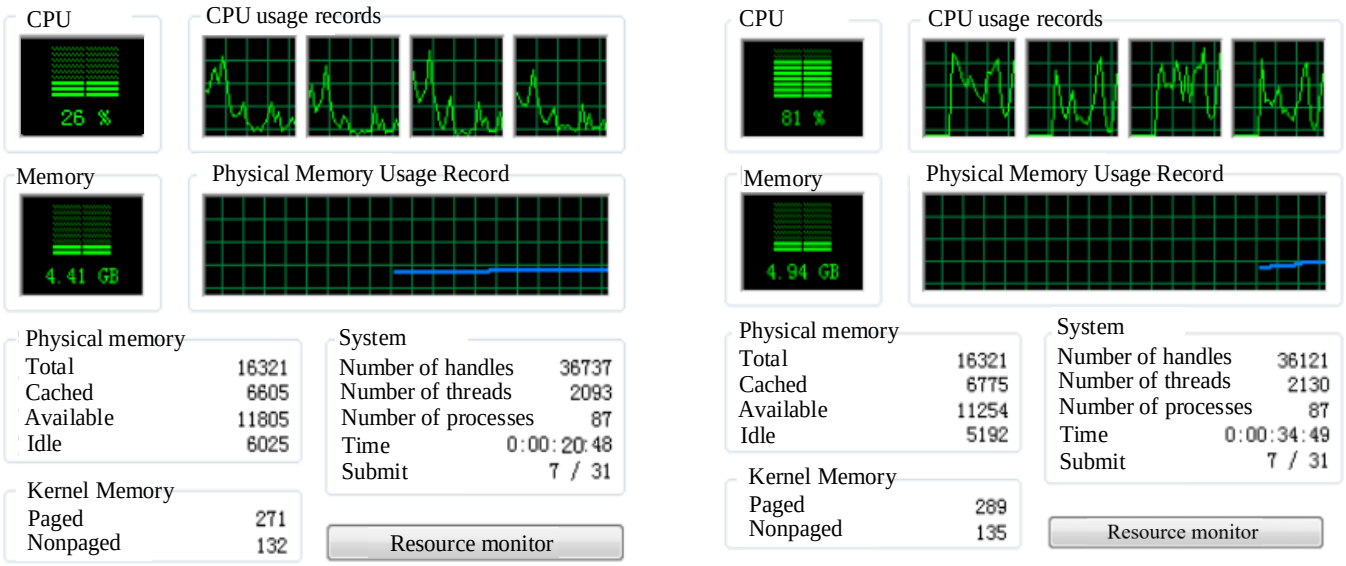


Fig. 9. Resource consumption of DCGAN and IDCGAN in various aspects.

D. Application Analysis of Style Transfer Network and Intelligent Design System

To address the limitations of real-time style transfer networks in terms of generated image quality, detail preservation, and style restoration, an improved real-time style transfer network was designed. The experiment first verified the optimization effect of the style transfer network based on AdaIN. For this purpose, batch normalization (BN) and instance normalization (IN) were introduced as control groups. The experiment was conducted using PyTorch 1.8 software on the Windows 10 platform, and the three models were trained for 62,500 iterations each. The results are shown in Fig. 10.

Fig. 10 displays the correlation of the loss and the iterations in the training based on three different normalization methods. From the figure, the loss values of the three models rapidly decrease in the first 5,000 iterations, then stabilize in the range of 50,000 to 45,000 iterations, and then decrease rapidly again. The final loss values of the models based on BN, IN, and AdaIN normalization converge to 1.02, 0.83, and 0.52, respectively. The real-time style transfer network based on AdaIN achieves the lowest loss value. To analyze the application of the proposed improved transfer model in this experiment, the trained model was applied to actual clothing design and evaluated by 36 professional fashion designers. Aesthetic quality scores and visual realism were used as evaluation criteria. The experimental results are shown in Fig. 11.

Fig. 11 provides a detailed display of the evaluation results of the images generated on AdaIN by 36 professional fashion designers. This chart clearly reflects that the proposed style transfer network has received widespread acclaim among the fashion designer community. Designers evaluated the images generated by the model rigorously and comprehensively from their professional perspectives. The results show that the 36 designers gave high aesthetic quality ratings to the images, with an average score of 96, indicating the excellent performance of the model in terms of aesthetic representation. Additionally, the designers highly recognized the visual realism of the images generated by the model, with an average score of 94.7. This score demonstrates the model's ability to successfully transfer the target style while preserving the original image content. To validate the superiority of the proposed intelligent design system in this study, comparative experiments were conducted with classical style transfer networks and standard real-time style transfer networks, as Fig. 12.

Fig. 12 displays the application effects of the three different style transfer networks. By observing Fig. 12(a), the images generated by the classical style transfer network are relatively blurry and lack detail. However, Fig. 12(b) shows that the real-time style transfer network has made significant progress compared to the classical style transfer network, but there is still room for improvement in terms of detail. In contrast, Fig. 12(c) demonstrates that the style transfer network proposed in this study preserves more details and generates style transfer results that are clearer. In conclusion, the style transfer network in this study has achieved significant improvements in image quality and detail expression.

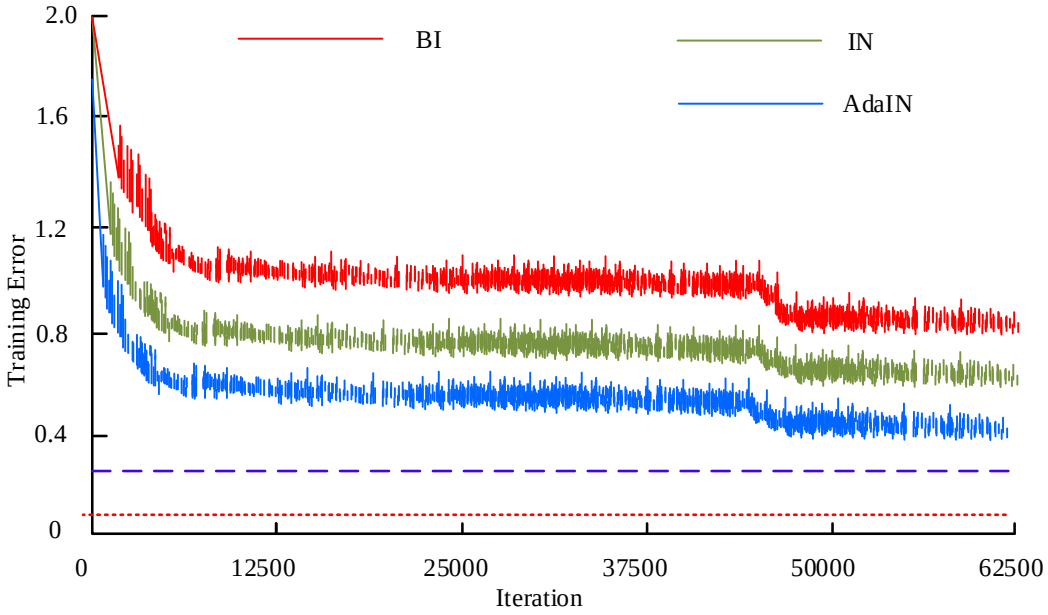


Fig. 10. Loss value variation curves of three models.

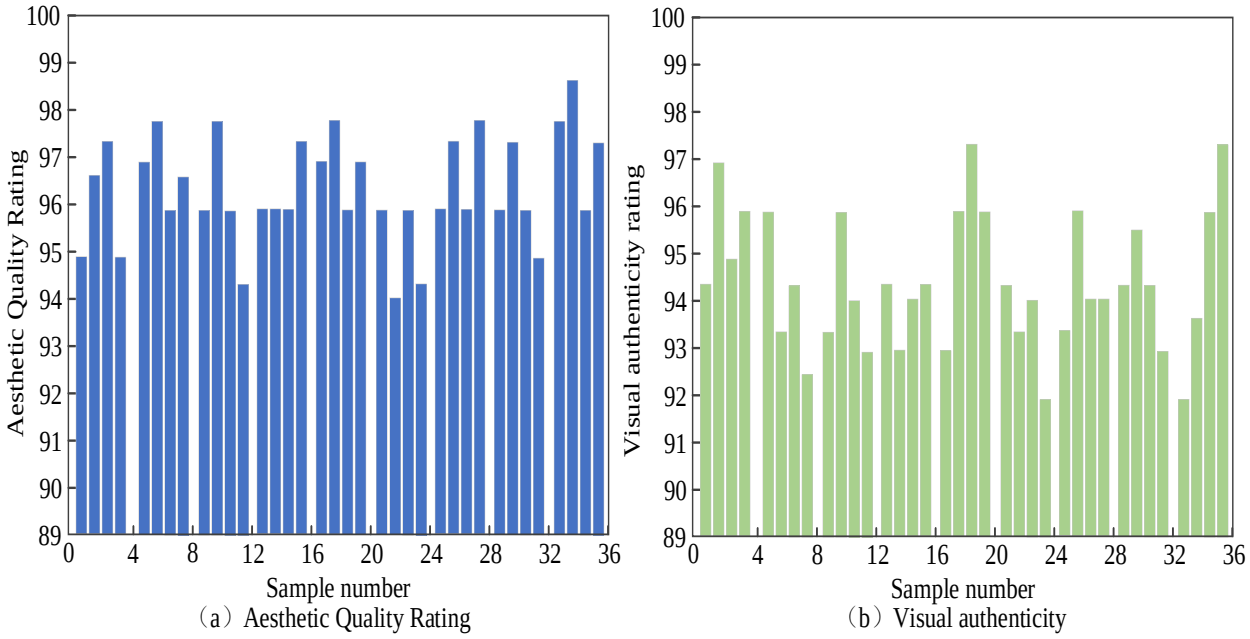
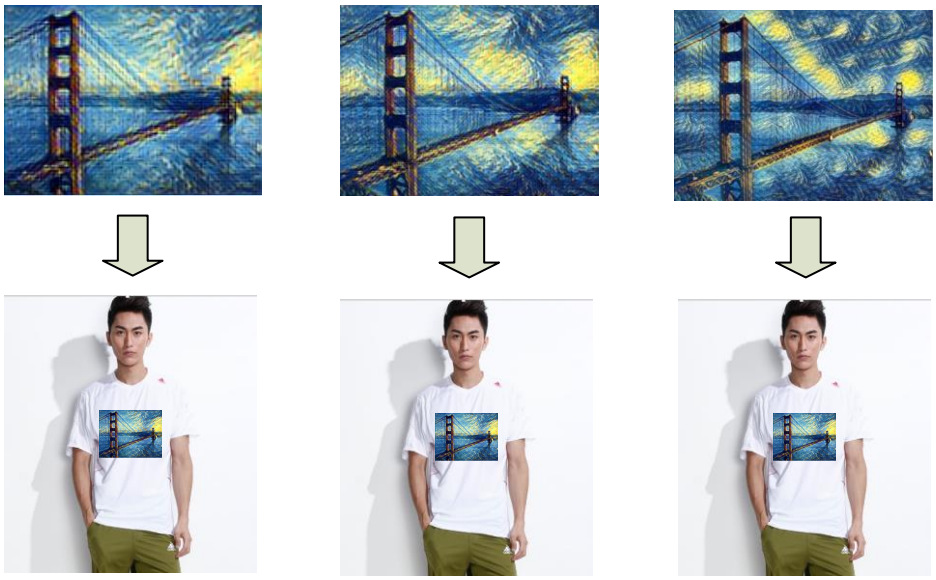


Fig. 11. Model aesthetic quality rating and visual authenticity rating.



(a) Classic style transfer

(b) Real time style migration

(c) Style transfer in this article

Fig. 12. Three styles of transfer application effects.

V. Conclusion

To address the issues of long manual design pattern creation time and low flexibility, an intelligent image design model based on the IDCGAN real-time style transfer algorithm was designed. IDCGAN The innovative content of the algorithm mainly includes the introduction of encoder and decoder network, the introduction of style vector, the method of multi-scale feature extraction and fusion, the use of conditional constraints and multi-task learning. These innovations enable the IDCGAN algorithm to generate more realistic and constrained fake images, and improve the efficiency and effect of the algorithm. For actually verifying the superiority of this improved algorithm compared to the standard DCGAN, experiments were conducted using PyTorch 1.4 software on a 64-bit Ubuntu system platform for 40,000 iterations. The results showed that after the model iterations converged, the recall rate of the DCGAN algorithm improved, but the precision significantly decreased. This indicates that the original DCGAN algorithm may have some noise or errors when generating samples. However, by using a deep feature extraction network to improve the algorithm, IDCGAN significantly enhanced the feature extraction capability of DCGAN. Additionally, the final loss values of the models based on BN, IN, and AdaIN normalization converged to 1.02, 0.83, and 0.52, respectively. The real-time style transfer network based on AdaIN achieved the lowest loss value. By comparison, IDCGAN can reduce physical memory usage by 45.7% to only 453MB, compared with DCGAN by 834MB. Moreover, in terms of CPU resource utilization, IDCGAN occupies only about 26% of the CPU resources, showing relatively low computational requirements. This is a very important advantage for scenarios where devices or environments are limited. Thus, this result indicates that IDCGAN not only generates more realistic false images, but also improves the efficiency of the algorithm. Finally, the proposed improved transfer model was subjected to application analysis, and the outcomes tells that the images generated by the classical style transfer network were relatively blurry and lacked detail. Furthermore, the aesthetic quality ratings given by the 36 designers were high, with an average score of 96, indicating that the model was well-received by the designers. On the other hand, the style transfer network proposed in this study preserved more details and generated clearer style transfer results. However, it should be noted that the model still required at least 45,000 iterations to stabilize during training, which is an aspect that needs improvement in future research. Future research will try to apply this method to more areas of image generation, such as art creation, interior design, game development, film and television special effects, etc.

References

[1] Oslund S, Washington C, So A. Multiview Robust Adversarial Stickers for Arbitrary Objects in the Physical World. Journal of Computational and Cognitive Engineering, 2022, 1(4): 152-158.

[2] Wang X, Cheng M, Eaton J. Fake node attacks on graph convolutional networks. Journal of Computational and Cognitive Engineering, 2022, 1(4): 165-173.

[3] Nimrah S, Saifullah S. Context-Free Word Importance Scores for Attacking Neural Networks. Journal of Computational and Cognitive Engineering, 2022, 1(4): 187-192

[4] Ding H, Cui Z, Maghami E, Chen, Y., Matinlinna, J. P., Pow, E. H. N., ... & Tsoi, J. K. H. Morphology and mechanical performance of dental crown designed by 3D-DCGAN. Dental Materials, 2023, 39(3): 320-332.

[5] Abd Al Rahman M, Danishvar S, Mousavi A. An improved capsule network (WaferCaps) for wafer bin map classification based on DCGAN data upsampling. IEEE Transactions on Semiconductor Manufacturing, 2021, 35(1): 50-59.

[6] Bian Y, Wang J, Jun J J,, & Xie, X. Q. Deep convolutional generative adversarial network (dcGAN) models for screening and design of small

molecules targeting cannabinoid receptors. Molecular pharmaceutics, 2019, 16(11): 4451-4460.

[7] Cheng K, Zhu L, Yao C,Yu, L., Wu, X., Zheng, X., ... & Lin, F.. DCGAN based spectrum sensing data enhancement for behavior recognition in self-organized communication network. China Communications, 2021, 18(11): 182-196.

[8] Li M, Tang H, Chan M D,Zhou, X., & Qian, X. DC-AL GAN: Pseudoprogression and true tumor progression of glioblastoma multiform image classification based on DCGAN and AlexNet. Medical physics, 2020, 47(3): 1139-1150.

[9] Ni J, Liu B, Li J, Gao, J., Yang, H., & Han, Z. Detection of carrot quality using DCGAN and deep network with squeeze-and-excitation. Food Analytical Methods, 2022, 15(5): 1432-1444.

[10] Jing Y, Liu X, Ding Y, Wang, X., Ding, E., Song, M., & Wen, S. Dynamic instance normalization for arbitrary style transfer[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(4): 4369-4376.

[11] Reimann M, Buchheim B, Semmo A,Döllner, J., & Trapp, M. Controlling strokes in fast neural style transfer using content transforms. The Visual Computer, 2022, 38(12): 4019-4033.

[12] Hollandi R, Szkalisity A, Toth T,Tasnadi, E., Molnar, C., Mathe, B., ... & Horvath, P. nucleAIzer: a parameter-free deep learning framework for nucleus segmentation using image style transfer. Cell Systems, 2020, 10(5): 453-458.

[13] Huang Z, Zhang J, Liao J. Style Mixer: Semantic-aware Multi-Style Transfer Network.Computer Graphics Forum. 2019, 38(7): 469-480.

[14] Xu K, Wang S, Jin Y, Che, Q., & Zhou, B. Object detection-oriented style transfer network for panchromatic remote sensing image. Journal of Applied Remote Sensing, 2023, 17(2): 026503-026503.

[15] Zhou L, Zhang T. AttCST: attention improves style transfer via contrastive learning. Journal of Electronic Imaging, 2023, 32(3): 033018-033018.

[16] Zhang L, Duan L, Hong X,Liu, X., & Zhang, X. Imbalanced data enhancement method based on improved DCGAN and its application. Journal of Intelligent & Fuzzy Systems, 2021, 41(2): 3485-3498.

[17] Zhang F, Wang X, Sun T,Xu, X. SE-DCGAN: a new method of semantic image restoration. Cognitive Computation, 2021, 13(4): 981-991.

[18] Li Q, Qu H, Liu Z, N., Sun, W., Sigg, S., & Li, J. AF-DCGAN: Amplitude feature deep convolutional GAN for fingerprint construction in indoor localization systems. IEEE Transactions on Emerging Topics in Computational Intelligence, 2019, 5(3): 468-480.

[19] Cho J, Moon N. Design of image generation system for DCGAN-based kids' book text . Journal of Information Processing Systems, 2020, 16(6): 1437-1446.

[20] Ikechukwu A V, Murali S, Deepu R, et al. ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images. Global Transitions Proceedings, 2021, 2(2): 375-381.

AI-Driven Optimization Approach Based on Genetic Algorithm in Mass Customization Supplying and Manufacturing

Shereen Alfayoumi[1], Neamat Eltazi[2], Amal Elgammal[3]
Department of Information Systems-Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt[1, 2, 3]
Department of Management, School of Business and Economics, NOVA LISBON Cairo Branch,
Knowledge Hub, New Administrative Capital, Cairo, Egypt[3]

***Abstract*—Numerous artificial intelligence (AI) techniques are currently utilized to identify planning solutions for supply chains, which comprise suppliers, manufacturers, wholesalers, and customers. Continuous optimization of these chains is necessary to enhance their performance. Manufacturing is a critical stage within the supply chain that requires continuous optimization. Mass Customization Manufacturing is one such manufacturing type that involves high-volume production with a wide variety of materials. However, genetic algorithms have not been used to minimize both time and cost in the context of mass customization manufacturing. Therefore, we propose this study to present an artificial intelligence solution using genetic algorithm to build a model that minimizes the time and cost which associated with mass customized orders. Our problem formulation is based on a real-world case, and it adheres to expert descriptions. Our proposed optimization model incorporates two strategies to solve the optimization problem. The first strategy employs a single objective function focused on either time or cost, while the second strategy applies the multi-objective function NSGAII to optimize both time and cost simultaneously. The effectiveness of the proposed model was evaluated using a real case study, and the results demonstrated that leveraging genetic algorithms for mass customization optimization outperformed expert estimations in finding efficient solutions. On average, the evaluation revealed a 20.4% improvement for time optimization, a 29.8% improvement for cost optimization, and a 25.5% improvement for combined time and cost optimization compared to traditional expert optimization.**

***Keywords*—Mass customization manufacturing; metaheuriatic search; genetic algorithm; optimization; supply chain management**

I. Introduction

Technological breakthroughs often give rise to new and persistent optimization dilemmas. To address these real-world challenges, metaheuristics (MHs), characterized as versatile and general-purpose methods, have been suggested as effective tools [1]. Metaheuristic optimization is focused on resolving real-world optimization problems by employing a range of metaheuristic algorithms, such as genetic algorithms, particle swarm optimization, bee algorithms, ant colony optimization, and memetic algorithms. Supply chain management poses a formidable task in the domain of continuous optimization using the power of metaheuristic optimization. It involves the simultaneous minimization of time, cost, and distance, or the maximization of quality and profit, as dictated by the problem's specifications. A supply chain is a cohesive group of organizations that are interconnected through the flow of materials, information, logistics, and finances. Each organization within this collective consists of enterprises responsible for manufacturing raw materials and components and offering services such as distribution, storage, wholesale, and retail. The ultimate customers are regarded as the concluding segment within this chain [2]. Typically, a supply chain encompasses diverse facilities such as suppliers of raw materials, manufacturers, warehouses, wholesalers, retailers, distribution hubs, and customers. The movement of materials and information occurs within and between these organizations [3].

In simpler terms, the supply chain comprises diverse components working together in a network that commences with raw material manufacturing and culminates in its shipment to storage facilities, distribution centers, and ultimately ensuring customer satisfaction [4]. The optimization of the supply chain network holds great importance when it comes to minimizing time and cost or maximizing profit. One area that requires particular attention is the manufacturing component of the supply chain. Manufacturing strategies, including make to stock, make to order, engineer to order, and mass customization, play a significant role in this regard. Mass customization revolves around producing a diverse array of products in large volumes. However, executing mass customization orders successfully poses various challenges, with a major focus on formulating a compelling value proposition that ensures customers' utmost satisfaction [5, 6], which creates the need for the optimization of the complete value creation process, the other challenge and the important one is that the mass customized products usually consume time and cost money more than the standard manufacturing forms.

In order to optimize the manufacturing lines inside the manufacturing floor, it is required to minimize the time and the cost of producing mass customized products which is the objective of this paper. The objective is not only the optimization of selecting the suitable manufacturing operations, but also the selection of proper or suitable supplier that help in minimizing time and cost. One of the AI methods for reaching this objective is genetic algorithms (GA). This includes formulating the supplying and manufacturing of mass

customization processes as an optimization problem, in addition to applying two optimization strategies, single objective time and cost separately, and multi-objective time and cost together.

The core motivation behind our proposed model is to apply genetic algorithms for the optimization of supply and manufacturing processes for mass-customized products. After constructing the model and implementing genetic algorithms, we observed substantial improvements, including a 20% reduction in time, a 30% reduction in cost, and a 25% reduction in both time and cost when compared to estimations provided by experts and consultants. This paper is structured into six sections. Section II provides an overview of existing research in the domains of supply chains and mass customization optimization from diverse viewpoints. Section III is the proposed model, followed by a discussion of Genetic Algorithms will be illustrated in the Section IV. Section V encompasses the discussion of results. The conclusions are detailed in the Section VI, and the presentation of future research directions in Section VII.

II. Related Work

Mass customization entails a highly complex supply chain with distinct features that can be classified into two key branches. The first branch encompasses the intricate relationship between the random information provided by clients' orders and the supply chain partners. This connection often gives rise to numerous scheduling conflicts and introduces dynamic or random elements into the process. The second aspect revolves around the collaborative benefits and the inherent risks within this intricate environment [7]. Therefore, this paper aims to investigate how to effectively manage these characteristics, analyze the advantages and risks associated with collaboration in mass customization, and present previous research on supply chain optimization, mass customization manufacturing and the application of GA either in supply chain or mass customization optimization in next three separate sections.

A. Supply Chain Optimization

The escalating global population and the increased demand for food, especially in aquaculture [8], have led to a surge in research focusing on the food supply chain. One notable study in this domain presented an innovative bi-objective and multi-period mathematical model for a closed-loop supply chain (CLSC) specifically tailored to the fish industry. The model is designed by using the multi-objective Keshtel algorithm (MOKA), NSGA-II, and MOSA. In addition, the Taguchi method is applied to harmonize these meta-heuristics to reach higher performance, and the ε-constraint method is used to solve small-sized problems to validate them. The results showed that the exact method cannot solve large-sized problems. The solutions are compared in terms of different performance metrics. Also, a case study with a trout CLSC in the north of Iran is investigated. The results and the case study showed that the implemented model can be applied to the suggested solution approach. The focal point of the food supply chain model mentioned is to enhance the chain's performance, prioritizing improvements regardless of time or cost implications.

A supply chain optimization problem could decide where to locate and relocate mobile and modular production units to convert biomass waste to energy [9]. Both deterministic and two-stage stochastic designs were introduced, accounting for the inherent uncertainty of how much and where biomass is produced. The framework was applied to case studies analyzing the states of Minnesota and North Carolina. Results from both states were that mobile production modules lead to reduce supply chain cost around 1–4%, or millions of dollars per year. Furthermore, this framework shows the benefit of mobile modules as a means of protection against uncertainty. Authors in that model directed their contribution to save cost by choosing the best location of the production units regardless the time.

A blood supply chain network (BSCN) [10] was formulated to minimize the total cost of the supply chain system for demand and transportation costs. The network stages considered for modeling was containing of blood donation clusters, major laboratory centers, permanent and temporary blood transfusion centers, and blood supply hubs. Other goals included determining the optimal number and location of potential facilities, optimal allocation of the flow of goods between the selected facilities and determining the most suitable transport route to distribute the goods to customer areas in uncertainty conditions. Given that the model was implemented by using NP-hard, the MFGO algorithm to solve the model with a priority-based solution. The results of the experiments' design showed the higher efficiency of the MFGO algorithm than the PSO algorithm in obtaining efficient solutions. Also, the mean of the objective function in robust approach was more than the one in the deterministic approach, while the standard deviation of the first objective function in the robust approach was less than the one in the deterministic approach at all levels of the uncertainty factor. In BSCN model time was not a factor in optimization process.

A location-inventory optimization model for supply chain (SC) configuration was presented in [11]. It included a supplier, several distribution centers (DCs), and several retailers. Customer demand and replenishment lead time were considered to be stochastic. Two classes of customer orders, priority and ordinary, were assumed based on their demand. The goal was to find the optimal locations for DCs and their inventory policy simultaneously. For this purpose, a two-phase approach based on queuing theory and stochastic optimization was developed. In the first phase, the stock level of DCs was modeled as a Markov chain process and is analyzed, while in the second phase, a mathematical program was used to determine the optimal number and locations of DCs, the assignment of retailers to DCs, and the order quantity and safety stock level at DCs. As solving this problem was NP-hard, a hybrid Genetic Algorithm (GA) was developed to make the problem computationally tractable. In location-inventory optimization model, the time and destination were considered to be factors in optimization process which were two dependable factors regardless the cost of the destination.

In time of pandemic (COVID 19), a novel multi-objective optimization model for the vehicle routing problem from suppliers of raw material to manufacturer were introduced within the realm of a factory-is-a-box framework [12]. The

key objective is the minimization of both the cumulative cost incurred while traversing network edges and the total cost accrued by visiting network nodes. This solution approach incorporates a specialized multi-objective hybrid metaheuristic algorithm that explicitly incorporates problem-specific characteristics. This model's primary objective was to optimize vehicle routing for raw material delivery to the manufacturer. In contrast, our proposed model is dedicated to optimizing the entire supply chain for raw materials and the manufacturing process, particularly for mass customization products.

B. Mass Customization Optimization

This section presents some authors who handled the manufacturing and mass customization manufacturing optimization. For optimizing manufacturing problems, a literature review was done to focus on reconfigurable manufacturing systems (RMS) optimization. This literature was classified in two scenarios. The first scenario, different optimization problems uprising in RMS were introduced and discussed. The second classification scenario presented solution approaches used to solve these problems. This work was intended to help scientists identify potential research areas in the domain of optimization for RMS. This literature showed the optimization process for reconfigurable manufacturing systems using different techniques such as mathematical programming (MP) models [13, 14], dynamic programming [15, 16], meta-heuristics [17] and heuristics [18]. RMS optimization concerned of maximizing the profit or quality or minimizing the cost regardless the time which is considering in our research. On the other hand, Mass Customization Manufacturing (MCM), existing research fingered to optimization.

A research was pointed to the fourth industrial revolution and the digital transformation of consumer marketplaces and its need in manufacturers to reshape their business models to deal with the continuous changing in customer needs and market fluctuations [19]. Currently, manufacturers are tending toward product variation strategies and more customer oriented methods to keep the competitive advantage in the Industry 4.0 environment, and mass customization is among the most famous implemented business models. Under such circumstances, an economical material supply to assembly lines has become a significant concern for manufacturers. Consequently, the proposed study concerned about optimizing the material supply to mixed-model assembly lines that contributed to the overall production cost efficiency, mainly by decreasing both the material holding and material transportation costs across production lines, while satisfying certain constraints. Given the complexity of the problem, a new two-stage heuristic algorithm is applied in such study to enable a cost-efficient delivery. To evaluate the efficiency and effectiveness of the proposed heuristic algorithm, a set of test problems were solved and compared versus the best solution found by a commercial expert. The results of the comparison reveal that the proposed heuristic offered reasonable solutions, thus presenting huge opportunities for production cost efficiency and manufacturing sustainability under the mass customization viewpoint. As seen, that research focused on minimizing the cost only without putting the factor of time into consideration.

The current global unpredictable market is characterized by increasing demand for highly customized products [20]. To thrive in this scenario, it becomes necessary to establish a closer interaction between product and manufacturing system, keeping the main focus on the customer. This paper presented a Modular Product Design (MPD) as a best strategy to produce a large product variety. MPD's configuration stage represented a key step for mass customization because it allows customers to be integrated into the value-creation process. Reconfigurable Manufacturing Systems (RMS) appear to be the most suitable manufacturing system to manufacture mass customized products due to their ability to be quickly reconfigured, adjusting their production capacity and functionality to fit new market demands. Pointing to integrate single customer needs with the decisions taken for the product and manufacturing process, this paper suggested a new 0-1 nonlinear integer programming model to optimize the configuration of modular products and RMS, driven by individual customer requests. A genetic algorithm based approach was proposed to solve this model, and its parameters were tuned with a two-full factorial design. A case study of customizable office chairs was used to illustrate the proposition, and several scenarios of customer requirements and RMS configurations were presented. Results showed that varying initial machines' configurations could highly affect the process plan and the total manufacturing costs; but, there was no confirmation that changes in initial design of configurations caused weighty effects. In summary, this work confirmed the relevance of integrating modular product and RMS configuration decisions for decreasing costs of producing mass-customized products. In RMS model, time was not considered in the optimization process.

A distributed approach for smart production management in a cellular manufacturing system was presented for offering mass-customized products [21]. This approach was based on three decision stages: factory-stage (master planning module), shop floor stage (bidding system) dealing with unexpected actions, and cell stage. The approach integrated planning, scheduling, and material handling allocation while considering real-time data from the supply chain. A mathematical model for factory-stage planning was proposed with two sequence-based resolution approaches implemented on two meta-heuristics, NSGAII and SMPSO.

C. Genetic Algorithm with Supply Chain Management and Mass Customization Manufacturing

A literature review of the application of GA on supply chain management (SCM) was published [22]. It consists of several complex processes and each process is equally important to maintain a successful supply chain. The literature review contained the eight processes of supply chain as given by Council of SCM Professionals. This literature review illustrated that there are no contributions of applying bi-objective function of minimizing the time and cost together which we focused on in our proposed model.

On the other hand, some models were designed to solve the problem of optimization of manufacturing sector [23].

However, very few models that concerned about mass customized manufacturing. Moreover, a few numbers of these models were implemented using genetic algorithm and no models were implemented using bi-objective function of minimizing time and cost in such sector [24, 25].

It was proven regarding the related work section that there were infrequently highlighted points in the optimization process in supply chain management, mass customization manufacturing and using genetic algorithm in these two fields. The following section will discuss the proposed model which takes into consideration what was neglected previously in the related work.

III. The Proposed Model

This research focuses on creating a model that integrates cost and time considerations in the optimization of mass customization. It emphasizes the collaboration between suppliers and manufacturers involved in the supply chains for mass customized products. The primary goal is to identify the most effective combinations of these entities to achieve desired objectives. To accomplish this, the study suggests the utilization of evolutionary algorithms, which are ideal for generating suitable combinations and yielding favorable results.

The suggested solution is based on using the evolutionary algorithms especially genetic algorithms to optimize the best scenario of selecting the best supplier, best operation type (either manual or automatic) and best number of manufacturing lines in order to minimize time, cost or both in a mass customized order. The definition of the mathematical formulation of the objective functions designed to obtain the optimal solution or scenario will be clarified in next paragraphs. Fig. 1 clarifies some abbreviations that will be used in the mathematical model of the proposed problem.

Before commencing a comprehensive explanation of our mathematical model, it's essential to recognize the importance of considering specific manufacturing rules. These rules play a critical role in clarifying the methods for calculating time and cost, and they are informed by the collective expertise of consultants. These rules are:

- Automatic Time = Manual Time/2.
- Automatic Manufacturing Cost = 1.6 * Manual Manufacturing Cost.

According to real-world manufacturing metrics, manual manufacturing consumes double the time of automated manufacturing, while the average cost of the automated process is 1.6 times that of manual manufacturing.

Generally, the mathematical formulation for the optimization process is being built according to an objective function and constrains that controls this function. Here, the suggested solution was divided into three choices according to the order demander priority. The customer may need to minimize the time only, cost only, or both of them. Table I illustrates the objective functions and the constraints of each priority. Within the context of the cost objective function, we come across TotSC, or total supply cost, which is subject to the influence of numerous variables such as supplier selection, category, color, and operation type. These factors are collectively evaluated to ascertain the overall cost. In the following three paragraphs, we will illustrate the three objective functions specified in Table I, encompassing a breakdown of their individual terms.

MLQ_n = Manufacturing Line Quantity for n (n: 1, 2, 3,..etc.)
MLT = Manufacturing Lead Time.
ML= Manufacturing Line.
SLT= Supplying Lead Time.
TotLT: Total Lead Time.
TotSC: Total Supplying Cost
TotMC: Total Manufacturing Cost
TotC: Total Cost.
SCO: Supplying Cost Order
TQ: Total Quantity
Op: Operation Type

Fig. 1. Abbreviations.

TABLE I. Objective Functions and Constraints

Priority	Objective Function	Subject to
Time	Min (Y) = $\sum MLT + SLT$	$1 \leq Order \leq 3$ $1000 \leq TQ \leq 6000$ $0 \leq Op \leq 1$ $1 \leq Supplier \leq 2$
Cost	Min (Y) = $\sum TotSC + TotMC + Overhead$	$1 \leq Order \leq 3$ $1000 \leq TQ \leq 6000$ $0 \leq Op \leq 1$ $1 \leq Color \leq 12$ $1 \leq Size \leq 2$ $1 \leq Supplier \leq 2$
Time and Cost	Min (Y) = $0.5(\sum (TotSC + TotMC + Overhead)) + 0.5(\sum (MLT + SLT))$	$1 \leq Order \leq 3$ $1000 \leq TQ \leq 6000$ $0 \leq Op \leq 1$ $1 \leq Color \leq 12$ $1 \leq Size \leq 2$ $1 \leq Supplier \leq 2$

When discussing the Time objective function, MLT, representing manufacturing lead time covering all phases of the manufacturing process (including assembly, painting, and packaging) is a key factor. Notably, manufacturing lines employing automatic processes outperform their manual counterparts in terms of time efficiency. In our model where the priority is for minimizing the time only, the quantity is distributed equally over the four manufacturing lines, so the time consumed is equal to any of manufacturing line. Furthermore, SLT, denoting supplying lead time, relates to the time required to deliver the components from two distinct suppliers, with Supplier A exhibiting superior delivery speed compared to Supplier B. In our model Supplier A takes half time of Supplier B. So, the total time will be MLT+SLT.

Within the context of the cost objective function, we come across TotSC, or total supply cost, which is subject to the influence of numerous variables such as supplier selection, category, color, and operation type. These factors are collectively evaluated to ascertain the overall cost.Conversely, TotMC, denoting total manufacturing cost, is predominantly governed by the choice between manual and automatic operation modes distributed over manufacturing lines, with manual operations being the more cost-efficient alterna additionally, there exists an overhead factor, representing a fixed monetary addition to the unit cost, covering various expenses like utilities (electricity, water, maintenance), and labor.

The third objective function can be regarded as a consolidation of the two objective functions discussed above. Nonetheless, in any multi-objective function, it is crucial to assign weights to individual terms during the optimization process. In this scenario, we have opted for an equal allocation of 50% weight to both time and cost, signifying their equal importance. Under these conditions, the quantity is allocated among the four manufacturing lines as detailed below:

- ML1 = Automatic = 2/5 from the Total Quantity.
- ML2= Manual = 1/5 from the Total Quantity.
- ML3= Manual = 1/5 from the Total Quantity.
- ML4= Manual = 1/5 from the Total Quantity.

This division into fifths is based on the principle that the automatic line processes double the quantity of the manual line within the same time interval.

Now that we have clarified the terms associated with each objective function in Table I, it's time to provide a comprehensive explanation of our model.

The suggested model consists of three scenarios of mass customization optimization process. The model depends on entering five inputs which are quantity, color, category, size, sub-size, and priority. The priority input is the key of which objective function will be executed.

Fig. 2 illustrates the inputs, GA processing and the output of these scenarios.

The suggested model has five inputs:

- Color: Black, Silver, White, Red, Blue, Green, Red, Brown, Pink, Purple, Golden, or Yellow.
- Category: Mountain, Tour, Road, or Folded bikes.
- Size: Child or Adult.
- Sub – Size: Small, Medium, or Large.
- Priority: Time only, Cost only, or Both.

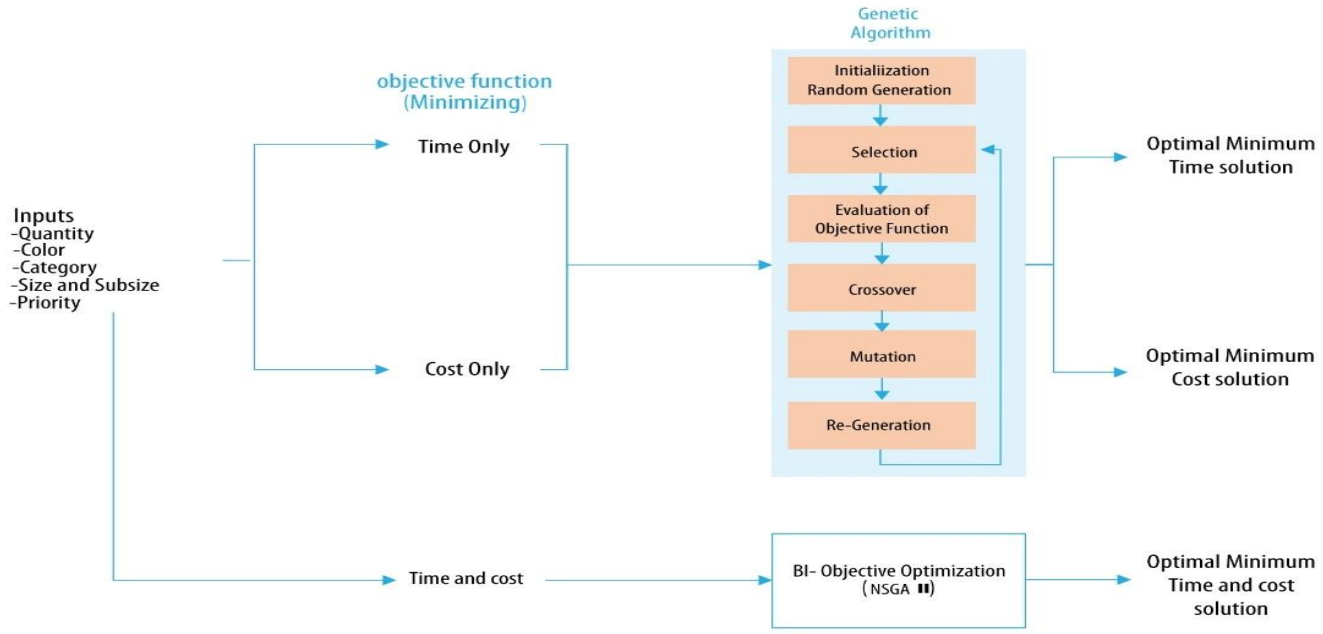


Fig. 2. The suggested optimization model.

The customer requests his/her order by determining the above inputs. However, the values of each input are limited which is considered as constrains in genetic algorithm point of view. These constraints (see Table I) are:

- Quantity : $1000 \leq TQ \leq 6000$
- Color: $1 \leq Color \leq 12$
- Category: $1 \leq Category \leq 4$
- Size: $1 \leq Size \leq 2$
- Sub-Size: $1 \leq Sub\text{-}Size \leq 3$
- Priority: Time, Cost, or Both.

From manufacturer side, there three values must be clarified which are the operation type, number of manufacturing lines and which supplier will supply the material. These values are:

- Operation Type (Op): Manual or Automatic.
- Supplier: A or B

The customer may request up to 3 different orders per one request. In addition, there are three constrains for optimization process but from the manufacturer side which are:

- Operation Type : $1 \leq Op \leq 2$
- ML=4
- Supplier = $1 \leq Supplier \leq 2$

The customer requests will be optimized using genetic algorithm to find the optimal solution of minimized time, cost or both of them. The proposed model is using two techniques of GA. For time only and cost only, single objective optimization technique is being used, while for both time and cost bi-objective optimization techniques is being implemented. Next section will explain the genetic algorithm and how the proposed model is being implemented by it.

IV. Genetic Algorithm

In recent times, meta-heuristic algorithms have gained significant popularity for addressing complex real-world challenges across various domains including engineering, manufacturing, economics, healthcare, and politics. Among these algorithms, the genetic algorithm (GA) stands out as a widely recognized approach, drawing inspiration from the process of biological evolution. Meta-heuristics can be categorized into two groups: single-solution based and population-based meta-heuristics [24]. GA falls into the category of population-based meta-heuristic algorithms.

The new populations are produced by iterative procedure of genetic operators on individuals existing in the population. The chromosome structure, selection, crossover, mutation, and evaluation of objective function computation are the basic elements of GA [24]. The GA chromosome representation and GA operators (Selection, Crossover and Mutation) will be explained and how GA was applied on the proposed solution.

A. Chromosome Representation

Genetic Algorithm draws inspiration from the evolutionary process, selecting elite solutions (chromosomes) for optimal outcomes in the search space. Chromosome in GA represents a solution and it is also called individual. Each chromosome consists of many genes according to the solution parameter.

The chromosome of our proposed solution consists of five genes. Fig. 3 represents the chromosome which represents the form of solution that the manufacturer will execute. First gene describes the supplier group that suits to the objective function. Rest of genes is the manufacturing lines (MLi) and if it will operate automatically or manually according to the objective function as well.

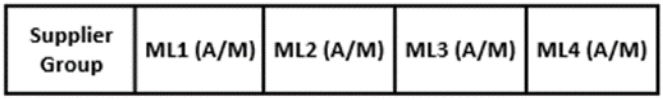


Fig. 3. The chromosome representation.

In the first gene of the Supplier Group, there are two categories. The first category is associated with higher expenses but faster raw material deliveries; whereas the second category offers cost savings but slower material delivery. The subsequent genes in the sequence represent the manufacturer's internal manufacturing lines. When these lines operate manually, manufacturing time increases but cost decreases.

Conversely, automatic operation reduces manufacturing time but increases cost. Within genetic algorithms, there are two terms called "genotype" and "phenotype" which signify the relationship between the proposed genetic chromosomes and their actual chromosomes post-processing. In the realms of artificial intelligence optimization and computer science, these two terms are fundamentally interchangeable. Hence, there is no inherent difference, eliminating the necessity to define a distinct ratio between them.

B. Selection Process

Selection is a first feature to go forward of finding the nearest solution, which commonly services Evolutionary Computation (EC) [26]. In general, there are many selection techniques in GA, such as roulette wheel, tournament, rank, Boltzmann, and stochastic universal sampling [24]. The most usable technique and it is used in the proposed solution is roulette wheel selection (RWS). It works on selecting specific solutions that will share in forming the next generation. It gives each solution or individual of the recent generation the probability to be selected in the next generation according to its proportionality to the objective function value [27].

C. Crossover

It is one the GA operators that is done between two chromosomes called parents by choosing randomly either single point or multiple points to swap what beyond these chosen point(s) [28]. The result of the crossover process is new modified chromosomes called off-springs. Fig. 4 describes the implementation of crossover types in the suggested solution. It describes a single crossover

implementation and a multi crossover implementation in the proposed solution.

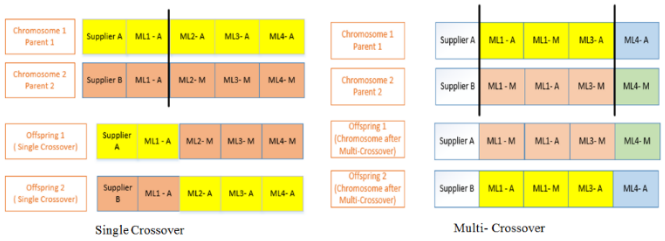


Fig. 4. Crossover operator single crossover, multi crossover.

D. Mutation

Mutation is another operator that keeps the genetic variety and diversity from one population to the succeeding population [29]. It is performed by choosing at least one or more genes randomly and changing their values. The value of the objective function is then recalculated. Fig. 5 gives an example of mutation process by choosing a random gene and changing its value according to the proposed solution.

Fig. 5. Applying mutation operator

The steps of implementing the GA can be clarified through the pseudo-code as shown in Fig. 6.

Parameter(s): C- set of Chromosomes

Output: super-offspring of set C

Initialization:
$t \leftarrow 0$
Initialize I_t *to random individuals from* C^*
Evaluate the solution objective value
while *stop criterion not met*
Do
    Select individuals from I_t
    Crossover individuals
    Mutate individuals
    Re- Evaluate the solution objective value
    $I_{t+1} \leftarrow$ *new created individuals*
    $t \leftarrow t + 1$
End Do
return (*super-offspring derived from best individual in* I_t)

Fig. 6. Genetic algorithm pseudocode.

V. Discussion and Results

To the best of our knowledge, this is the first study to solve the problem through applying the optimization methodology; moreover, the problem was formulated mathematically according to the real manufacturing case, and according to the consultants inside the manufacturer and experts in mass customization manufacturing field. So, they are our reference and benchmark in our proposed study. In a practical case study focused on a mass customization bicycle manufacturer, the proposed model was put into action. Customers had the option to place orders for bicycles in diverse categories, sizes, and colors, as previously mentioned. Moreover, they were able to order bicycles in large quantities.

The objective is to optimize the combination of supplying and manufacturing processes from the manufacturer to execute the orders in minimum time, minimum cost or both to achieve the customer goal. The three mathematical models and GA were implemented using MATLAB [30]. Two of these mathematical models used the single objective GA.

However, the third model was implemented using NSGAII technique [31] to solve the multi-objective optimization problem. Table II illustrates the values of each parameter used in GA implementation. Each genetic algorithm code asks for assigning values of 4 main parameters. These parameters are population size [24], number of generations, crossover rate and mutation rate.

Each objective function is tested on three different quantities 2000, 3500 and 5500 bicycles with different categories, colors, sizes and sub-sizes. Table III illustrates the enhancement of solutions in form of enhancement of GA generations.

The results in Table III shows the modifications of solutions for time only and cost only columns versus the number of generations. While the last column shows the modifications of time and cost together over the number generations using NSGAII technique. In bi-objective function mode, time is being represented in Y-axis and cost is in X-axis. In Table III, the graphs in each row determine the values of the most optimum time, cost or time and cost with every round of generation (100 rounds) until reaching the minimum optimum solution.

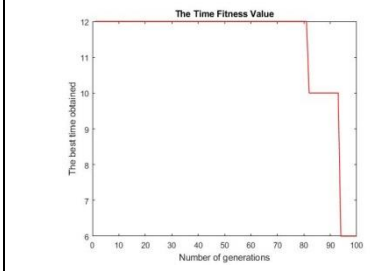
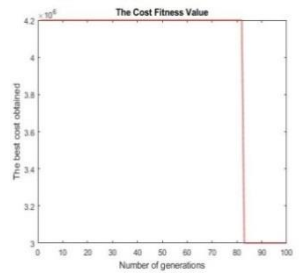
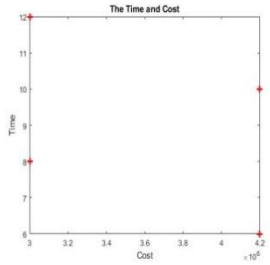
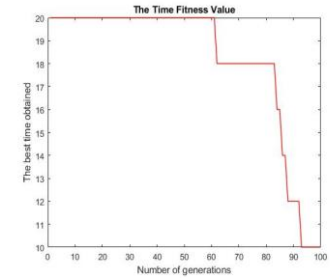
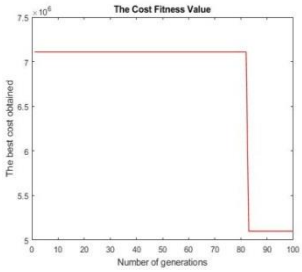
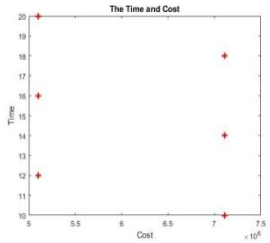
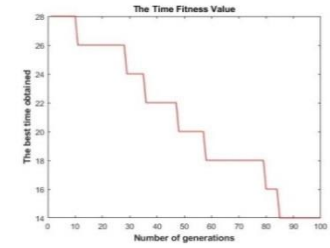
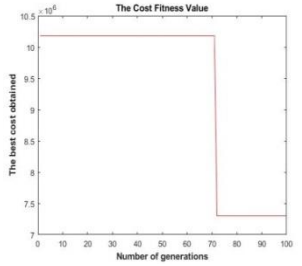
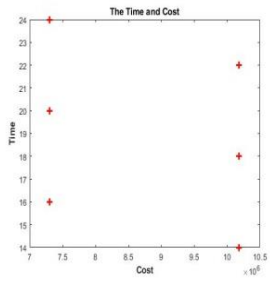
When examining the time column graphs, it becomes evident that values are progressively minimized with each generation. Conversely, the cost column experiences abrupt reductions, primarily because cost is subject to fixed prices determined by various factors, leading to the characteristic sharp curve.

After the observation of 100 solutions per each iteration, eight forms of solutions were noticed but different objective function values. Fig. 7 presents these forms of these solutions (chromosomes that mentioned in Fig. 3).

TABLE II. GA Parameter Setting

Parameter	Value
Population Size	100
Number of Generation	100
Crossover Rate	0.85
Mutation Rate	0.01

TABLE III. ENHANCEMENT OF GA GENERATIONS FOR THREE DIFFERENT CASES

Quantity	Time	Cost	Time and Cost
2000			
3500			
5500			

Solution 1	Supplier A	ML1 - A	ML1- A	ML3- A	ML4- A
Solution 2	Supplier B	ML1 - A	ML1- A	ML3- A	ML4- A
Solution 3	Supplier A	ML1 - A	ML1- A	ML3- A	ML4- M
Solution 4	Supplier B	ML1 - A	ML1- A	ML3- A	ML4- M
Solution 5	Supplier A	ML1 - A	ML1- A	ML3- M	ML4- M
Solution 6	Supplier B	ML1 - A	ML1- A	ML3- M	ML4- M
Solution 7	Supplier A	ML1 - M	ML1- M	ML3- M	ML4- M
Solution 8	Supplier B	ML1 - M	ML1- M	ML3- M	ML4- M

Fig. 7. Forms of solutions.

The evaluation of the results involved consulting production management experts in the bicycle manufacturing industry. Their estimations, obtained using traditional computation techniques, were compared with the outcomes generated by GA and NSGAII, as shown in Table IV. The table presents numerical data for three different cases, each involving different quantities. The evaluation considered time, cost, and a combination of both for each quantity. The results section of the table includes the expert estimation, GA results, and the percentage improvement achieved by our proposed models over the expert estimation. The findings clearly indicate that treating time and cost as a multi-objective functions leads to optimal results. Additionally, it was observed that higher quantities ordered corresponded to greater enhancements achieved through GA.

The expert estimations documented in Table IV are cited in order to facilitate a comparison between our GA-generated results and the authentic data stored within the database of a prominent bicycle manufacturer.

Table V illustrates the average of enhancement percentage of GA and NSGAII versus the experts estimates.

TABLE IV. SAMPLE OF NUMERIC RESULTS

Solution Forms	Time (Week)			Cost (x10^6)			Time (Week) and Cost (x10^6)					
	Expert Estimation	GA Result	%	Expert Estimation	GA Result	%	Expert Estimation		NSGAII Result		%	
							Time	Cost	Time	Cost	Time	Cost
	Quantity : 2000 Bicycles											
Solution 1	8	6	25	6.3	4.2	33.3	8	6.3	6	4.2	25	33.3
Solution 2	10	6	40	5.9	3	49.1	10	5.9	8	3	20	49.1
Solution 3	9	6	33.3	5.5	4.2	23.6	9	5.5	6	4.2	33.3	23.6
Solution 4	11	10	9.1	5.1	3	41.2	11	5.1	8	3	27.3	41.2
Solution 5	10	10	0	4.7	4.2	10.6	10	4.7	10	4.2	0	10.6
Solution 6	12	10	16.6	4.4	3	31.8	12	4.4	12	3	0	31.2
Solution 7	11	10	9.1	4.3	4.2	2.4	11	4.3	10	4.2	9.1	2.3
Solution 8	13	12	7.7	3.5	3	14.3	13	3.5	12	3	7.7	14.3
Quantity : 3500 Bicycles												
Solution 1	14	10	28.5	11.5	7.11	38.2	14	11.5	10	7.1	28.6	38.3
Solution 2	16	12	25	10.0	5.1	49	16	10.0	12	7.1	25	29
Solution 3	16	12	25	9.8	7.11	27.4	16	9.8	12	5.1	25	48
Solution 4	18	14	22.2	9.2	5.1	44.5	18	9.2	14	5.1	22.2	44.6
Solution 5	20	16	20	8.6	7.11	17.3	20	8.6	14	7.1	30	17.4
Solution 6	20	18	10	7.8	5.1	34.6	20	7.8	16	7.1	20	9
Solution 7	22	18	18.2	7.2	7.11	1.4	22	7.2	18	5.1	18.2	29.2
Solution 8	24	20	16.6	5.6	5.1	9	24	5.6	20	5.1	16.7	8.9
Quantity : 5500 Bicycles												
Solution 1	20	14	30	17.5	10.1	42.3	20	17.5	14	10.2	30	41.7
Solution 2	22	16	27.2	16.3	7.3	55.2	22	16.3	16	7.3	27.3	55.2
Solution 3	24	18	25	15.5	10.2	34.2	24	15.5	18	10.2	25	34.2
Solution 4	26	20	23.1	14.3	7.2	49.6	26	14.3	20	7.3	23.1	49
Solution 5	28	22	21.4	13.5	10.3	23.7	28	13.5	22	10.2	21.4	24.4
Solution 6	30	24	20	12.3	7.2	41.5	30	12.3	24	7.3	20	40.7
Solution 7	32	26	18.8	11.5	10.3	10.4	32	11.5	24	10.2	25	11.3
Solution 8	34	28	17.6	10.5	7.2	31.4	34	10.5	24	7.3	29.4	30.5

TABLE V. GA AND NSGAII AVERAGE OF REFINEMENT PERCENTAGE

Quantity	2000	3500	5500	Total Average
Time	17.6%	20.7%	22.9%	20.4
Cost	25.8%	27.7%	36.03%	29.8
Time and Cost	20.5%	25.6%	30.5%	25.5

VI. CONCLUSION

In this paper, a genetic-based approach was introduced for the supply and manufacturing of mass customization products. The approach integrated the supply and manufacturing phases to achieve the best possible solution with minimal time and cost.

The study presented two techniques utilizing Genetic Algorithms: one for solving single objectives of time and cost, and another for optimizing both objectives simultaneously using the multi-objective NSGAII.

Experimental results, including numerical and graphical analyses, demonstrated the significant improvements of the proposed approach compared to traditional factory methods. On average, the proposed model enhances time optimization by 20.4%, cost optimization by 29.8%, and both time and cost optimization by 25.5%.

VII. FUTURE WORKS

Optimization constitutes a vast and dynamically evolving field that accommodates numerous artificial intelligence technologies. In the context of our planning model, we have the flexibility to refine and tailor it further by adopting novel hybrid heuristics and metaheuristic search techniques, including adaptive algorithms, self-adaptive algorithms, or combinations of existing methods. Moreover, our future work includes exploring the potential of applying machine learning

as an optimizer within our model. On a parallel front, the domain of supply chain optimization offers a rich landscape for exploration. We aim to extend the horizons of optimization by considering the comprehensive optimization of the entire supply chain network, employing new evolutionary algorithms, machine learning, and deep learning approaches.

ACKNOWLEDGMENT

We thank everyone who helped us in manufacturing field to apply, test and evaluate our proposed model in minimizing cost and time for supplying and manufacturing mass customization products.

REFERENCES

[1] J. M. Cruz-Duarte, I. Amaya, J. C. Ortiz-Bayliss, S. E. Conant-Pablos, H. Terashima-Marín, and Y. Shi, "Hyper-Heuristics to customise metaheuristics for continuous optimisation," Swarm and Evolutionary Computation, vol. 66, p. 100935, Oct. 2021.

[2] M. Fathi, M. Khakifirooz, A. Diabat, and H. Chen, "An integrated queuing-stochastic optimization hybrid Genetic Algorithm for a location-inventory supply chain network," International Journal of Production Economics, vol. 237, p. 108139, Jul. 2021.

[3] Nozari, Najafi, Fallah, and Lotfi, "Quantitative Analysis of Key Performance Indicators of Green Supply Chain in FMCG Industries Using Non-Linear Fuzzy Method," Mathematics, vol. 7, no. 11, p. 1020, Oct. 2019.

[4] H. Kazemipoor and M. Ghanbarzadeh, "Solve a New Robust Bi-Objective Model for Designing Blood Supply Chain Network by NSGA II and Imperialist Competitive Algorithm," International Journal of Innovation in Engineering, vol. 1, no. 1, pp. 1–21, Jan. 2021.

[5] J. C. de Man and J. O. Strandhagen, "An Industry 4.0 Research Agenda for Sustainable Business Models," Procedia CIRP, vol. 63, pp. 721–726, 2017.

[6] G. Baranauskas, "Digitalization impact on transformations of Mass Customization concept: Conceptual modelling of online customization frameworks," Mark Manag Innov, 3, p. 120-132, 2020.

[7] M. Jin, H. Wang, Q. Zhang, and Y. Zeng, "Supply chain optimization based on chain management and mass customization," Information Systems and e-Business Management, vol. 18, no. 4, pp. 647–664, Jan. 2019.

[8] M. Fasihi, R. Tavakkoli-Moghaddam, S. E. Najafi, and M. Hajiaghaei, "Optimizing a bi-objective multi-period fish closed-loop supply chain network design by three multi-objective meta-heuristic algorithms," Scientia Iranica, vol. 0, no. 0, Oct. 2021.

[9] A. Allman, C. Lee, M. Martín, and Q. Zhang, "Biomass waste-to-energy supply chain optimization with mobile production modules," Computers & Chemical Engineering, vol. 150, p. 107326, Jul. 2021.

[10] J. Ghahremani-Nahr, H. Nozari, and M. Bathaee, "Robust Box Approach for Blood Supply Chain Network Design under Uncertainty: Hybrid Moth-Flame Optimization and Genetic Algorithm," International Journal of Innovation in Engineering, vol. 1, no. 2, pp. 40–62, Jul. 2021.

[11] M. Fathi, M. Khakifirooz, A. Diabat, and H. Chen, "An integrated queuing-stochastic optimization hybrid Genetic Algorithm for a location-inventory supply chain network," International Journal of Production Economics, vol. 237, p. 108139, Jul. 2021.

[12] J. Pasha et al., "Exact and metaheuristic algorithms for the vehicle routing problem with a factory-in-a-box in multi-objective settings," Advanced Engineering Informatics, vol. 52, p. 101623, Apr. 2022.

[13] M. Rabbani, M. Samavati, M. S. Ziaee, and H. Rafiei, "Reconfigurable Dynamic Cellular Manufacturing System: A New Bi-Objective Mathematical Model," RAIRO - Operations Research, vol. 48, no. 1, pp. 75–102, Jan. 2014.

[14] Y.-C. Choi and P. Xirouchakis, "A holistic production planning approach in a reconfigurable manufacturing system with energy consumption and environmental effects," International Journal of Computer Integrated Manufacturing, vol. 28, no. 4, pp. 379–394, Apr. 2014.

[15] P. A. Borisovsky, X. Delorme, and A. Dolgui, "Balancing reconfigurable machining lines via a set partitioning model," International Journal of Production Research, vol. 52, no. 13, pp. 4026–4036, Oct. 2013.

[16] L. Tang and Y. Meng, "Data analytics and optimization for smart industry," Frontiers of Engineering Management, Aug. 2020.

[17] P. Borisovsky, "Genetic Algorithm for One Machining Line Balancing Problem with Setup Times," In 2020 Dynamics of Systems, Mechanisms and Machines (Dynamics), pp. 1-5, November 2020.

[18] S. Ghanei and T. AlGeddawy, "An Integrated Multi-Period Layout Planning and Scheduling Model for Sustainable Reconfigurable Manufacturing Systems," Journal of Advanced Manufacturing Systems, vol. 19, no. 01, pp. 31–64, Mar. 2020.

[19] M. Fathi and M. Ghobakhloo, "Enabling Mass Customization and Manufacturing Sustainability in Industry 4.0 Context: A Novel Heuristic Algorithm for in-Plant Material Supply Optimization," Sustainability, vol. 12, no. 16, p. 6669, Aug. 2020.

[20] R. C. Sabioni, J. Daaboul, and J. Le Duigou, "An integrated approach to optimize the configuration of mass-customized products and reconfigurable manufacturing systems," The International Journal of Advanced Manufacturing Technology, vol. 115, no. 1–2, pp. 141–163, May 2021.

[21] E. Maalouf, J. Daaboul, J. Le Duigou, and B. Hussein, "Production management for mass customization and smart cellular manufacturing system: NSGAII and SMPSO for factory-level planning," The International Journal of Advanced Manufacturing Technology, vol. 120, no. 9–10, pp. 6833–6854, Apr. 2022.

[22] S. K. Jauhar and M. Pant, "Genetic algorithms in supply chain management: A critical analysis of the literature," Sādhanā, vol. 41, no. 9, pp. 993–1017, Sep. 2016.

[23] J. M. Yao, "Scheduling optimisation of co-operator selection and task allocation in mass customisation supply chain based on collaborative benefits and risks," International Journal of Production Research, vol. 51, no. 8, pp. 2219–2239, Apr. 2013.

[24] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," Multimedia Tools and Applications, vol. 80, Oct. 2020.

[25] N. BAYĞIN and M. KARAKÖSE, "Genetic Algorithm-Based Optimization of Mass Customization Using Hyperledger Fabric Blockchain," Turkish Journal of Science and Technology, Aug. 2022.

[26] Z. Huang, "A feature selection approach combining neural networks with genetic algorithms," AI Communications, vol. 32, no. 5–6, pp. 361–372, Mar. 2020.

[27] K. Jebari, and M. Madiafi, "Selection methods for genetic algorithms," International Journal of Emerging Sciences, 3(4), pp. 333-344, Dec. 2013.

[28] K. B. Ali, A. J. Telmoudi, and S. Gattoufi, "Improved genetic algorithm approach based on new virtual crossover operators for dynamic job shop scheduling," IEEE Access, 8, 213318-213329, 2020.

[29] C. Liu and A. Kroll, "Performance impact of mutation operators of a subpopulation-based genetic algorithm for multi-robot task allocation problems," SpringerPlus, vol. 5, no. 1, Aug. 2016.

[30] S. Hrehová and Jozef Husár, "Selected Application Tools for Creating Models in the Matlab Environment," Springer eBooks, pp. 181–192, Jan. 2022.

[31] A. Aminmahalati, A. Fazlali, and H. Safikhani, "Multi-objective optimization of CO boiler combustion chamber in the RFCC unit using NSGA II algorithm," Energy, vol. 221, p. 119859, Apr. 2021.

Application Model Construction of Emotional Expression and Propagation Path of Deep Learning in National Vocal Music

Zhangcheng Tang*
Music and Dance Academy, Hunan First Normal University, ChangSha, 410205, China

***Abstract*—Emotional expression is important in Chinese national vocal music art. The emotional expression in national vocal music is based on the art of national vocal music, with distinct characteristics and requirements. The ultimate goal is to spread the expression of various emotions in the national vocal music art. Promoting the spread of national vocal music singing art using modern media is an urgent requirement for the inheritance and development of national vocal music singing art. With the rapid development of science and technology, integrating deep learning and traditional music has become the general trend. It has been gradually applied to melody recognition, intelligent composition, virtual performance, and other aspects of traditional music and has achieved good results, but also hidden behind a series of ideas and technical and ethical issues. In this paper, the application of deep learning has been discussed and prospected. The recognition rate of emotional expression in national vocal music is 92 %. In terms of communication, combined with the deep learning algorithm, this paper analyzes the characteristics and requirements of emotional expression in the art of national vocal music singing and puts forward a new method of promoting the development of the art of national vocal music singing, hoping to attract more attention and enhance the social awareness of the application field, to promote the steady development of Chinese traditional music in the information age.**

***Keywords*—Deep learning; national vocal music; innovation; emotion; dissemination**

I. INTRODUCTION

Emotional expression has always been particularly valued in national vocal music. The emotional expression running through the art of national vocal music is the soul of the art of national vocal music. The art of national vocal music without emotional expression will lose its significance. The emotional expression ability of singers is the focus of national vocal music singing art. The emotional expression of Chinese national vocal music singing art has accumulated rich levels and diversity in the development of thousands of years, with its unique characteristics and requirements. As the pearl in the art treasure house, the inheritance of national vocal music singing art is related to the integrity of traditional art [1], [2]. While analyzing the emotional expression of national vocal music singing art, continuously improving its artistic and emotional expression ability and meeting the current social and cultural needs, the inheritance of national vocal music singing art has also become a social problem. In the face of the new era of openness and diversification, how to take effective communication paths to make the national vocal music singing art occupy a place in the field of communication and inherit it for a long time is a problem that must be thought and solved simultaneously in addition to paying attention to the emotional expression of the national vocal music singing art.

Music education is gradually developing towards intelligence and online in the Internet era. As an emerging discipline, deep learning is mainly used to research, develop, and extend the method theory of human intelligence. Nowadays, many things are related to deep learning, such as fingerprint recognition, intelligent search, face recognition, language translation, and automatic planning. Deep learning and hearing has a great relationship, and deep learning for music education has a unique advantage. The breakthrough of material civilization also accompanied the development of music art. Deep learning will change music education's teaching mode and theory, especially in teaching means and methods, to provide positive practical value for music education in the Internet era [3].

Combining deep learning algorithms and traditional national vocal music has become an important research direction in the future. The singing of traditional national vocal music will usher in great changes under the impetus of deep learning. It makes the emotional expression of traditional folk music singing easier to identify and push to the audience who like to change the type of folk music so that it is not limited to the nation. The technology of retrieving music based on music's emotional attributes has also been developed. Many kinds of music are stored in the network music database. The essence of music information retrieval is music recognition and classification. Most music end users often like a certain type of music, and the diversity of music gives it unique attributes [4], [5]. Therefore, a music recognition and classification system can help people retrieve and manage music more effectively. MIR contains many sub-tasks, such as music emotion recognition, instrument recognition, genre recognition, and author recognition [6].

The traditional extraction of music emotion features often lacks the temporal structure and related semantics of music because it is usually extracted in frames, and the deep semantics of chords, melody, and rhythm in music that change with time is important for music emotion recognition. To analyze the correlation between temporal information and emotional expression in the research process, some scholars have done verification experiments in 2014: First, the music is

transformed into a feature vector of the time structure, and then the Gaussian mixture model, Markov and Hidden Markov model and other generative models are used to form the temporal characteristics of the music. Finally, the model is passed through the Probability Product Kernel for music emotion recognition. Traditional folk music has been rapidly upgraded, making full use of the convenience of the mobile Internet and spreading through more channels. Therefore, the mobile Internet has unprecedented advantages for promoting music works and the comprehensive utilization of traditional Internet, mobile communication networks, mobile phone ringtones, and other wired and wireless platforms for communication [7]–[10]. This mode of creation, production, dissemination, and acceptance of music products mainly relies on digitalization, reflecting technological and cultural progress and transforming and upgrading communication modes. The Internet and mobile terminal technology have achieved unprecedented development speed since the 21st century, especially in the past five years. It has not only formed a pattern of national coverage of online music but also gradually replaced the traditional music platform as the main channel in the rapid development of this mode of communication and platform. The mode of communication has developed into one of the most fashionable ways of music communication. Some scholars have applied deep learning to music content analysis, especially style, and artist recognition, producing meaningful features from the preprocessed spectrum. The deep learning features have higher accuracy than those standard Mel Frequency Cepstral Coefficients features for music style recognition. There is also a convolutional neural network (CNN) that MFCC features vectors and inputs them into three hidden layers. Finally, it can automatically extract image features for classification, indicating that CNN can capture changing image information features [11]–[13].

The role of national vocal music art in the mass media has also changed; specific singing techniques and emotional expression also began to change. Although from the perspective of modern communication, this is the development and progress of national vocal music art, to some extent, it can also be said to be desalination. In the modern media environment, maintaining a proper relationship between the dissemination of national vocal music singing art and the ' native ' cultural roots and maintaining unique characteristics in emotional expression, artistic technology, and other aspects is a problem facing the inheritance of national vocal music singing art. National vocal music technology should effectively use modern media to explore new communication paths. Based on this article, the convolutional neural network in deep learning is combined to realize the recognition and propagation path optimization of emotional expression in national vocal music. Firstly, the emotional feature extraction of vocal music signals is carried out, and then the time and frequency domain features are analyzed. The different emotions of vocal music signals in national vocal music are labeled and classified through feature extraction. The signal feature is the input index, and the emotional label is the output feature. Finally, the effective identification of emotions in national vocal music singing based on a deep learning algorithm is obtained, which provides efficient national vocal music classification based on a machine learning algorithm for following different categories of listeners and realizes targeted dissemination for different users.

II. Research on Deep Learning in National Vocal Music

A. National Vocal Music Multi-Vocal Emotion Parameter Recognition Design

In the expression of national vocal music, it is necessary to present different national vocal music styles and emotions by concretizing the abstract space of national vocal music. Effectively determining the similarity of two music points is an important part of concreting the abstract space of national vocal music. Music emotion recognition has become a popular trend. Under this condition, it is necessary to study how to enhance the accuracy of music emotion recognition and provide support and help for music search [14].

Currently, the mainstream music emotion recognition classification structure is shown in Fig. 1. There are three main steps:

1) Select the emotion model (continuous emotion model and discrete emotion model).

2) Preprocessing music, extracting useful music features and information as input.

3) Input into the recognition model for emotion recognition.

The most important part is extracting music emotion features and constructing the recognition classification model. Previous studies often used single emotional features or classifiers based on traditional machine learning models. A single emotional feature often does not have unity and cannot fully express musical emotions. It needs to be re-extracted when performing different recognition tasks. Although the recognition effect has been improved, the efficiency is too low. The traditional machine learning model only in the unique music feature recognition results; strange music is not ideal, and poor generalization ability shows the main breakthrough in the second and third steps.

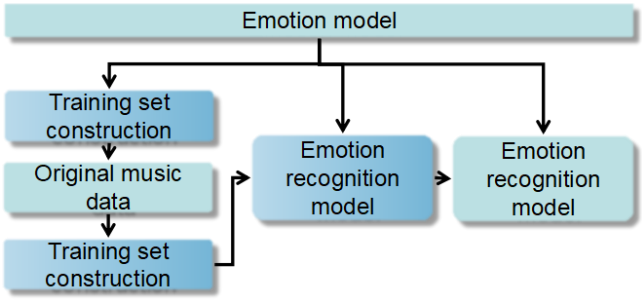


Fig. 1. Deep learning music emotion expression recognition basic structure.

Previously, music expression produced a series of music with the same style, whether relying on artificial technology or automation technology. In the abstract space of music, it isn't easy to have a variety of different styles of national vocal music at the same time. The reason for this phenomenon is that there is a certain probability of large similarity between national vocal music points with similar or different styles in the abstract space, which leads to an uneven transition between

different styles of national vocal music. Therefore, in the expression of national vocal music, the determination of the style of national vocal music creation needs to adopt some attributes as the characteristics of learning, and the style of national vocal music is numerically expressed. The objective function expresses the style value, and the national vocal music style is expressed based on the value of different objective functions.

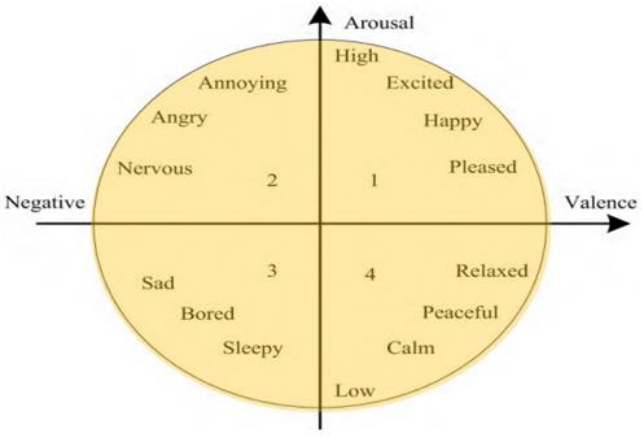


Fig. 2. National vocal music emotional space.

Different emotions are distributed in the two-dimensional space shown in Fig. 2, which is the emotional space. Based on the two-dimensional emotional space, in the expression of national vocal music, it is only necessary to project the emotional characteristics of music into the two-dimensional emotional space, take the coordinate scale value in the emotional space as the input information, and then use the artificial neural network to train the sample information. In this process, first, judge the various national vocal music elements in the training set, give the emotional category to determine its position in the emotional space, one-to-one correspondence with the emotional characteristics in the figure, obtain the emotional scope of the national vocal music, and determine the emotion and style of national vocal music creation.

The data of national vocal music is regarded as a time series, two different national vocal music sequences Y1 and Y2, their characteristic quantities are (a1μ, b1μ) and (a2μ, b2μ), and μ=1, 2,..., n represents the time length of national vocal music Y1 and Y2. At each time point μ, the Euclidean distance D (μ) between (a1μ, b1μ) and (a2μ, b2μ) is calculated, and then the similarity between the national vocal sequence Y1 and Y2 can be defined as μ = 1, 2,..., n.

$$S_{Y_1Y_2} = \frac{1}{n}\sum_{\mu=1}^{n} D(\mu) \tag{1}$$

However, the above similarity calculation process only applies to the same time length of two national vocal music sequences. For two different time lengths of national vocal music sequence, let the two-time series are X = (x1, x2,..., xi) and Y = (y1, y2,..., Yi), and set the metric function k (x1, y1), can measure the distance between component x1 and y1, namely (a1μ, b1μ) and (a2μ, b2μ) Euclidean distance.

B. A New Model of National Vocal Music Communication Under Deep Learning

1) The traditional 'live' mode of transmission of national vocal music mainly is the early context of national vocal music; people can sing works, watch, listen to activities in indoor and outdoor places, and participate in live communication activities. These modes of transmission, especially the performance of the 'stage' as a platform to sing, appreciate the more common way. At this time, the music performers and the audience of the works are in the same space-time environment for close emotional and artistic information exchange, which is a materialized, information-sharing, face-to-face communication mode. This traditional communication mode, which has lasted for many years, emphasizes that both the dissemination of song information and the acceptance of songs in the activities show the characteristics of face-to-face and the same time and space. This traditional mode of creative song transmission does not even need the assistance of other media. The instant communication and feedback of national vocal music and emotional information can be realized between the information disseminator and the song receiver, which has a strong feature of real-time information sharing.

However, there have been new changes in the current model - with live video support; people can achieve instant information exchange and communication in different times and spaces. Disseminators of songs can achieve 'off-site instant' information exchange through mobile Internet through performances anywhere. In the high development of modern mobile media, the "on-site" communication of this kind of creative song has developed into a new "on-site" communication mode relying on live video broadcast, which has become one of the most important forms of contemporary creative song communication.

2) New deep learning model in disseminating national vocal music. Deep learning in various national vocal music communication modules has been innovative from the basic information dissemination acquisition-generation distribution process. From the perspective of music information collection, a deep neural network (DNN) has many applications in music information retrieval (MIR), including beat estimation, melody mining, music emotion recognition, etc. If music is regarded as a binary data stream, Convolutional Neural Network (CNN) can efficiently mine the feature attributes of musical moments. A recurrent neural network (RNN) can analyze the characteristics of music at different times, mining its more macro features, such as emotion. QQ music, Netease cloud music, and other platforms ' listening song recognition function is the basic application of MIR. Deep learning has made more innovative attempts from the generation of music information. Nowadays, deep learning technology uses generative adversarial networks (GAN) and reinforcement learning (RL) to inject data into music and dynamically evaluate it in real-time. Deep learning can complete the evaluation of music emotion only by relying on the binary

representation of music; it can also judge the good and bad after creating and generating music and modifying it. Deep learning technology to generate music is a process of repeated creation and reflection. From the distribution of music information, the network distribution platform of music has become the world of deep learning. Music platforms rely mainly on recommendation algorithms to accurately distribute music resources to corresponding users. Deep learning represented by recommendation algorithms can process a large amount of data faster and more accurately and make more accurate recommendations for users to get a better experience.

C. Opportunities and Challenges brought by Deep Learning to National Vocal Music

1) Deep learning enables people who do not understand music theory to create music and have ownership. Then, the problem arises. If the creator creates music without knowing how it is created, it goes beyond the scope of ordinary 'music creators'. At the same time, if deep learning music creation tools create new music based on imitating existing music, then defining similarity becomes a problem. A similar piece of music, the ownership of its creative subject, is the owner of the tool, the creator of the original music, or belongs to the tool? These are issues that need careful consideration.

2) Deep learning brings great convenience to music creation, but many creations can create, far from 'can create well'. Like other arts, music creation needs to think, feel, and understand. Only in this way can it be truly appreciated and endowed with value. The music constructed by deep learning technology only proves that deep learning can generate music. This music can only be used as experimental products, which cannot make waves and even reduce the quality of the whole music.

3) Music copyright is a behavior of endowing value for art, a very important issue on the road of music commercialization. Attaching importance to copyright is to protect creators and make them have a healthy creative environment. Deep learning allows online music resources to be easily collected and quickly spread on a large scale, which increases the difficulty of tracing the path of music transmission. With such speed and breadth of transmission, the traceability of music copyright and other audio and video resources, software copyright, etc., can only be out of reach. In addition, the difficulty of positioning the creative subject will lead to copyright issues. If a piece of music cannot clarify its subject attribution, it cannot discuss its proper attribution.

4) Deep learning technology originates from the deep neural network constructed by computer scientists to imitate the ' neurons ' of the human brain, which promotes the rapid development of music communication. Music's collection, generation, and differentiation are more intelligent with the participation of deep learning. Our in-depth understanding of deep learning technology will make related concepts' definitions more mature. After jumping out of the 'learning' mode, deep learning may truly perform music creation with machine rationality.

From the current technical level, artificial intelligence has achieved computational and perceptual intelligence, but cognitive intelligence is still insufficient and needs to continue developing into cognitive intelligence. To achieve breakthroughs in national vocal music, it is necessary to combine new things such as artificial intelligence, deep learning, and machine learning to achieve more efficient emotional expression and classify national music with different emotions to share with audiences of different preferences.

III. Analysis and Implementation of National Music Deep Learning System

As the main component of music, audio plays a vital role in music emotion recognition. As an important part of music information retrieval, music emotion classification based on audio has attracted more and more attention. The second-level emotional expression in national vocal music mainly refers to the singer's realization of the emotional state through deeper excavation, such as through sound, expression, and movement form [15]. The article builds the model shown in Fig. 3.

A. Network Structure

The first few layers of the CNN network structure are feature extractors that automatically get image features through supervised training; the last layer is classified and identified by the SoftMax function. The CNN network structure is shown in Fig. 4, which shares the basic framework with the traditional AlexNet. It contains eight layers; the first five layers are convolutional layers alternating with the pooling layer, and the remaining three layers are fully connected layers for classification. The input images of the CNN network are harmonic spectra and shock spectra separated by HPSS and the spectra of the original music signals. The input image size is normalized to 256 * 256, and then it is input into the first convolution filter. In the deep network structure, the first convolution layer uses 96 kernels with a size of 11*11 and a step size of 4 pixels (the distance between the receptive field centers of adjacent neurons in the same kernel mapping) to filter the input image. Next, the max pooling layer takes the output of the first convolution layer as input and filters with 96 kernels with a size of 3*3, and the response is normalized. Using these five convolutional layers, 256 feature maps of size 6*6 were finally obtained, fed to three fully connected layers containing 4096, 1000, and 10 neurons, respectively. The final identification result is the output of the last fully connected layer, as shown in Fig. 4.

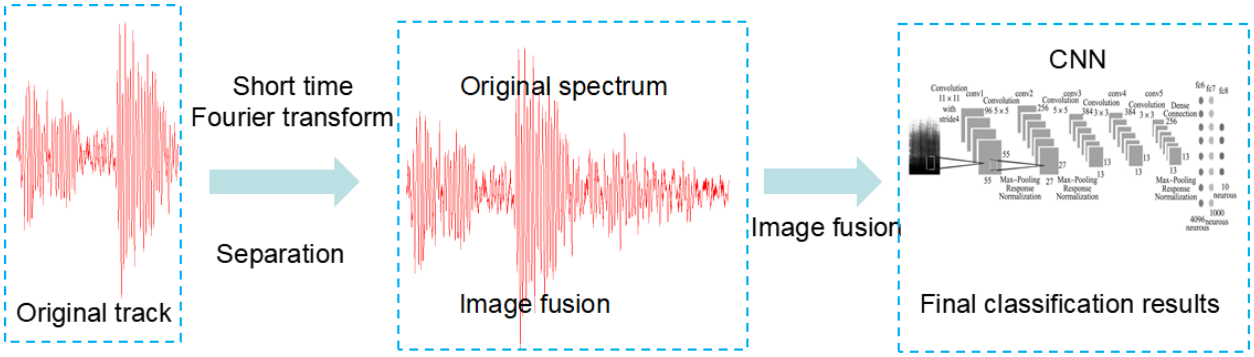


Fig. 3. CNN classification research model.

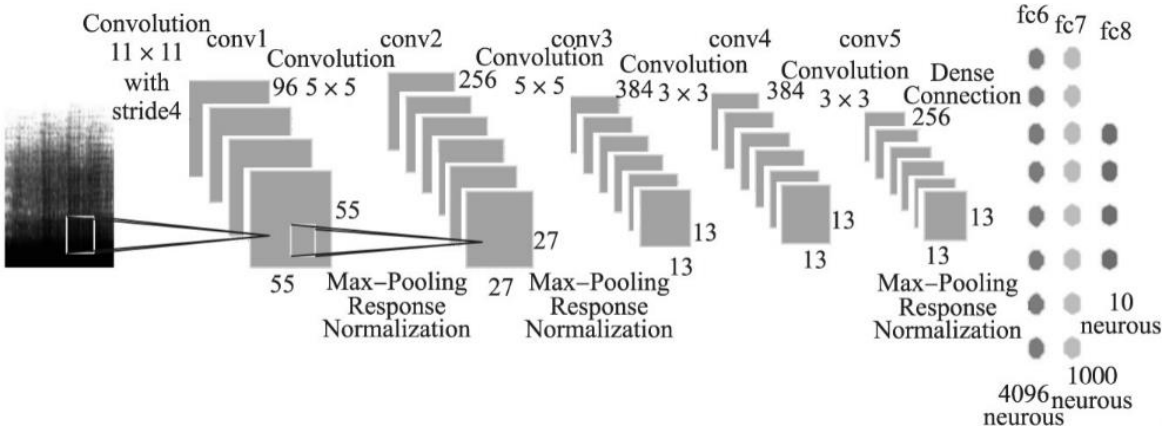


Fig. 4. CNN model.

B. Methods of Network Training and Learning

The Pooling layer uses Max Pooling, and the convolutional and pooling layers appear alternately in the convolutional neural network. Using the stochastic gradient descent (SGD) training network model, it is found that smaller weight attenuation is very important for model learning. Weight attenuation can decrease the training error in the model. Therefore, in the experiment in this paper, fine-tuning to 0.0005 usually, dropout and momentum can improve the learning effect. Since the convergence for all layers using dropout is time-consuming, the dropout value is set to η=0.5, μ= 0.9, λ = 0.0005 in the fully connected layer in the experiment of this paper. There are three fully connected layers in the network structure. The last fully connected layer, i.e., the eighth layer, is the output layer, and the output of the seventh layer is its input, which contains m neurons corresponding to m types of music styles. The output probability is p = [p1, p2,..., pm] T. The SoftMax regression formula is as follows.

$$p_j = \frac{\exp\left(X_8^j\right)}{\sum_{i=1}^{m}\exp\left(X_8^i\right)} \tag{2}$$

where, Xj 8 is the input of the SoftMax function, j is the current class being computed, j = 1,..., m ; pj represents the true output of class j.

C. Model Results and Analysis

In the experiment, the Caffe framework is utilized to train the CNN model to realize the recognition of music style. Using the recognition rate as the performance index, the music signal on the Chinese folk music collection database is the input index, and the 8 types of emotions are the output indexes. Each style category contains multiple audio recordings. The folk music emotion classification model is in Fig. 5.

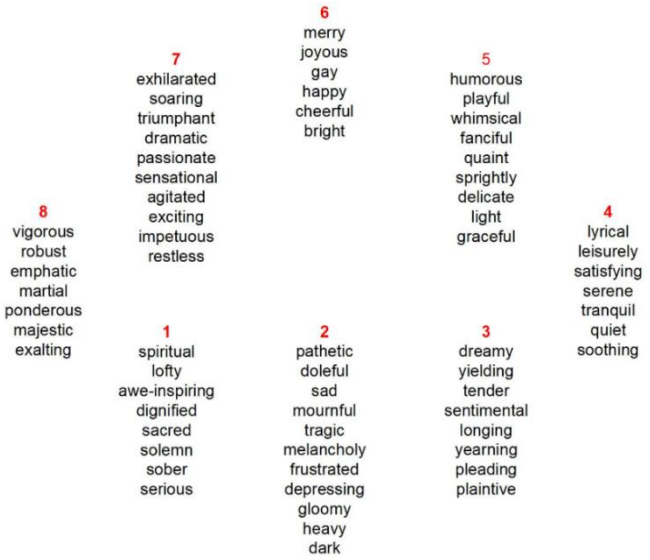


Fig. 5. Folk music emotion classification model.

TABLE I. TRAINING-RELATED HYPERPARAMETERS

Hyperparameters	η	Size	μ	λ	Dropout
Value	0.02	15	0.9	0.0004	0.45

The parameter is to be adjusted when the error rate of the training set becomes small enough and stable. The hyperparameters obtained through this adjustment process are summarized in Table I.

Relation image between iteration times and hyperparameters is shown in Fig. 6 to Fig. 9.

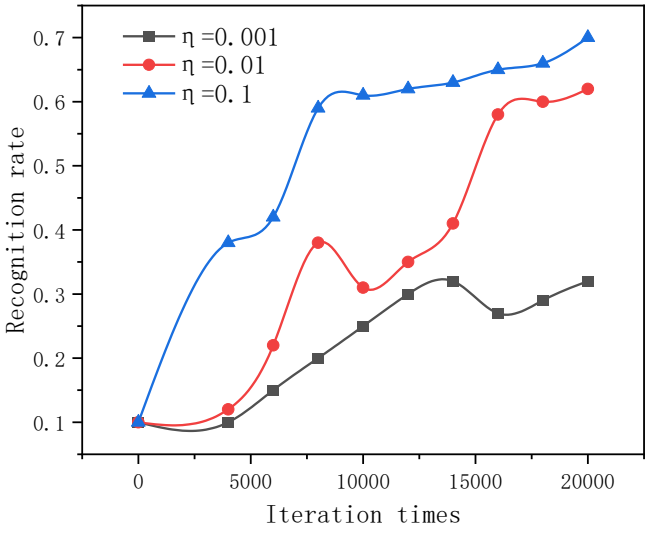


Fig. 6. Learning rate and number of iterations.

Fig. 6 shows that the learning process will be very slow, and the recognition rate is unstable when the training samples are 20000 iterations, and the learning rate η is relatively small as 0.001. Appropriately increasing η can effectively improve learning efficiency. At the same time, if η is too large as 0.1, the learning process will be unstable, and the classification performance will be decreased. Fig. 8 and Fig. 6 illustrate the influence of momentum μ and weight attenuation λ, respectively. Fig. 7 indicates that using momentum μ can accelerate the learning process well.

As shown in Fig. 9, Dropout is a technique for preventing overfitting during the training of neural networks. In the literature, dropout with a 70 % reduction in output is applied to the last fully connected layer. This experiment adopts this technique and scrambles the training data in each round to reduce overfitting. In such research, the output of hidden layer neurons is usually set to 0 with a certain probability, so such neurons will not play any role in forward and backpropagation.

In this experiment, using the hyperparameters set in Table I, the recognition rate is about 73 % without data expansion. The index to realize the correct classification of national vocal music emotion is on the diagonal of the matrix in Table II.

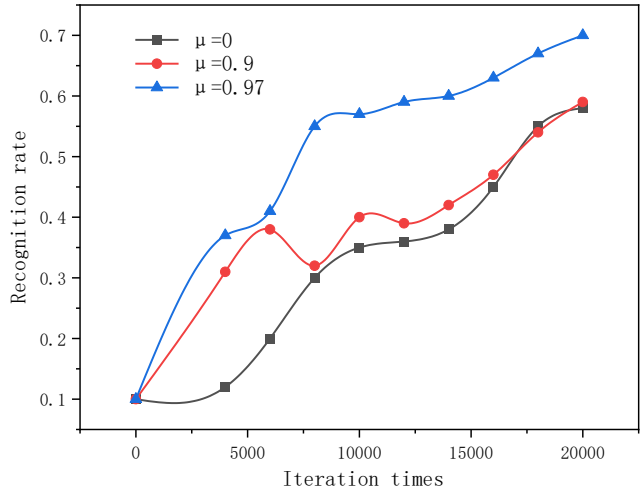


Fig. 7. Momentum coefficient and number of iterations.

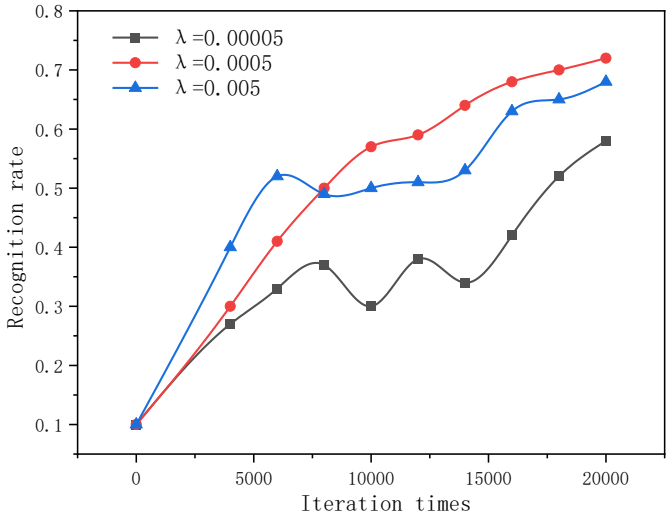


Fig. 8. Attenuation coefficient and iteration times.

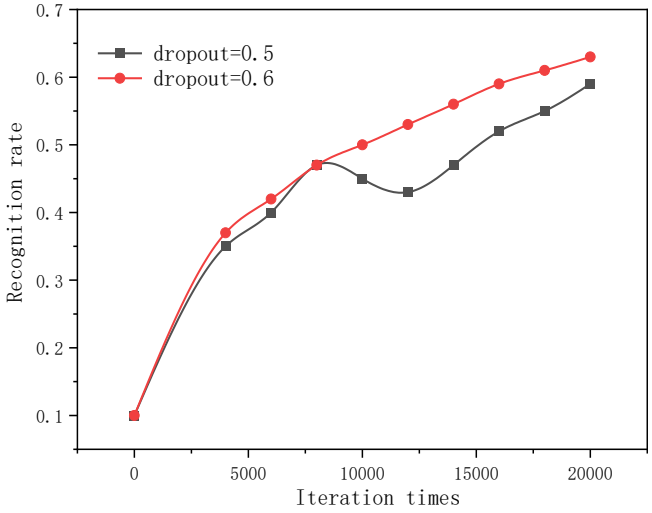


Fig. 9. Dropout value and the number of iterations.

TABLE II. CONFUSION MATRIX

Emotion categorization	1	2	3	4	5	6	7	8
1	82.5	8.3	0.0	0.0	0.0	7.3	0.0	0.0
2	10.2	74.6	0.0	0.0	0.0	12.5	0.0	0.0
3	0.0	4.5	91.7	0.0	0.0	0.0	0.0	0.0
4	4.0	4.1	4.3	79.1	0.0	0.0	0.0	8.3
5	0.0	0.0	0.0	0.0	75.6	0.0	0.0	0.0
6	3.3	8.2	0.0	0.0	0.0	76.2	0.0	0.0
7	0.0	0.0	0.0	0.0	20.2	0.0	92.0	0.0
8	0.0	0.0	0.0	12.0	0.0	4.0	0.0	82.7

D. Research on National Vocal Music Dissemination Mode of Different Emotion Types

After getting different emotional types of folk music, the emotional types of music can be classified and put on the corresponding label in the current mainstream music software. Take the dissemination of national music in microblogs as an example. After researching and analyzing the elements in the communication mechanism, it is not difficult to find that the music communication mode in microblogs is similar to the epidemic communication theory on complex networks. On this basis, by transforming the classic infectious disease communication model, a model consistent with the music communication information in microblogs is constructed to describe the communication process of music information in microblogs and a series of behavioral characteristics to facilitate the study of music communication mode in microblog in the future. Disseminating music information in Weibo, when the communicator sends out the music information, it can be seen by its friends and fans on Weibo. When the recipient receives the information, it can have one or more behaviors, such as forwarding, commenting, and liking, paying attention to private letters, and collecting the choice of music information feedback. Music knowledge, hobbies, friend intimacy, mood, and other conditions influence this choice. Similarly, the complexity of the information itself also determines its communication effect. Whether different music information has a certain value is the basis of its broadcast. The source of music information also restricts the spread of this music. The music information sent by the disseminator with high popularity and great influence is more likely to be disseminated among the audience. This music information can be more identified as meaning or value. Affected by this, the disseminator has a certain probability to choose to forward the content of the entry containing the music information. It can be regarded as an 'infected person', and the music information transmitted can be compared to the 'virus' of an infectious disease. The audience is not interested in the music information contained in this microblog and is regarded as an 'immune person'. They may hinder the praise of friends, spit in the comments or even directly ignore this information. That is the termination of music transmission. Because the music communication network in microblogs is complex and diverse, the scope of the music audience is wide, and the behavior is random and uncertain. To facilitate the research, the music communication is set in microblogs that can only spread from the music communicator to the music audience as its fans. The communicator is set as a node; lines connect the relationship among them, and the music information can only be transmitted through lines. Combined with the influence of the factors in the music communication mechanism, the classical SIR model is improved, and the music communication mode in the mechanism is modeled and studied in Fig. 10.

When the public in the micro-blog does not receive the music information, its node is S after it is 'infected' by the music information, it will be converted to the I state or IR state and the probability of converting into two states is uncertain. If it is transformed into state I, it continues to propagate along the line through the node it is concerned with and eventually transforms into state R; if it is transformed into IR state, it may need to be transformed into R state, or I state through the influence of some other information. In Weibo, music information often has certain timeliness. After some time, the heat of information will gradually decrease, so the characteristics of relay propagation will decrease. At a certain point, the IR state will also change to the R state until only two nodes, S and R, are on Weibo. In studying the music transmission mode in the music transmission mechanism in microblogs, the SIR model of music transmission is constructed by improving the SIR model of classical infectious diseases. It fully considers the rules of music transmission mode in microblogs as much as possible and discusses the influence of nodes and transmission lines. Based on the SIR model of classical infectious diseases, it can combine the dynamic principle of infectious diseases and complex networks to further study the music transmission mode in the microblog. It provides a reference basis and maybe another new direction for music transmission research.

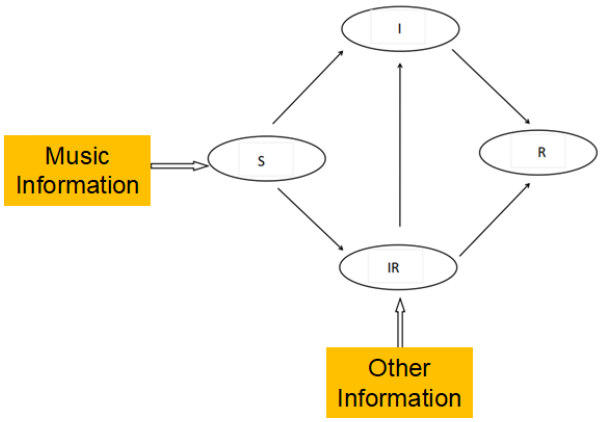


Fig. 10. National vocal music propagation model.

IV. Conclusion

Deep learning accelerates the scope and speed of information dissemination and expands the impact of various disciplines on music dissemination. Based on the innovation of deep learning in music communication, the present work analyzes the challenges brought by deep learning to music communication and looks forward to the future development of the rational application of deep learning in music communication activities. The efficient identification of emotional expression in national vocal music singing through the CNN model has been realized, and the identification accuracy is as high as 92%. With the help of microblog software in China, a new national vocal music promotion model has been designed. The article research is a new attempt at national vocal music research and has certain research value.

Music emotion recognition is an important research direction in music information retrieval. The research of music emotion involves many interdisciplinary such as music and psychology. In this paper, before the model construction, some Chinese folk music was mobile phoned, and a database was established, including eight different types of folk music audio with different emotional characteristics. Firstly, the music features are filtered and extracted, then the deep learning network model is used to process the music, and the features are used as the input of the CNN model to realize the emotional classification of Mongolian music. The final study shows that the accuracy of emotional identification of national vocal music is as high as 92 %, and the new national vocal music communication mode can achieve more efficient national music sharing. Although many researchers have conducted some research on some of these sub-areas and have achieved initial results, music emotion recognition is still in its infancy; there is still a lot of research space. The music emotion recognition work based on Mongolian music faces many problems due to limited conditions and late start.

Availability of Data and Materials

The datasets used in this paper are available from the corresponding author upon request.

Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work.

Authorship Contribution Statement

Zhangcheng Tang: Writing-Original draft preparation Conceptualization, Supervision, Project administration.

References

[1] W. Hongdan, S. SalmiJamali, C. Zhengping, S. Qiaojuan, and R. Le, “An intelligent music genre analysis using feature extraction and classification using deep learning techniques,” Computers and Electrical Engineering, vol. 100, p. 107978, 2022.

[2] D. Xiaofeng, “RETRACTED: Application of deep learning and artificial intelligence algorithm in multimedia music teaching,” Journal of Intelligent & Fuzzy Systems, vol. 38, no. 6, pp. 7241–7251, 2020.

[3] C. Kereliuk, B. L. Sturm, and J. Larsen, “Deep learning and music adversaries,” IEEE Trans Multimedia, vol. 17, no. 11, pp. 2059–2071, 2015.

[4] C. Kereliuk, B. L. Sturm, and J. Larsen, “Deep learning, audio adversaries, and music content analysis,” in 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), IEEE, 2015, pp. 1–5.

[5] J. Nam, K. Choi, J. Lee, S.-Y. Chou, and Y.-H. Yang, “Deep learning for audio-based music classification and tagging: Teaching computers to distinguish rock from bach,” IEEE Signal Process Mag, vol. 36, no. 1, pp. 41–51, 2018.

[6] S. Zhao et al., “Emotion-based end-to-end matching between image and music in valence-arousal space,” in Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 2945–2954.

[7] H. Hong and W. Luo, “The method of emotional education in music teaching,” The International Journal of Electrical Engineering & Education, p. 0020720920983559, 2021.

[8] N. Thammasan, K. Moriyama, K. Fukui, and M. Numao, “Continuous music-emotion recognition based on electroencephalogram,” IEICE Trans Inf Syst, vol. 99, no. 4, pp. 1234–1241, 2016.

[9] M. B. Er and I. B. Aydilek, “Music emotion recognition by using chroma spectrogram and deep visual features,” International Journal of Computational Intelligence Systems, vol. 12, no. 2, pp. 1622–1634, 2019.

[10] Y. Hu, “Music Emotion Research Based on Reinforcement Learning and Multimodal Information,” Journal of Mathematics, vol. 2022, pp. 1–9, 2022.

[11] N. V. Kimmatkar and B. V. Babu, “Novel approach for emotion detection and stabilizing mental state by using machine learning techniques,” Computers, vol. 10, no. 3, p. 37, 2021.

[12] S. Nag, M. Basu, S. Sanyal, A. Banerjee, and D. Ghosh, “On the application of deep learning and multifractal techniques to classify emotions and instruments using Indian Classical Music,” Physica A: Statistical Mechanics and its Applications, vol. 597, p. 127261, 2022.

[13] S. Chowdhuri, “PhonoNet: Multi-stage deep learning for raga preservation in hindustani classical music,” J Acoust Soc Am, vol. 146, no. 4, p. 2947, 2019.

[14] R. A. Wittmann-Price and M. Godshall, “Strategies to promote deep learning in clinical nursing courses,” Nurse Educ, vol. 34, no. 5, pp. 214–216, 2009.

[15] R. T. Dean and J. Forth, “Towards a deep improviser: a prototype deep learning post-tonal free music generator,” Neural Comput Appl, vol. 32, pp. 969–979, 2020.

Using Generative Adversarial Networks and Ensemble Learning for Multi-Modal Medical Image Fusion to Improve the Diagnosis of Rare Neurological Disorders

Dr.Bhargavi Peddi Reddy[1], Dr K Rangaswamy[2], Doradla Bharadwaja[3],
Mani Mohan Dupaty[4], Partha Sarkar[5], Dr. Mohammed Saleh Al Ansari[6]
Associate Professor, Dept of CSE, Associate Professor, Vasavi College of Engineering, Hyderabad, India[1]
Associate Professor, Department of Computer Science and Engineering (Data Science),
Rajeev Gandhi Memorial College of Engineering and Technology (Autonomous) Nandyal, Andhra Pradesh, India[2]
Assistant Professor, Information Technology, Prasad V Potluri Siddhartha Institute of Technology, Vijayawada, India[3]
Assistant Professor, Dept. of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur District -522502, Andhra Pradesh, India[4]
Department of Electronics and Communication Engineering, National Institute of Technology,
Mahatma Gandhi Road, Durgapur, West Bengal, India[5]
Associate Professor, College of Engineering-Department of Chemical Engineering, University of Bahrain, Bahrain[6]

Abstract*—The research suggests a unique ensemble learning approach for precise feature extraction and feature fusion from multi-modal medical pictures, which may be applied to the diagnosis of uncommon neurological illnesses. The proposed method makes use of the combined characteristics of Convolutional Neural Networks and Generative Adversarial Networks (CNN-GAN) to improve diagnostic accuracy and enable early identification. In order to do this, a diverse dataset of multi-modal patient medical records with rare neurological disorders was gathered. The multi-modal pictures are successfully combined using a GAN-based image-to-image translation technique to produce fake images that effectively gather crucial clinical data from different paradigms. To extract features from extensive clinical imaging databases, the research employs trained models using transfer learning approaches with CNN frameworks designed specifically for analyzing medical images. By compiling unique traits from each modality, a thorough grasp of the core pathophysiology is produced. By combining the strengths of several CNN algorithms using ensemble learning techniques including voting by majority, weight averaging, and layering, the forecasts were also integrated to arrive at the final diagnosis. In addition, the ensemble approach enhances the robustness and reliability of the assessment algorithm, resulting in increased effectiveness in identifying unusual neurological conditions. The analysis of the collected data shows that the proposed technique outperforms single-modal designs, demonstrating the importance of multi-modal fusion of pictures and feature extraction. The proposed method significantly outperforms existing methods, achieving an accuracy of 99.99%, as opposed to 85.69% for XGBoost and 96.12% for LSTM. The proposed method significantly outperforms existing methods, achieving an average increase in accuracy of approximately 13.3%. The proposed method was implemented using Python software.

Keywords*—Multi-modal medical images; ensemble learning; CNN; GAN; neurological disorders; image-to-image method; transfer learning; feature extraction

I. INTRODUCTION

A neurological condition called Alzheimer's disease (AD) causes diminished cognitive abilities as well as memory and mobility problems. As civilization gets older, this illness affects a growing number of elderly people. According to research, poorer nations have a significantly greater incidence of AD than advanced economies do [1]. MCI gradually develops into AD as the disorder progresses, and early mild cognitive impairment (EMCI) and late mild cognitive impairment (LMCI) are transitional states among healthy normal persons and those with Alzheimer's. Therefore, it is crucial to understand the best way to properly diagnose MCI and AD. The prevalence of brain illnesses has increased recently all across the world. One of the world's most prevalent neurological conditions is Alzheimer's disease (AD) that primarily manifests clinically as diminished memory and loss of mental abilities, along with difficulties with language and abnormalities of movement. AD is currently the fifth most common cause fatalities in the United States. The Alzheimer's Disease Association of USA published an article in 2018. The information provided by the Center for Health Statistics on the rate of change in death from a variety of hazardous illnesses in the US [2]. The number of fatalities from various risk conditions has increased. Moreover, 123% rise in AD prevalence has been reported. Another study found that in 2050, there will be around one million new instances of dementia caused by Alzheimer's, with a fresh case being identified every 33 seconds. One of the main illnesses endangering the well-being of elderly people and having an impact on social sustainability is AD. Presently, only a few medications have proven effective for the medical management

of AD, and no medication has been proven shown to slow or stop the progression of the disease.

Brain tumors currently have the highest early cost per the patient among malignancies. Tumors can develop in humans of any age due to the enormous cell growth in some areas of the cerebral cortex. Brain lesions are caused by uncontrolled growth of cerebral or central spine tissue that can impair cognitive function [3]. These enormous cells of tumors can be divided into two distinct groups: malignant and non-malignant lymphocytes depending on the region, magnitude, and location. The primary and subsequent tumor sites are the intense regions of cancerous cells. The initial tumor region is defined as the first, harmless stages of cancerous cells. According to similar research, numerous medical CBIR approaches have recently been introduced. Most advanced CBIR recovery techniques utilize just one kind of illumination [4]. One method to find the necessary healthcare images over huge collections of images is to employ resemblance contrast, and extraction techniques can allow the user to select the image category first. A CBIR algorithm may gain a lot from effective picture classification because it could eliminate the need to look over pointless images, cutting down on the total amount of images the software must look through.

Non-invasive cerebral imaging techniques have become popular in the past few decades for diagnosing AD. According to these benefits, this method of imaging is both extremely efficient and benign to human cerebral regions [5]. With regard to these benefits, doctors frequently select non-invasive clinical imaging of the brain as one of the most significant medical assessment tools. The images from multiple modes can emphasize various topographical features and geographical aspects of the brain due to the various image philosophies and techniques. Researchers are able to categorize and recognize sufferers more accurately via acquiring these traits, enabling early identification and treatment of diseases. Although modern deep learning techniques are efficient in assessing clinical images of disease progression in AD. In earlier times, scientists have employed spectral evaluation and time-frequency distribution methods like the discrete wavelet transformation method (DWT) and the Fourier transformation method (FT). It is difficult to establish a generic approach for studying different feelings, nevertheless, considering how complicated and personal the mental state is. Frequency component knowledge is insufficient for classifying human feelings for non-stationary EEG signals since it changes over time [6]. Consequently, a continuous wavelet transform (CWT) is yet another method utilized to obtain the complete understanding of frequency of signals in the temporal and spatial arena.

Nowadays, numerous neuroimaging techniques offer different kinds of data. A single method may not always yield enough data to pinpoint the distinguishing characteristic needed to locate AD in a patient. This makes it challenging for therapists to detect AD in its infancy by looking solely at the evidence from a single therapy [7]. Brain structural inequality, neurochemical, and behavioral investigations have investigated and studied a number of biomarkers and provided a variety of information. To enhance the diagnostic efficiency of a computer-aided diagnosis (CAD) framework, it is crucial to incorporate the complementing elements from several modalities. Medical image fusion integrates multiple images collected with different methods to improve the image's quality while maintaining the minute details of a single image. The method of fusion enhances visual data and simplicity, which aids in the diagnosis and evaluation of the condition by physicians [8]. Due to the rapid advancement of high-tech and modern devices, diagnostic imaging has become an integral component of a vast array of applications, including evaluation, investigation, and treatment.

Data that covers multiple kinds and situations is referred to as multimodal information. The main goal of techniques employed for combining multimodal information is to combine the data with characteristics of various dimensions and dispersion into a worldwide space of features so that the information can be expressed more accurately [9]. When performing tasks like categorization and estimation, this homogeneity can be utilized. For instance, data from significant bio banks like the UK Biobank, the Million Veterans Program, and the National Institutes of Health All of Us effort include specific to patient's genome data, diagnostic investigations, and behavioral information from electronic medical records and surveys. Consequently, it is crucial to effectively mix all of the fusion procedures [10]. With minimal computing difficulty, an effective multi-modal fusion technique ought to maximize collaboration among different modes. The Siamese networks have demonstrated outstanding efficiency in a variety of programs, including facial and recognition of handwriting. Furthermore, people transitioning to Alzheimer's disease are identified utilizing an array of recurrent neural network. Although their method is intriguing, it lacks a joint education component that does not calculate ROI-based statistics, which can be inaccurate and cause detection mistakes. To overcome these issues the research produces an ensemble learning approach for multi-modal medical image fusion and feature extraction using convolutional neural networks and generative adversarial networks in diagnosing rare neurological disorders.

The key contributions of the research are given as follows:

- The goal of data collection is to gather multi-modal diagnostic information for neurological illnesses using an ensemble learning approach built on the CNN-GAN technique.
- In order to assure data quality, a thorough data cleaning procedure is used to remove artifacts, noise, and superfluous data.
- The proposed research employs Convolutional Neural Networks, which are recognized for their capacity to automatically collect pertinent organizational traits, are used for feature extraction.
- This work utilizes Generative Adversarial Networks (GAN), which trains on a wider variety of samples to prevent overfitting, are used for data augmentation and picture fusion. In this framework, the discriminator separates created fused pictures from actual images, while the generator aims to create better fused images

that can trick the discriminator and promote adversarial learning.

- When diagnosing unusual neurological illnesses, ensemble learning is used to train several classification models utilizing chosen features, improving accuracy, dependability, and generality.
- To further improve diagnostic accuracy, particularly for illnesses with a low prevalence, ensemble learning incorporates anticipated results from many simulations.

The remaining sections of the research are given as follows. Literature Review is given in Section II. Problem Statement is discussed in Section III. Then Section IV that discusses about the proposed method for diagnosing neurological disorders. Results and Discussions are discussed in Section V. Conclusion is discussed in Section VI.

II. Related Works

Khan et al. [11] discussed in the paper that the body's center of control is the cerebral cortex. As years goes on, more and more brand-new brain disorders have been identified. As a result of the variety of neurological conditions, current diagnosing or detection methods are growing difficult and remain an unsolved scientific issue. Early diagnosis of brain disorders can have a significant impact on efforts to treat them. Artificial intelligence (AI) has been increasingly prevalent in the past few decades, altering virtually every aspect of research, including neuroscience. The deployment of AI in clinical studies has improved the accuracy and precision of neurological conditions diagnosis and treatment. In this research, researchers evaluate current advances in machine learning and deep learning for the detection of four distinct brain disorders, including Parkinson's illness, seizures, brain tumors, and Alzheimer's disease.

Brain disorders are mostly brought on by aberrant brain cell proliferation, which can harm the neural network of the central nervous system and finally result in aggressive cancer of the brain said by Musallam, Sherif, and Hussein [12]. Significant challenges exist when utilizing a Computer-Aided Diagnosis (CAD) technology to make an accurate diagnosis that would allow for effective medication, particularly when it comes to accurately identifying various illnesses in magnetic resonance imaging (MRI) images. In this research, a novel Deep Convolutional Neural Network (DCNN) design for effective identification of tumors such as meningioma, and pituitary tumors is put forward with a three-step pre-processing method to improve the presentation of MRI images. For quick instruction with a greater rate of comprehension and simple startup of the component measurements, the structure leverages sequential normalization. The method of detecting brain deviations, which is crucial for determining the extent to which they are present in MRI scans, is regarded as the biggest downside.

Hashem et al. [13] exposed in the paper that in terms of anatomy, the palate and face house 30–40% of the human body's motor and sensory neurons, showing the intimate link between the cavity in the mouth and the nervous system in general. The duties of an oral surgeon are directly related to the detection of orofacial symptoms of neurological conditions. In order to effectively recognize, diagnose, and make the right choices while managing these related neurologic disorders, dental practitioners must become acquainted with these specific presentations. These symptoms should be thoroughly evaluated with cutting-edge methodologies because it's crucial to spot any associated neurological conditions before they have major repercussions. Additionally, much-planned and successful innovative techniques are required for all types of rehabilitation therapies as well as dental treatment for individuals with neurological conditions.

A comparatively recent innovation called the Internet of Medical Things (IoMT) enables the transmission of health information across a safe network of wearables and healthcare sensors. The disadvantage of this research is that IoMT for dental care concentrates mostly on proactive maintenance methods by identifying the root of tooth decay at the earliest opportunity and disseminating information with the oral surgeon and the consumer round-the-clock. The precise identification of Alzheimer's disease (AD) has been made possible by the integration of multi-modal data, such as magnetic resonance imaging (MRI) and positron emission tomography (PET), which both provide complimentary functional and structural data said by Ning et al. [14]. While their fundamental connections might offer additional distinguishing traits for AD detection, the majority of the present approaches merely combine multi-modal characteristics in the initial location. The issue of the overfitting problem brought on by highly dimensional multimodal information that continues to be intriguing. In order to do this, researchers suggest a relation-induced multidimensional shared representational method of learning for identifying AD. The suggested approach combines reduction of dimension, classification simulation, and representational training into a single architecture.

Rathore et al. [15] discussed in the paper that classifying various patterns of attack for profound implants in the brain is the goal of the deep learning approach. The restricted accessibility of labelled information and the substantial patient-to-patient variation in cerebral implantation impulses pose challenges, though. It is crucial to guarantee the model's accuracy and comprehension for healthcare providers and patients as well as its durability over prospective hostile assaults. In order to allow efficient operation on cerebral implantation gadgets, immediate processing and hardware constraints have to be conquered. Issues over patients' privacy and confidentiality, as well as security concerns, must be carefully taken into account. In order to provide an efficient and safe deep learning system for categorizing assault behaviors in profound cerebral devices, it is going to be essential to overcome these challenges.

Around one percent of the global population suffers from epilepsy, a persistent neurological illness that is defined by an increased frequency of uncontrolled convulsions. The majority of the present epilepsy detection techniques depend heavily on historical patient information, which makes them ineffective when trying to identify fresh patients in a patient-independent environment Zhang et al. [16]. To get across this issue, researchers offer a successful framework for detecting seizures

caused by epilepsy that improves from epileptic events while reducing inter-patient interference. To identify the original seizure-specific information from the unprocessed non-invasive electroencephalography (EEG) recordings via adversarial learning, a sophisticated deep neural network framework is suggested.

The research examines the significance of accurate diagnosis and early treatment of a variety of brain illnesses and emphasizes the contribution of AI and deep learning to increasing diagnostic precision. Brain tumor identification or facial signs of neurological diseases, IoT in dental care, deep learning for brain implant security, and enhanced epilepsy diagnosis are some of the subjects discussed. The study on utilizing technology to improve our comprehension and treatment of neurological diseases is expanding, and these studies add to that body of knowledge.

III. Problem Statement

From the above discussion the literature reveals that the Siamese networks have demonstrated outstanding efficiency in a variety of programs, including facial recognition [17]. The current approach of detecting uncommon neurological illnesses is promising, but it has flaws in terms of accuracy and robustness, largely because it lacks a collaborative education component that includes trustworthy indicators other than ROI-based data. ROI-based data may be inaccurate and might result in missed diagnoses of neurological illnesses that are infrequent. This research suggests an ensemble learning strategy that incorporates multi-modal medical picture fusion and feature extraction approaches to solve these problems and improve the accuracy and reliability of diagnosis. This method uses generative adversarial networks (GANs) and convolutional neural networks (CNNs) to offer a more thorough and efficient diagnosis for uncommon neurological illnesses.

IV. Proposed Methodology

The approach employed for the research concentrates on creating an innovative technique to identify rare neurological disorders by utilizing the combined strength of Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) for multi-modal medical imaging. The Harvard Atlas Brain dataset is utilized for the process of testing and training. This dataset contains the modality of MRI brain images. This is secondhand to validate the efficiency of the investigation and to diagnose rare neurological disorders. Then, the gathered data are cleaned and any undesirable artifacts, noise, or unimportant data are removed by employing preprocessing procedure. Then, the ensemble learning based on CNN-GAN is utilized for the process of diagnosing neurological disorder. Moreover, it is hired in demand to upsurge accuracy. The recommended CNN-GAN outline's architectural illustration is publicized in Fig. 1.

A. Data Collection

The process of collecting knowledge, data or evaluations from multiple sources to use in investigation, evaluation, decision-making or any other type of specified objective is known as data collection [18]. The Harvard Atlas Brain dataset is utilized for the process of testing and training. This dataset contains the modality of MRI brain images. This is secondhand to validate the efficiency of the investigation and to diagnose rare neurological disorders.

B. Pre-processing

In data analysis and machine learning, pre-processing is a critical phase where unprocessed data is cleaned, converted and structured to render it appropriate for future analysis or model training [19]. Preprocessing aims to enhance the integrity of the data eliminate noise and discrepancies and guarantee that the information arrives in a manner that can be utilized successfully for its intended purpose.

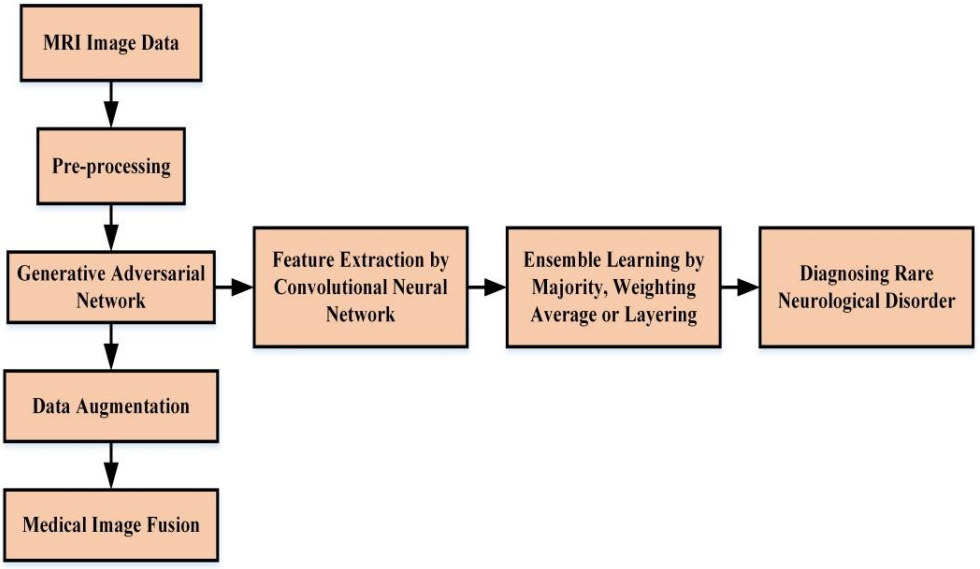


Fig. 1. Proposed ensemble learning based CNN-GAN method.

C. Generative Adversarial Network (GAN)

A machine learning model called a generative adversarial network (GAN) consists of two distinct neural networks such as a discriminator and a generator. Although the system for discrimination attempts to separate genuine data from produced data, the generator creates samples of artificial data. The generator seeks to provide convincing samples to deceive the discriminator, and the discriminator attempts to increase its capacity to distinguish between actual and false information. They are trained in an adversarial manner. GANs can produce high-quality and different information samples via this competitive approach, which makes them valuable for projects like image synthesizing, data enhancement, and style transfer. In Fig. 2, the diagrammatic representation of GAN is depicted.

GANs are employed in multi-modal fusion for the diagnosis of neurological disorders for the reason they are efficient at capturing complicated trends and variations from diverse sources of data, integrating information from multiple sources, performing data enhancement, enabling transfer learning, and aiding in the creation of accurate images or depictions of neurological conditions. By utilizing these skills, GANs improve illness diagnosis' precision, durability, and generalization. The equation for generator and discriminator is given in Eq. (1) below.

$$min_G\, max_D * L(D,G) = F_x{\sim}P_{data}(x)\big[log\big(D(x)\big)\big] + F_y{\sim}P_y(y)[log(1 - D(G(y)))] \quad (1)$$

Here, y is represented as the input of the image; D is the denoted as the Discriminator; G is represented as the generator; F is denoted the expectation operator; x is represented as the real samples. This research's efficacy is considerably increased by including generative adversarial networks by tackling major difficulties in identifying uncommon neurological illnesses. When working with little or unbalanced datasets, as is frequently the case with uncommon conditions, GANs excel at producing synthetic data that closely resembles actual medical pictures. In order to ensure a more varied and representative set of medical pictures for training, GANs can help supplement the dataset, which will ultimately result in a more robust model.

1) Data augmentation by GAN: It is a collection of methods for changing duplicates of already-existing data or creating fresh copies of the database intentionally utilizing the data already present [20]. It serves as normalization during machine learning model training and lessens over fitting.

2) Medical image fusion using GAN: Various medical images possess certain distinctive qualities that call for concurrent observation for clinical diagnosis [21]. Therefore, multi-modality image fusion is used to merge the characteristics of several image detectors into a single image. A cutting-edge method called medical image fusion utilizes GAN to combine data from many healthcare imaging techniques into a single, improved representation. The GAN develops to produce combined images that capture crucial elements from each modality by being trained on pairs of medical images from multiple sources. To provide realistic results with an adversarial loss and to protect important anatomical components with a content loss, this technique improves a loss function. The created composite image can be very helpful in enhancing medical analysis, helping to diagnose diseases, organize treatments, and track patients. Utilizing GANs for medical image fusion has the ability to deliver more and thorough educational data, thereby enabling medical practitioners to make better choices for the welfare of their patients. Before implementing an AI-driven strategy into the healthcare sector a thorough validation and compliance to legal requirements are essential.

D. Feature Extraction by Convolutional Neural Network (CNN)

A key step in deep learning for pattern and image recognition applications is feature extraction and here it is done by Convolutional Neural Networks (CNN). CNNs were created in order to automatically pick up organizational and discriminating characteristics from the raw input images. Small filters are convolved throughout the input image in a sequence of convolutional layers in order to gather regional patterns and characteristics. The boundaries, surfaces, and contours that make up the graphical world are captured by these features [22]. The feature maps are then down sampled by pooling layers, which reduces the computational burden while preserving crucial information. Fully connected layers receive the outcomes of the characteristics learned and use them for categorization or similar tasks in the future. With contemporary effectiveness in challenges like recognition of objects, segmenting an image, and image analysis for medicine, CNNs have achieved impressive results in a variety of applications that use computer vision. This is attributable to their capacity to spontaneously acquire and retain significant characteristics from images. The CNN-based multi-modal fusion technique harnesses the advantages of each method, making up for the limits associated with particular techniques and offering a more thorough understanding of the neurological state by combining data from various modalities. This comprehensive representation improves the model's capacity to provide more accurate and reliable diagnoses, allowing physicians to learn more about the condition of the patient and formulate effective therapies. The diagrammatic representation of CNN is depicted in Fig. 3.

E. Ensemble Learning

A machine learning technique known as ensemble learning integrates the projections of various designs, known as base learners and it is utilized to produce an end result that is more accurate and trustworthy. Ensemble learning tries to enhance efficiency, generalization, and dependability by utilizing the variety of distinct models [23]. The base learners may consist of multiple algorithms or modifications of the exact same algorithm that have been trained on various data groups. The ensemble aggregates its forecasts via techniques like vote by majority, weighted average or layering. Ensemble learning is utilized extensively over multiple fields, such as regression, classification, detection of anomalies and has been demonstrated to be highly efficient in solving complicated issues and producing better final results comparing to individual methods.

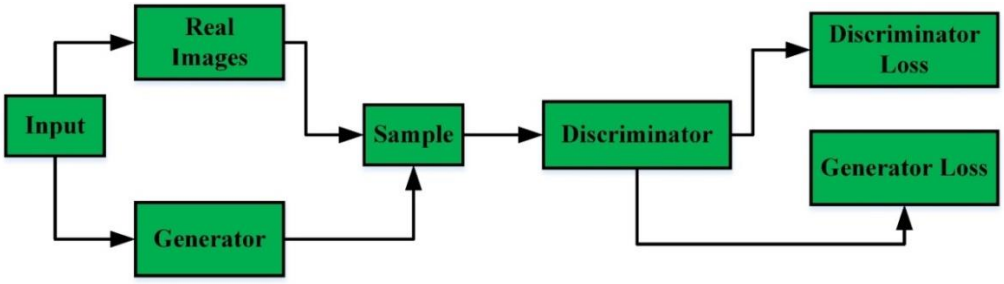


Fig. 2. Generative adversarial network.

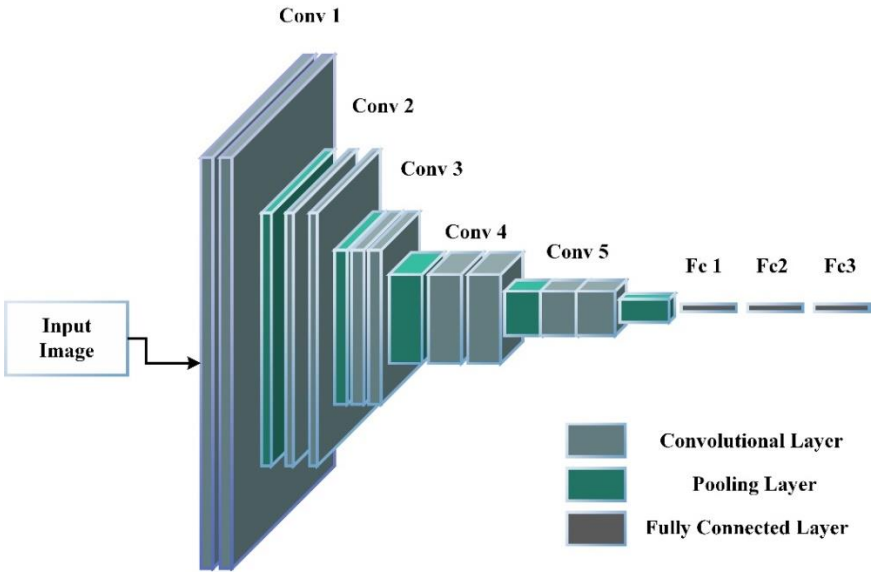


Fig. 3. Convolutional neural network.

1) Vote of majority: A straightforward and well-liked technique for merging the projections of base learners in the classification issue is the vote of majority. The classification for an input provided is predicted by every base learner in the ensemble and the final guess is the collective label that all base learners agree on. In the event of a tie, the category label having the greatest level of probability or likelihood will win. Although the base learners possess a low correlation and prone to engage in various errors, the vote of majority works well because it lessens the effect on particular errors.

2) Weighted average: In problems with regression where the base learners estimate constant outcomes instead of binary class labels, weighted average is frequently utilized. The ultimate prediction is derived by adding the weighted predictions after multiplying each base learner's forecast by the weight. The weights may be chosen independently or by using methods like the cross-validation or minimization.

3) Layering: The process of layering sometimes referred to as stacking, is a sophisticated ensemble learning technique that entails teaching a meta-model, or stacker, to draw knowledge from the forecasts of numerous base learners. Utilizing the provided data as a starting point, the base learners produce guesses; these hypotheses then serve as new characteristics for the meta-model. After learning via these preliminary estimates, the meta-model can subsequently anticipate the outcome. The utilization of stacking enables the ensemble to make the most of the abilities of various base learners and may enhance effectiveness. The algorithm of the proposed ensemble learning is given below and followed by that Fig. 4 depicts the flowchart of the proposed method.

Algorithm 1: Proposed CNN-GAN

Input: Multi Modal MRI Scan Image
Output: Diagnosis of Rare Neurological Disorder
Load data for provided image //Data Collection
Cleaning of data, Elimination of noise //Pre-processing
Data Augmentation, Medical Image Fusion //GAN
for G training L **do**
for D training N **do**
Select r patches from the testing data and generated data
G and D equation is given in (1)
Update Discriminator
Select r patches from the testing data
Update Generator
end for
Feature Extraction //CNN
Vote of majority, weighted average and layering //Ensemble Learning

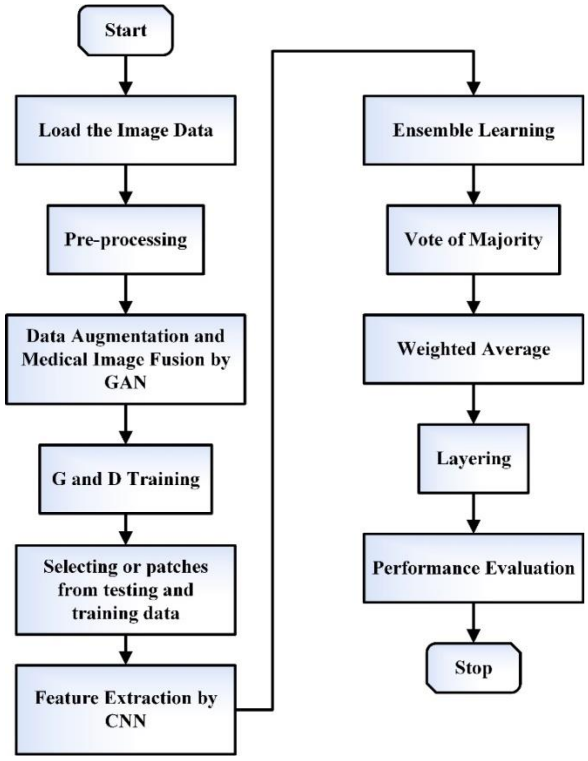


Fig. 4. Flowchart of proposed ensemble learning based CNN-GAN method.

V. Results and Discussions

The recommended method has been inspected by means of some datasets. Here, Ensemble Learning based Convolutional Neural Network and Generative Adversarial Network framework (CNN-GAN) is secondhand in this investigation for the diagnosis of multi modal fusion neurological disorders. Python is the programming language that was employed to construct and carry out the computational methods described in the suggested technique, which were executed utilizing Python software. The description of the recommended technique is reflected by certain features such as Accuracy, Recall, Precision, F1 Score, ROC, Training Accuracy and Training Loss.

A. Performance Metrics

1) Accuracy: The statistic that is the simplest to understand is accuracy, which is expressed as a proportion of all instances that occurred when a data set has been correctly classified. It provides a comprehensive indication of overall correctness. The accuracy calculation is offered in Eq. (2) below.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2)$$

2) Precision: Precision is the ratio of exactly predicted favorable outcomes to all other instances. The precision equation is offered in Eq. (3) below,

$$Precision = \frac{TP}{(TP+FP)} \quad (3)$$

3) Recall: Remember to compute the proportion of correctly anticipated beneficial actions that actually materialized amongst all favorable situations. It increases up how well the algorithm can categorize every favorable circumstance. In Eq. (4), the recall formula is presented.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

4) F1 Score: Precision and recall are effectively added to create the F1 score. It combines both dimensions to provide a single number that provides a precise evaluation of a representation's effectiveness. In Eq. (5), the F1-Score equation is distributed.

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision+Recall)} \quad (5)$$

In Table I and Fig. 5 it shows the values of Accuracy, Precision, Recall and F1 Score of the existing i) XGBoost method [24] ii) LSTM method [25] and the proposed method produces more accuracy of about 99.99%, precision of about 99.13%, recall of about 98.21% and f1 score of about 98.99% than the existing methods.

TABLE I. Comparison Table of Accuracy, Precision, Recall and F1 Score

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
XGBoost	85.69	57.62	63.21	60.22
LSTM	96.12	97.87	90.32	92.96
Proposed Method	99.99	99.13	98.21	98.99

B. Training Accuracy and Training Loss

The proportion of occurrences in the training dataset that were properly anticipated is known as training accuracy, and it shows the extent to which the model works on the data it was developed on. High training precision alone is not enough to guarantee effective generalization to new data because over fitting could lead to poor generalization. The difference among the forecasts made by the model and the actual goals achieved over training is quantified as training loss. By changing the parameters of the model, the goal is to increase the effectiveness of the model on the training data while minimizing the training loss. Considering optimization excessively for training data it can result in inadequate results on new, unforeseen data, finding a balance between low training loss and high generalization is a significant difficulty in machine learning. For the algorithm to produce precise forecasts on new information and to gauge its learning progress, it is critical to track both measures. Fig. 6 depicts the diagram of Accuracy vs. Loss graph.

C. Area under the ROC Curve (AUC-ROC)

A binary classification model's efficacy can be evaluated graphically utilizing the Receiver Operating Characteristic (ROC) curve. For the algorithm's predictions, it displays the true positive rate (sensitivity) versus the false positive rate (1-specificity) across different threshold values. The ROC curve enables users to see the trade-off among the model's tendency to attribute negative examples wrongly and its capability of correctly recognizing positive instances.

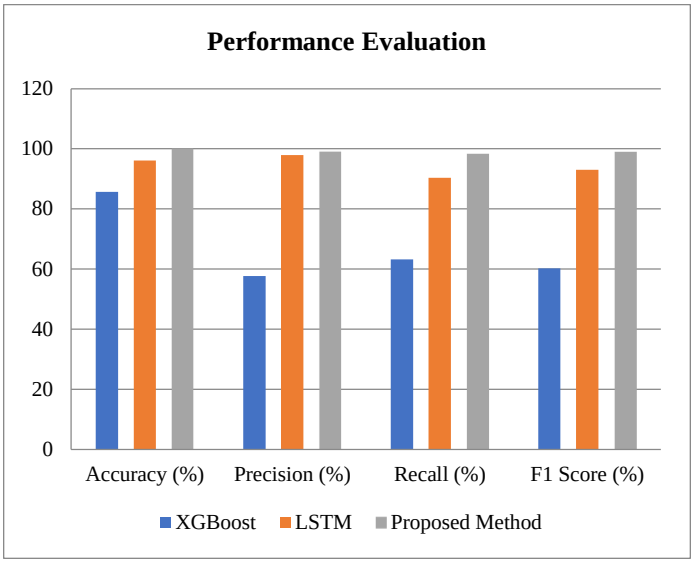


Fig. 5. Comparison graph of accuracy, precision, recall and f1 score.

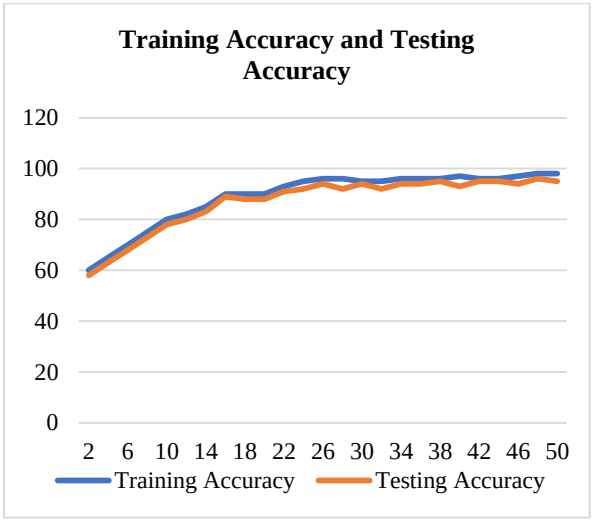


Fig. 6. Accuracy vs loss.

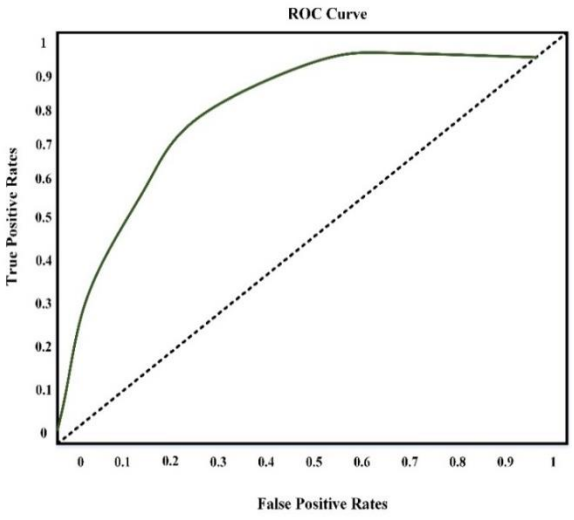


Fig. 7. AUC-ROC.

An ROC curve that crosses the top-left corner of the plot, which denotes a high degree of sensitivity with a low false positive rate, represents the ideal classifier. An AUC-ROC number below 1 implies high classification abilities, while a value near to 0.5 denotes random or poor performance. The area under the ROC curve (AUC-ROC) is an individual scalar statistic calculated from the curve that provides an overview of the model's general efficacy. ROC analysis is utilized frequently in machine learning to evaluate and contrast the performance of classifiers in a variety of programs, including detecting fraud, identifying anomalies, and medical diagnosis and its diagram is depicted in Fig. 7.

D. Discussions

The cumulative corpus of research highlighted highlights the changing face of neurological healthcare, with a focus on the amalgamation of cutting-edge technologies and artificial intelligence (AI). Khan et al. address the growing difficulty of diagnosing and treating an increasing range of brain illnesses,

arguing for the central function of the cerebral cortex as the body's centre of control. The potential transformative influence on diagnostic and treatment precision is highlighted by their investigation of AI applications, particularly in machine learning and deep learning, for the early detection of illnesses like Parkinson's, seizures, brain tumours, and Alzheimer's. Musallam, Sherif, and Hussein [12] offer insightful observations about the connection between neurological illnesses and abnormal brain cell growth. They also suggest a unique Deep Convolutional Neural Network (DCNN) that can be used with advanced pre-processing techniques to improve tumour identification. On the other hand, there is still a significant drawback in the proper diagnosis of many disorders from magnetic resonance imaging (MRI) images. The investigation by Hashem et al. [13] of the relationship between the neural system and the oral cavity emphasizes the importance of oral surgeons in identifying orofacial symptoms of neurological disorders. Although the significance of state-of-the-art techniques for comprehensive assessment is emphasized, the restriction is the narrow emphasis on preventive dental care approaches, which may obscure more general dental health issues. Although novel, the use of the Internet of Medical Things (IoMT) to dentistry is criticized for its restricted focus on preventative maintenance, especially when it comes to detecting tooth decay. Ning et al.[14] 's proposal to integrate multi-modal data for Alzheimer's disease detection tackles the issue of overfitting in highly dimensional data. Deep learning for the purpose of classifying attack patterns in cerebral implants is being investigated by Rathore et al. [15] their work highlights issues with labelled information accessibility, patient-to-patient variability, and the critical need for accuracy, comprehension, and security considerations, including patient privacy. With an emphasis on epilepsy detection, Zhang et al.[16] expose the shortcomings of methods that rely on patient history data and present a deep neural network architecture that is highly intelligent and can learn from epileptic episodes to improve seizure diagnosis. As a whole, these studies highlight the exciting possibilities that artificial intelligence (AI) and technical advancements hold for improving neurological healthcare. However, they also highlight ongoing obstacles that require additional study and improvement before being put into practice.

A novel ensemble learning approach for precise feature extraction and feature fusion from multi-modal medical images to identify uncommon neurological illnesses is suggested in this research. Convolutional neural networks (CNNs) and generative adversarial networks (GANs) are utilized in the suggested strategy in a CNN-GAN combined strategy, with the goal of improving the diagnostic precision and enabling rapid diagnosis. In a broad collection of multi-modal health records, crucial clinical data is effectively combined from multiple approaches by constructing artificial images via GAN-based translation. The assessment of the method's robustness and reliability are increased by the ensemble technique, which employs voting by majority, weighted average, and stacking to combine estimates in order to arrive at the ultimate diagnosis. As a result, it is more effective in identifying unusual neurological illnesses. The technique functions better than single-modal concepts showing the significance of multi-modal fusion of images and feature extraction, possibly offering a quick and reliable detection of unusual neurological disorders, with a combined approach resulting to about 99.99% diagnostic accuracy. To verify the generality and usefulness of this suggested strategy in actual clinical circumstances, further verification on a bigger and more varied dataset will be required.

VI. Conclusion and Future Works

Convolutional neural networks (CNNs) and generative adversarial networks (GANs) are being combined in this study to improve diagnostic accuracy and enable early detection. In order to diagnose unusual neurological illnesses, this work offers a unique ensemble learning technique that successfully mixes and segregates characteristics from multi-modal medical pictures. Critical health data is effortlessly incorporated from diverse medical records using GAN-based translation. The research extracts distinguishing characteristics from sizable clinical imaging databases by utilizing transfer learning inside specialized CNN frameworks, giving a thorough knowledge of the basic pathophysiological concepts. Voting, weight averaging, and stacking ensemble approaches all effectively include hypotheses, greatly improving the identification of uncommon neurological illnesses. With a diagnostic reliability of over 100%, this combination strategy performs better than single-modal approaches. This novel approach has a great deal of promise to transform medical image processing, possibly enabling quick and precise diagnosis, and eventually enhancing patient outcomes. But further testing and real-world applications are necessary. Future research can broaden the ensemble learning strategy to incorporate more sophisticated fusion methods and apply it to different medical specialties. Comprehensive research studies and validation across a larger and more varied patient group are necessary to evaluate the method's practical efficacy and generalizability.

References

[1] Z. Jiao, S. Chen, H. Shi, and J. Xu, “Multi-Modal Feature Selection with Feature Correlation and Feature Structure Fusion for MCI and AD Classification,” *Brain Sciences*, vol. 12, no. 1, p. 80, Jan. 2022, doi: 10.3390/brainsci12010080.

[2] X. Hao *et al.*, “Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of Alzheimer’s disease,” *Medical Image Analysis*, vol. 60, p. 101625, Feb. 2020, doi: 10.1016/j.media.2019.101625.

[3] S. Maqsood, R. Damaševičius, and R. Maskeliūnas, “Multi-Modal Brain Tumor Detection Using Deep Neural Network and Multiclass SVM,” *Medicina*, vol. 58, no. 8, p. 1090, Aug. 2022, doi: 10.3390/medicina58081090.

[4] M. H. Abid, R. Ashraf, T. Mahmood, and C. M. N. Faisal, “Multi-modal medical image classification using deep residual network and genetic algorithm,” *PLoS ONE*, vol. 18, no. 6, p. e0287786, Jun. 2023, doi: 10.1371/journal.pone.0287786.

[5] Z. Kong, M. Zhang, W. Zhu, Y. Yi, T. Wang, and B. Zhang, “Multi-modal data Alzheimer’s disease detection based on 3D convolution,” *Biomedical Signal Processing and Control*, vol. 75, p. 103565, May 2022, doi: 10.1016/j.bspc.2022.103565.

[6] M. A. Asghar *et al.*, “EEG-Based Multi-Modal Emotion Recognition using Bag of Deep Features: An Optimal Feature Selection Approach,” *Sensors*, vol. 19, no. 23, p. 5218, Nov. 2019, doi: 10.3390/s19235218.

[7] S. Sharma and P. K. Mandal, “A Comprehensive Report on Machine Learning-based Early Detection of Alzheimer’s Disease using Multi-modal Neuroimaging Data,” *ACM Comput. Surv.*, vol. 55, no. 2, pp. 1–44, Feb. 2023, doi: 10.1145/3492865.

[8] A. S. Vijendran and K. Ramasamy, "Optimal segmentation and fusion of multi-modal brain images using clustering based deep learning algorithm," *Measurement: Sensors*, vol. 27, p. 100691, Jun. 2023, doi: 10.1016/j.measen.2023.100691.

[9] S. Amal, L. Safarnejad, J. A. Omiye, I. Ghanzouri, J. H. Cabot, and E. G. Ross, "Use of Multi-Modal Data and Machine Learning to Improve Cardiovascular Disease Care," *Front. Cardiovasc. Med.*, vol. 9, p. 840262, Apr. 2022, doi: 10.3389/fcvm.2022.840262.

[10] Y. Dai, Y. Gao, and F. Liu, "TransMed: Transformers Advance Multi-Modal Medical Image Classification," *Diagnostics*, vol. 11, no. 8, p. 1384, Jul. 2021, doi: 10.3390/diagnostics11081384.

[11] P. Khan *et al.*, "Machine Learning and Deep Learning Approaches for Brain Disease Diagnosis: Principles and Recent Advances," *IEEE Access*, vol. 9, pp. 37622–37655, 2021, doi: 10.1109/ACCESS.2021.3062484.

[12] A. S. Musallam, A. S. Sherif, and M. K. Hussein, "A New Convolutional Neural Network Architecture for Automatic Detection of Brain Tumors in Magnetic Resonance Imaging Images," *IEEE Access*, vol. 10, pp. 2775–2782, 2022, doi: 10.1109/ACCESS.2022.3140289.

[13] M. Hashem, S. Vellappally, H. Fouad, M. Luqman, and A. E. Youssef, "Predicting neurological disorders linked to oral cavity manifestations using an IoMT-based optimized neural networks," *IEEE Access*, vol. 8, pp. 190722–190733, 2020.

[14] Z. Ning, Q. Xiao, Q. Feng, W. Chen, and Y. Zhang, "Relation-Induced Multi-Modal Shared Representation Learning for Alzheimer's Disease Diagnosis," *IEEE Trans. Med. Imaging*, vol. 40, no. 6, pp. 1632–1645, Jun. 2021, doi: 10.1109/TMI.2021.3063150.

[15] H. Rathore, A. K. Al-Ali, A. Mohamed, X. Du, and M. Guizani, "A novel deep learning strategy for classifying different attack patterns for deep brain implants," *IEEE Access*, vol. 7, pp. 24154–24164, 2019.

[16] X. Zhang, L. Yao, M. Dong, Z. Liu, Y. Zhang, and Y. Li, "Adversarial Representation Learning for Robust Patient-Independent Epileptic Seizure Detection." arXiv, Jan. 31, 2020. Accessed: Aug. 03, 2023. [Online]. Available: http://arxiv.org/abs/1909.10868

[17] C. Ostertag, M. Visani, T. Urruty, and M. Beurton-Aimar, "Long-term cognitive decline prediction based on multi-modal data using Multimodal3DSiameseNet: transfer learning from Alzheimer's disease to Parkinson's disease," *Int J CARS*, vol. 18, no. 5, pp. 809–818, Mar. 2023, doi: 10.1007/s11548-023-02866-6.

[18] M. Adeel Azam, K. Bahadar Khan, M. Ahmad, and M. Mazzara, "Multimodal Medical Image Registration and Fusion for Quality Enhancement," *Computers, Materials & Continua*, vol. 68, no. 1, pp. 821–840, 2021, doi: 10.32604/cmc.2021.016131.

[19] X. Bi, W. Zhou, L. Li, and Z. Xing, "Detecting Risk Gene and Pathogenic Brain Region in EMCI Using a Novel GERF Algorithm Based on Brain Imaging and Genetic Data," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 8, pp. 3019–3028, Aug. 2021, doi: 10.1109/JBHI.2021.3067798.

[20] Rejusha and Vipin Kumar, "Artificial MRI Image Generation using Deep Convolutional GAN and its Comparison with other Augmentation Methods," in *2021 International Conference on Communication, Control and Information Sciences (ICCISc)*, Idukki, India: IEEE, Jun. 2021, pp. 1–6. doi: 10.1109/ICCISc52257.2021.9484902.

[21] Y. Fu, X.-J. Wu, and T. Durrani, "Image fusion based on generative adversarial network consistent with perception," *Information Fusion*, vol. 72, pp. 110–125, 2021.

[22] Y. Liu, B. Yan, R. Zhang, K. Liu, G. Jeon, and X. Yang, "Multi-Scale Mixed Attention Network for CT and MRI Image Fusion," *Entropy*, vol. 24, no. 6, p. 843, Jun. 2022, doi: 10.3390/e24060843.

[23] L. A. W. Gemein *et al.*, "Machine-learning-based diagnostics of EEG pathology," *NeuroImage*, vol. 220, p. 117021, Oct. 2020, doi: 10.1016/j.neuroimage.2020.117021.

[24] L. Herath, D. Meedeniya, M. A. J. C. Marasingha, and V. Weerasinghe, "Autism spectrum disorder diagnosis support model using Inception V3," in *2021 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, Colombo, Sri Lanka: IEEE, Sep. 2021, pp. 1–7. doi: 10.1109/SCSE53661.2021.9568314.

[25] N. Chintalapudi, G. Battineni, M. A. Hossain, and F. Amenta, "Cascaded Deep Learning Frameworks in Contribution to the Detection of Parkinson's Disease," *Bioengineering*, vol. 9, no. 3, p. 116, Mar. 2022, doi: 10.3390/bioengineering9030116.

Creating a Framework for Care Needs Hub for Persons with Disabilities and Senior Citizens

Guillermo V. Red, Jr[1], Thelma D. Palaoag[2], Vince Angelo E. Naz[3]
Computer Studies Department, Bicol University Polangui, Albay, Philippines[1]
College of Information Technology and Computer Science, University of the Cordilleras, Baguio City, Philippines[2]
Computer Studies Department, Bicol University Polangui, Albay, Philippines[3]

***Abstract*—Patient satisfaction is an assessment that assesses how effectively a company's goods or services fulfil consumer expectations. This study aims to design an architectural framework for a care needs hub for people with disabilities and senior citizens. Using systems modelling for crafting architectural frameworks, the researchers used a 4+1 view model with UML to intensively describe the features of the care needs hub. Quality attributes were used to indicate how well the system would satisfy the needs of the stakeholders beyond its basic functions. The design includes the system's functional and non-functional features, as well as their corresponding diagrams drawn in a unified modelling language in accordance with the 4+1 view model, to assist the system's developer in mapping the system's functionalities correctly and accurately. Architecture models and design patterns are developed and executed to understand how the system's primary components fit together, how messages and data move effectively across the system, and how other structural issues work. The proposed model includes verified and validated development paradigms and architectural and design patterns that may help accelerate the development process. The architecture and design patterns fulfil all of the system's criteria. The researchers designed a comprehensive tool for the completion of the development of the care needs hub, which would greatly help the developers of the system in crafting the correct features and data abstractions needed to build and implement the said system. This research aims to develop an innovative solution that addresses the current challenges faced by persons with disabilities and senior citizens in accessing care services and provides a comprehensive and accessible platform for their care.**

Keywords*—Care need framework; persons with disability; CareAide; 4+1 view model; CareNeed

I. INTRODUCTION

Health and disability are inextricably linked to the housing requirements of senior families. Physical as well as mental performance tends to deteriorate with age, raising the prevalence of limitations linked to movement and activity, self-care, and the capacity to operate a home, all of which may impede older individuals' capabilities to live freely in society [1]. This study examined current levels of disabled people, health developments that may affect the above rates in the future, and the forecasting of both the number and size of prospective families in which one or more people are highly likely to have a disorder in order to best explain the accommodation and care requirements for the older community by 2035. The next section examines our aging population's financial well-being to better understand their ability to fulfil their own housing and care demands.

The World Health Organization says that "health" is a state of complete physical, mental, and social health, not just the absence of illness or incapacity [2]. For our purposes, we're curious about how health, or the lack thereof, impacts housing demands. As a result, we investigate health in terms of how it impacts people's ability to do daily self-care and housekeeping duties autonomously, since such tasks are inextricably linked to mobility about the home as well as prospective demands for help and care. In that facility, not being able to do a key daily task on your own is seen as a useful limitation or handicap.

The phrase "activities of daily living" (ADLs) means personal care responsibilities such as showering, dressing, using the toilet, transporting, and eating. Accessible housing may partially address ADLs inside the home; for instance, issues with bathing and toileting might be improved by the installation of walk-in showers, grab bars, and raised toilet seats. In contrast, those with ADL issues often need the aid of caregivers [3]. "Instrumental activities of daily living" (IADLs) are supportive housing skills related to a person's ability to interact with his or her environment, incorporating tasks like marketing, preparing food, cleaning, mobility, financial management, medication administration, and communication utilization [4]. Physical weakness is a common source of IADL-related difficulties; for instance, laundry and routine house-work might require more endurance or ability than an individual possesses. Nevertheless, some responsibilities, including bill payment, food preparation, and medication administration, may have a stronger correlation with mental wellbeing. People with disabilities are often assessed using a combination of ADL and IADL tests, and their impairments may be both neurological and physical. In this research, we also look at mobility issues such as difficulty traveling, walking, and transporting. Certain movement issues may be resolved by using assistive equipment, such as wheelchairs or walkers. Physical improvements to the house may also be used to solve mobility issues. A single-floor living layout, for example, may reduce the difficulty of ascending steps.

We utilize the National Institute on Aging's 2014 Health and Retirement Study (HRS), a continuous information gathering project, to explore the housing repercussions of impairment in the older population. HRS is unique among many disability assessments for the reason that it provides thorough information on healthcare, including functional limitations for any and all persons in the primary participant's

home, allowing us to evaluate disability and residential consequences for both individuals and households [5]. This research may help us determine the number of elderly families affected by impairment, individual attributes (such as ownership and family structure), as well as the scope of the demand for home modifications and assistance. To implement movement limitations, ADLs, and IADLs, we selected specific activities from the disability research that are evaluated by HRS and categorized them into the following categories: movement liabilities, soul functional limitations, and domestic behavior impairments. Mobility restrictions include difficulties walking, transferring in and out of bed, and climbing stairs. In the literature, the ability to transfer is classified as an ADL because.

Neither moving nor climbing stairs are often considered ADLs or IADLs [4]. Nevertheless, we formed the subcategory of flexibility difficulties in order to examine older people's ability to move into residential houses. The literature classifies eating, Dressing, using the toilet, and bathing independently as ADLs, which are among the four self-care limitations. Limitations at work and at home have included the requirement for assistance with IADLs such as meal preparation, shopping, money management, housework, traveling, using the telephone, and medication administration.

An extensive body of work on ADL and IADL limitations that has developed because the two indicators were established in the 1960s and 1970s also indicates that the frequency of impairment drastically rises with age [6, 7]. There is a correlation between income, educational attainment, ethnic origin, relationship status, and the incidence of disability in the elderly. The unmarried, Hispanics, and non-Hispanic blacks have the highest disability rates, as do individuals with low incomes or poor educational attainment. 5. Our examination of HRS data confirms these findings.

As 41 percent of older adults aged 65 to 79 must have had at least one personality, family activity, or movement limitation, but the proportion increases to almost 71 percent for those aged 80 and beyond. Household activity disability is the most prevalent impairment (see Table I). This is a broad category, with elevated numbers generated most often by observed difficulty with housekeeping and driving, both of which are much more prevalent among older age groups [8, 9].

When addressing the consequences of impairment for older individuals' living arrangements, individual occurrence is less important than family occurrence. For instance, if just one member of a newlywed family has a transportation impairment, the housing unit must be adapted to accommodate that individual, regardless of whether the other spouse doesn't seem to [9].

Disability rates are higher for minority families across all three forms of impairment. Hispanic households had the greatest percentage of mobility disabilities among older households (48%), followed by non-Hispanic Asians or other households (41%). Hispanic homes had the greatest percentage of house-hold activity disabilities (62%), followed by non-Hispanic black households (60%).

In the same way, minority families had a higher rate of self-care handicaps than non-Hispanic white households: 31% of non-Hispanic black households and 35% of Hispanic households aged 65 and older have a self-care handicap, compared to 21% of non-Hispanic white households of the same age [8].

Lower-height appliances may eliminate the need for users to raise their hands above shoulder level. Individuals who use mobility aids, such as wheelchair users, may benefit from wider hallways and entrances that allow them to move around the house more easily. Customer happiness is a metric that reflects how well a company's products or services meet consumers' aspirations. It ranks among the most influential indicators of future sales, customer happiness, and loyalty [11]. Since the coronavirus pandemic began, delivery apps have become increasingly important for both business owners and their customers as more individuals order takeout and groceries. There are delivery apps to get individuals food, groceries, and other necessities, but somehow, they have left a segment of the population in need of care and nursing, the PWDs and seniors who have been abandoned by their loved ones for various reasons, without any outside assistance that can provide the same level of trust and security that have developed in some merchants and delivery apps. Based on the 2015 Census of Population and Housing, 1.44 million Americans, or 1.57 percent of the current population of 92.1 million, have a disability. People with disabilities (PWD) were 935,551 in CPH in 2000, or 1.23 percent of the population. [12-14] Region IV-A has the most PWD among the 17 regions, at 193 thousand. The National Capital Region (NCR) ranked second with 6 million people with disabilities (PWD). The Cordillera Administrative Region (CAR) has the fewest people with disabilities (26 thousand) [2]. Thirteen regions had a higher proportion of PWD than the national average. The top five (1.58 percent) were as follows: Region VI (1.95 percent), Region IVB and Region V (1.85 percent each), Region VIII (1.75 percent), Region II (1.72 percent), Region I (1.64 percent), CAR (1.63 percent), Region XI and Region VII (1.60 percent each), and CARAGA (1.60 percent) [15–20].

TABLE I. DISABILITY MEASUREMENT CHART [9, 10]

Disability related to	Difficulty with
Movement	Walking
	Transferring in and out of Bed
	Climbing Stairs
Personal care	Eating
	Dressing
	Toileting
	Bathing
In-house activities	Meal Preparation
	Food Shopping
	Using Telephone
	Taking Medication
	Money Management
	Housework
	Driving

In 2015, men comprised 50.9% of all PWDs, while females comprised 49.1%. Based on these statistics, there are 104 disabled males for every 100 disabled women. Men with disabilities outnumbered females in the 0–64 age categories. The age range of 0 to 14 years had the greatest surplus of men, with a sex ratio of 121 males per 100 females. In the age range of 65 and above, however, a greater number of females than men have disabilities. This is because females have a higher survival rate than males. In this age group, there were 70 males with impairments for every 100 females. One in five people with disabilities were between the ages of 0 and 14; three (59.0 percent) resided between the ages of 15 and 64; and one (22.1 percent) was 65 or older [21]. Individuals with impairments were more likely to be 5 to 19 years old and 45 to 64 years old. Youngsters between the ages of 10 and 14 were the largest age group among the overall impaired population by five-year age group (7.2 percent). This was followed by those aged 15 to 19 (6.9%), 5 to 9 (6.7%), and 50 to 54 years (6.9%). (6.6 percent) [22–24].

The new model indicates that, with the statistics presented, the researcher is considering creating an app that would allow caregivers, nurses, and other individuals to showcase their services, along with fees and locations, and for PWDs and senior citizens to search for these types of services. Ranging from simple tasks such as pushing their wheelchair to specific locations, such as in the park or outside of their home, to more intensive tasks such as bathing, giving them medicine, and taking care of them as a whole. The researcher introduces the CareAide+ app, a mobile software that allows these consumers to experience and care for those who have been unintentionally or intentionally neglected. The purpose of this study is to design a framework for the development of the CareAide+ mobile application, which is a care-needs hub for people with disabilities and senior citizens. This platform will serve as a hub for people with disabilities and senior citizens who seek care that ranges from simple to intensive; at the same time, it will also create a hub for caregivers, freelance practitioners of caregiving, nurses, and other related services. The hub will open up different types of services and different disabilities that most normal people will never know about. This application can also be used by the loved ones of persons with disabilities or seniors in order for them to check all the necessary things. This CareAide+ app would open a different market that would connect this type of person, find what they need, and showcase what they can offer.

The increasing aging population and the prevalence of disabilities among individuals of all ages have created a growing need for accessible and comprehensive care services. The current pandemic has highlighted the difficulties faced by people with disabilities and senior citizens in accessing essential care services. Therefore, developing an app for a care needs hub can help address these challenges by providing a centralized platform that offers a range of care services and resources for people with disabilities and senior citizens. The research motive for creating an app for a care needs hub for people with disabilities and senior citizens could be to: (a) explore the current challenges faced by persons with disabilities and senior citizens in accessing care services, including issues related to accessibility, affordability, and availability. (b) Identify the specific needs and preferences of persons with disabilities and senior citizens regarding care services, such as personal care, medical support, and social interaction. (c) Evaluate the effectiveness of existing care services and resources, including government-funded programs, private initiatives, and community-based organizations. (d) Examine the potential of technology, such as mobile apps, to enhance the accessibility and quality of care services for persons with disabilities and senior citizens. (e) Develop a user-centered design for the app that considers the unique needs and preferences of persons with disabilities and senior citizens, including accessibility features such as voice recognition and text-to-speech capabilities. (f) Conduct user testing and feedback sessions to assess the usability and effectiveness of the app and identify areas for improvement. (g) Assess the impact of the app on the quality of life and well-being of persons with disabilities and senior citizens, as well as their caregivers and families.

Overall, this research aims to develop an innovative solution that addresses the current challenges faced by persons with disabilities and senior citizens in accessing care services and provides a comprehensive and accessible platform for their care needs.

In this paper, the researchers presented our work part by part and described our relations with others work, which is an overview of the work presented in Section I of the Introduction. Working methods and new models are described in the Section II. Methodology is described in Section II. Results and outcomes are described in Section III. A summary of this research is described in Section IV.

II. METHODOLOGY

To completely envision the functional requirements and necessary processes from the target users' perspectives, a qualitative research design was used in the study. This research design was used for the reason that it is mainly focused on the "why" rather than the "what" of the given situation. The 4+1 view model was also utilized to fully forge the designs for the architectural framework of the study. 4+1 is a view model used for "describing the architecture of software-intensive systems, based on the use of multiple, concurrent views" [31].

The gathered data was mainly from observations and unstructured interviews. An in-depth, unstructured interview with random people with disabilities from different group chats on social media and selected senior citizens was conducted to identify and understand the requirements needed to be incorporated into the study. Through these unstructured interviews, the researchers learn more about the different types of disabilities in different forms, what their current situation is, and how and when they get the care they need for the specifics and for the things that they do not understand about the care given by a non-care practitioner family member. To support the unstructured interviews, observations were made on the situational impacts of people with disabilities and senior citizens. From this data, Fig. 1 shows 4+1 view model was adopted with unified modeling language.

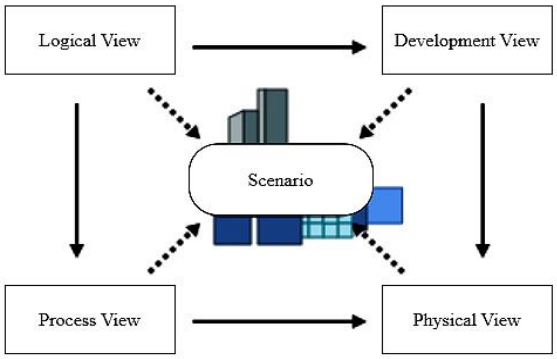


Fig. 1. 4+1 Architectural view model.

A. Related Work

Using a neuromuscular interface controller, a Human Adaptive Mechatronics (HAM) device is controlled. Electromyogram (EMG) signals are used to evaluate muscle movement. A few electrodes placed on the skin's surface can detect the change in signal intensity caused by the voluntary movement of body parts. The EMG signal is employed as the control signal and provides coordination of movement. The signal's amplitude varies in response to muscle activity. The operation of the HAM device is dependent on the amplitude changes in the produced EMG signal. The availability of a cybernetic loop helps balance the system's control error and delayed reaction time. This study describes the operation of a human adaptive neuromuscular interface controller in a mechatronics device with simulations conducted in real time for the various muscle activities, primarily between two actions: voluntary motion and desired motion [25].

The ATM (automated teller machine) system exists to enable a computerized banking network shared by a consortium of banks that includes both human cashiers and ATMs [26]. Each bank maintains its own accounts and processes financial transactions using its own computer. The ATMs interact with the consortium's central computer, which clears transactions with the relevant banks. The ATM interacts with the user, connects with the central computer to conduct transactions, distributes cash, and produces receipts. The system needs proper documentation and security measures. The system must appropriately manage concurrent access to the same account. On their own computers, banks will supply their own software [27].

The architecture of software is concerned with abstraction, deconstruction, and composition, as well as style and aesthetics. In addition, it addresses the design and execution of the software's high-level structure. Architects are responsible for constructing structures utilizing a variety of architectural components in well-considered shapes. These aspects meet the system's primary functionality and performance needs, in addition to non-functional criteria such as dependability, scalability, portability, and system availability [28].

The 4+1 View Model is intended to describe the architecture of software-intensive systems using numerous concurrent views. Each of the five viewpoints focuses on a particular aspect of the system and is detailed with an accompanying notation. These diverse perspectives permit the respective resolution of the issues of the different parties. It may display the architecture from many perspectives and provide the required perspective to various stakeholders. Using an architecture-centered, scenario-driven, iterative development methodology, these perspectives are created [29].

The physical view focuses mostly on non-functional needs and depicts the mapping of software to hardware. The configuration of the hardware has been shown in the physical view. It is necessary to map the many identified components, such as networks, processes, tasks, and objects, onto the various nodes. Certain configurations are used for development and testing, while others are utilized to create the system for different locations or clients [30].

The suggested technique is to assist in resolving the disparity between CRPD guiding principles and regional and national legal practices of interpretation. This technique intends to execute a rhetorical strategy tied to the Thirdness thesis in order to develop a new disability legal culture. Through a thirdness theory, the law of disability, conceived as a unitary figure and inspired by a rhetorical methodology, can contribute to the theoretical evolution of the methodology of legal interpretation without limiting itself to raising fundamental questions of justice philosophy [30], such as a new theory of the concept of a legal person. So, the rediscovery of the rhetorical approach might result in a metamorphosis of legal thinking capable of transcending certain aporia inherent to the positivistic idea of law. In addition, the reintroduction of the rhetorical approach may make it feasible to comprehend the structural function of fiction juris in legal reasoning. A few concluding remarks might be made on the link between ethics and rhetoric with respect to the rhetorical structure of thirdness in the trial, in which the judge is positioned after the parties [31].

Utilizing the Delphi research approach, a search strategy of sites, publications, and research papers was done to design a survey comprising questions about the information, abilities, and actions required to aid a person with an intellectual impairment who has been having mental health issues. A panel of experts evaluated these issues over the course of three polling rounds to determine if they should be included in the recommendations [30].

In total, 53 experts completed all three survey rounds (a retention rate of 67%). During the course of three rounds, 202 items were evaluated, resulting in 170 recommended items that have been put into the guidelines. The recommendations emphasize the necessity of recognizing the distinctive indicators of mental health issues in individuals with an intellectual impairment and of providing appropriate assistance, understanding, and compassion for these individuals. The recommendations will also strengthen caregivers' abilities to confront concerning or economically restrictive behaviors or to access professional assistance whenever necessary. The criteria are going to be employed to create a course on psychological disorders and first aid [31].

III. Results and Discussion

After measured consideration of all gathered data, the following system models were crafted to create a framework for the study, shown in Table II.

A. Scenarios

A limited selection of use cases, or scenarios, are used to show an architecture's definition, resulting in a fifth perspective. The purpose of the diagram was to assist in understanding how the proposed system would function in relation to the actors. They are used for high-level system condition analysis. Fig. 2 depicts the system's overall use case diagram from a higher-level perspective.

The PWD/senior, healthcare professionals, hospitals, administrators, and customer support are the characters in the use case diagram in Fig. 2. The diagram's purpose is to help users update their inquiries and grievances so that the customer service module can deal with them later. Use cases associated with it at a more fundamental level of abstraction include filing complaints about abuse, getting compensation, scheduling meetings with management, and contacting management.

B. Logical View

- The conceptual perspective is centered on the service's final functioning. A static view of the program is provided by the class diagram. One of the key factors in its use during building is that it can be directly mapped with object-oriented languages [32].
- Fig. 3 displays the class diagram for the suggested application. The objective is to make it easier to understand the subtleties of the interactions between various classes. Each class consists of a rectangular box that operates in a different method and accepts attributes [33].

TABLE II. Users Stories from Public Survey

As a	I want to...	so that...
Care Giving Individual	Add a service	I can showcase my services to the needs of PWDs and Seniors and have a decent fee for it
Care Giving Individual	Specify my service	My clients/patients will have accurate knowing of the services and care they need for them or their patient
Care Giving Individual	Upload my credentials and experiences	My clients/patients will have an idea of my credibility to take this service and gain their trust
Care Giving Individual	Write a description of my services	I can describe and thoroughly explain what are this type of services
Care Giving Individual	Specify the fees or amount of my services	My client/patient will have accurate expectations
Care Giving Individual	Specify my location range	My client/patient will have accurate knowing of my location
Care Giving Individual	Specify and list my policies for my service	My client/patient understands my expectations for them.
Care Giving Individual	Respond to messages from clients/patients	I can quickly respond on inquiries about my service listing
Care Giving Individual	Accept requests	I can accept or reject requests to whom I want to have my services
Care Giving Individual	See past reviews of the requester (as client/patient)	I can assess the quality of the requester before I allow them to have my service rendered
Care Giving Individual	See a message history between myself and the requester	I can remember what was communicated between me and my client/patient
Care Giving Individual	Leave a review of a client/patient with an overall rating (1-5)	I feel assured the client/patient will treat me with respect.
Client/patient	Browse services	I can look for certain services upon my needs
Client/ patient	See past reviews of services of care giving individuals, and other services by the same care giving individuals	I can assess the quality of the services and the Care Giving Individual as to the matter of professionalism
Client/ patient	Send a message to the care giving individual as to the services offered	I can ask the care giving individual about anything not covered in the services page.
Client/ patient	See the schedule of a care giving individual if available for	I can plan for the said service offered
Client/ patient	Send a request for available schedule for the service	I can quickly lock down a service offered by the care giving individual
Client/ patient	Include an introduction of myself	I can make the care giving individual feel comfortable about having me as a client/patient and increase the likelihood that they will approve my request.
Client/ patient	Leave a review of the service	I feel assured that the care giving individual will care about my experience.
Client/ patient	Review the accuracy of the service offered (1-5)	I feel assured the care giving individual will be truthful in their service offered page description.
Client/ patient	Review the communication of the host (1-5)	I feel assured the care giving individual will be responsive to my messages.
Client/ patient	Leave a written review (in freeform text)	I can describe my experience in my own words.

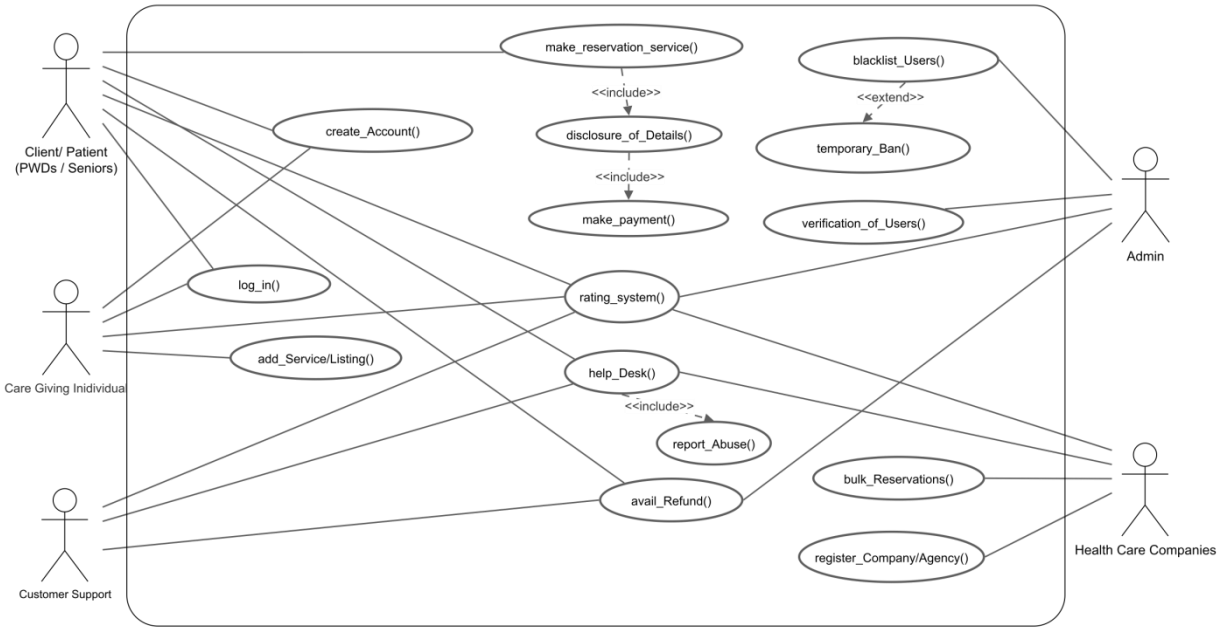


Fig. 2. Use case diagram of CareAide+.

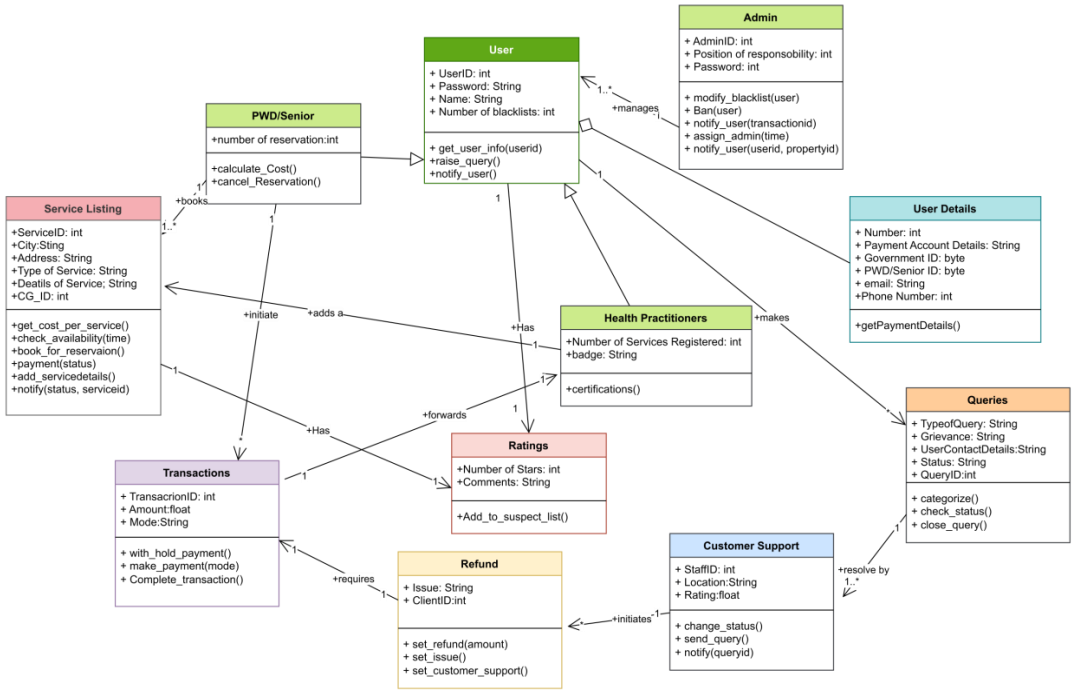


Fig. 3. Class diagram for the CareAide+.

C. Process View

The process perspective focuses on the performance of the service and is concerned with the flexible aspects of the system. It also describes system activities and interactions.

Fig. 4 shows graphical representations that simulate the logic of a complex method, function, or action and describe the intricate details of a UML use case. The flow charts attempted to illustrate the idea of the project's key processes, including developing services, utilizing facilities, and paying for interactions.

According to the flowchart in Introducing a Platform, the current variables are client, service description, administrator, and user information. The administrator must confirm the patient's identity and financial information when the user adds a support showing, which disables the Care Review component and notifies the administrator via the Communication Lineup. If the status is successful, the client or patient is informed and the property listing is updated; if it is unsuccessful, the client or patient is informed. A reference is used to notify the user of a status change in order to avoid cluttering the diagram and make it easier to understand.

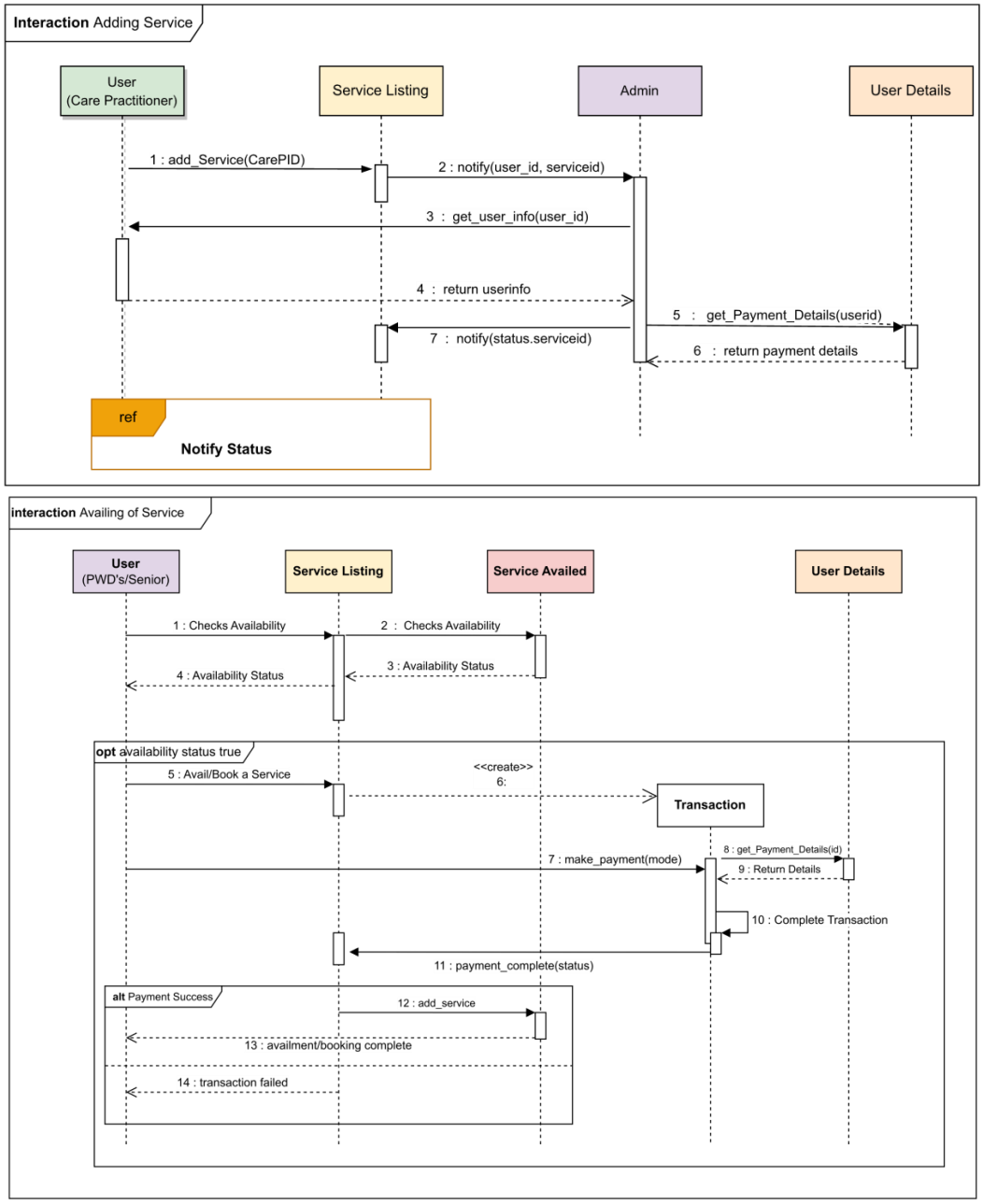


Fig. 4. Sequence diagram for the CareAide+.

In the Availment of a Service, the objects involved are the user, the service listing, the service availed, and the user details. The steps involved in a user requesting a service are shown in this flowchart. The user must first check whether the service is accessible on the date or at the time specified, and only then may the user use the service. When a service is requested, a new instance of the transaction class is created.

The actual perspective, also known as the implementation perspective, illustrates the infrastructure as a technical team would see it. Both the physical connections between software components and their layer-by-layer structure are of interest. Fig. 5 shows how the Layered and Model View Controller architecture pattern worked in the implementation of the visualization services of the CareAide+ application. The Presentation layer adopts the View component to display various visualization layouts and widgets, especially user reviews and ratings. The application or business layer implements the controller component in processing requests and then communicates the result to the persistence layer, where the model component retrieves and updates data with the database layer. The database layer then sends the updates to the persistence layer through the model component until they are displayed in the presentation layer through the view component.

The forecasting model contains data along with the explanation associated with that as well. It describes data transmitted among controlling sections or even other relevant circuitry. A control system module, for example, will fetch

client or patient details from the database. Before putting the data back into the repository or using it to create data that is somewhat similar to it, it tries to manipulate the data. A view is an interface element that contains and presents data. Presentations are made using the information obtained from the model data. A viewpoint asks the model for information in order to give the client a special highlight. The Control System is the system element in charge of interactivity. The controller evaluates the inputs from the user's input devices before deciding how the model and views should react. The model receives instructions from a controller to modify its state (for instance, by saving a specific document). The controller can modify how a view is displayed by sending commands to views that are related to it. For example, scrolling while turning the page.

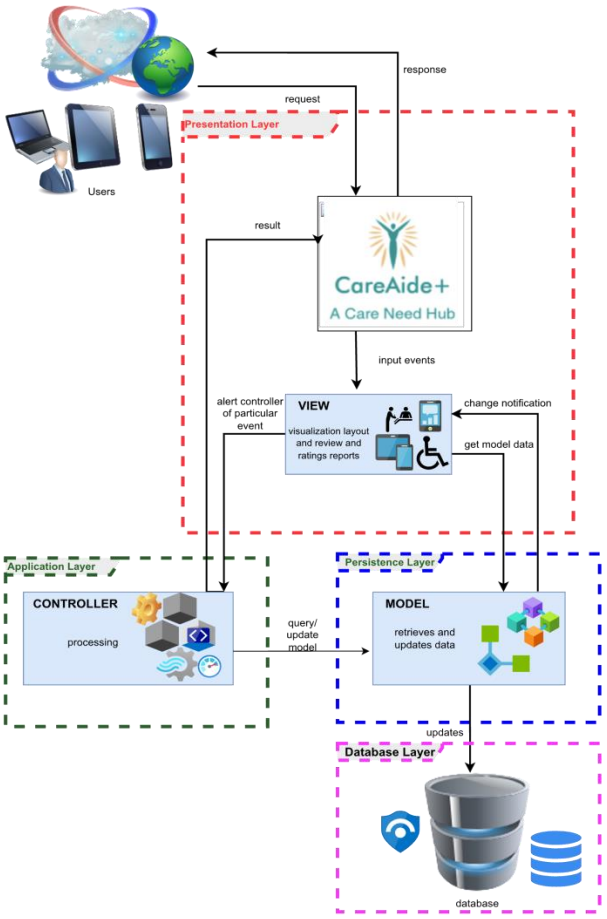


Fig. 5. Architecture design pattern using model-view-controller pattern mapped in a layered architecture for the CareAide+.

The MVC Pattern Interface has advantages such as simple code that is easy to grow and expand. MVC model elements may be tested without the involvement of the user. Assistance for various clienteles is now simpler. It is possible for various components to evolve. By dividing an application into three parts, it helps to avoid complexity. Components include the model, view, and controller. It only uses the Front Controller design pattern, which uses a single controller to route requests from custom applications. It presents the toughest development assistance. It works well for websites that are supported by large teams of web designers and engineers. It guarantees a distinct separation of concerns (SoC) [34]. Search engine optimization (SEO) is friendly since, each class and object is self-contained, you can test each one separately. The MVC design pattern [35] makes it possible to logically group related controller operations. The drawbacks are as follows: Understanding, modifying, unit testing, and reusing this model are challenging. Because the framework introduces additional levels of abstraction that force users to adapt to the MVC decomposition requirements, navigating the framework can occasionally be difficult. No explicit validation support is provided. Data inefficiency and complexity have grown. the difficulty of using MVC with a modern graphical user interface [36]. The use of many programmers is required for parallel programming. It is necessary to be familiar with a variety of technologies and management of several codes in the controller.

IV. Conclusion

The crafted architecture design for the care needs hub application would be a comprehensive tool for the completion of the development. This would greatly help the developers of the system in building the correct features and data abstractions needed to build and implement the said system. The researchers then recommend conducting an in-depth analysis of the features to be included in the application and considering adding payment alternatives or revising payment modules to other modes, like non-monetary things, if it would be possible. According to the research, a hub of helping hands is very important for senior citizens around the world to survive their daily lives. Our model helps to increase humanity, reliability, and confidence all over the world for overaged people. It is possible that in the future, with the help of artificial intelligence, it will be developed for automatic operation. If we are able to develop a model for the care hub that operates on its own, it will be more accurate and have lower costs for individuals.

References

[1] Y. Wang and Z. Liu, “A workflow based self-care management system,” Conf. Proc. IEEE Eng. Med. Biol. Soc., vol. 2006, pp. 558–561, 2005. doi: 10.1109/IEMBS.2005.1616472.

[2] A. Lanata, G. Valenza, M. Nardelli, C. Gentili, and E. P. Scilingo, “Complexity index from a personalized wearable monitoring system for assessing remission in mental health,” IEEE J. Biomed. Health Inform., vol. 19, no. 1, pp. 132–139, 2015. doi: 10.1109/JBHI.2014.2360711.

[3] X. Cao, E. Klinger, A.-S. Douguet, and P. Fuchs, “Issues in the design of a virtual Instrumental Activity of Daily Living (vIADL) for Executive Functions exploration,” in 2009 Virtual Rehabilitation International Conference, 2009. doi: 10.1109/ICVR.2009.5174244.

[4] V. Buso, L. Hopper, J. Benois-Pineau, P.-M. Plans, and R. Megret, “Recognition of Activities of Daily Living in natural ‘at home’ scenario for assessment of Alzheimer’s disease patients,” in 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2015. doi: 10.1109/ICMEW.2015.7169861.

[5] Y. Zeng, Q. F. Jia, and J. Zhou, “Does policy of delayed retirement affect individual health,” in 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2017. doi: 10.1109/IEEM.2017.8290019.

[6] D. Portela, M. Almada, L. Midão, and E. Costa, “Instrumental activities of daily living (iADL) limitations in Europe: An assessment of SHARE data,” Int. J. Environ. Res. Public Health, vol. 17, no. 20, 2020. doi: 10.3390/ijerph17207387

[7] B. Roehrig, K. Hoeffken, L. Pientka, and U. Wedding, "How many and which items of activities of daily living (ADL) and instrumental activities of daily living (IADL) are necessary for screening," Crit. Rev. Oncol. Hematol., vol. 62, no. 2, pp. 164–171, 2007. doi: 10.1016/j.critrevonc.2006.10.001.

[8] N. Farmer, G. R. Wallen, L. Yang, K. R. Middleton, N. Kazmi, and T. M. Powell-Wiley, "Household cooking frequency of dinner among Non-Hispanic black adults is associated with income and employment, perceived diet quality and varied objective diet quality, HEI (Healthy Eating Index): NHANES analysis 2007-2010," Nutrients, vol. 11, no. 9, p. 2057, 2019. doi: 10.3390/nu11092057.

[9] M. P. Lawton and E. M. Brody, "Assessment of older people: Self-maintaining and instrumental activities of daily living," Gerontologist, vol. 9, no. 3 Part 1, pp. 179–186, 1969. doi: 10.1093/geront/9.3_Part_1.179.

[10] M. P. LaPlante, "The classic measure of disability in activities of daily living is biased by age but an expanded IADL/ADL measure is not," J. Gerontol. B Psychol. Sci. Soc. Sci., vol. 65, no. 6, pp. 720–732, 2010. doi: 10.1093/geronb/gbp129.

[11] P. L. Wright et al., "Epithelial reticulon 4B (Nogo-B) is an endogenous regulator of Th2-driven lung inflammation," J. Exp. Med., vol. 207, no. 12, pp. 2595–2607, 2010.

[12] M. J. Sirgy, "Macromarketing metrics of consumer well-being: An update," J. Macromarketing, vol. 41, no. 1, pp. 124–131, 2021. doi: 10.1177/0276146720968096.

[13] S. Chatterjee, D. Goyal, A. Prakash, and J. Sharma, "Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application," J. Bus. Res., vol. 131, pp. 815–825, 2021. doi: 10.1016/j.jbusres.2020.10.043.

[14] S. Mitra, W. Chen, J. Hervé, S. Pirozzi, and J. Yap, "Invisible or mainstream? Disability in surveys and censuses in low- and middle-income countries," Soc. Indic. Res., vol. 163, no. 1, pp. 219–249, 2022. doi: 10.1007/s11205-022-02879-9.

[15] E. Williams et al., "Perspectives of basic wheelchair users on improving their access to wheelchair services in Kenya and Philippines: a qualitative study," BMC Int. Health Hum. Rights, vol. 17, no. 1, 2017. doi: 10.1186/s12914-017-0130-6.

[16] B. M. Altman and E. K. Rasch, "Purpose of an international comparable census disability measure," in International Measurement of Disability, Cham: Springer International Publishing, 2016, pp. 55–68. https://doi.org/10.1007/978-3-319-28498-9_4.

[17] M. Schneider and A. Montes, "The Asian testing experience," in International Measurement of Disability, Cham: Springer International Publishing, 2016, pp. 123–135. https://doi.org/10.1007/978-3-319-28498-9_8.

[18] B. M. Altman and H. Meltzer, "Developing tools to identify environmental factors as context for disability: A theoretical perspective," in International Measurement of Disability, Cham: Springer International Publishing, 2016, pp. 183–206. https://doi.org/10.1007/978-3-319-28498-9_12.

[19] M. Marella, A. Devine, G. F. Armecin, J. Zayas, M. J. Marco, and C. Vaughan, "Rapid assessment of disability in the Philippines: understanding prevalence, well-being, and access to the community for people with disabilities to inform the W-DARE project," Popul. Health Metr., vol. 14, no. 1, p. 26, 2016. https://doi.org/10.1186/s12963-016-0096-y.

[20] E. E. Rotas and M. Cahapay, "Managing the mental health of persons with disabilities amid the COVID-19 pandemic in the Philippines: Specific factors and key actions," Eur. J. Environ. Public Health, vol. 5, no. 2, p. em0077, 2021. https://doi.org/10.21601/ejeph/10954.

[21] L. G. Martin, "The aging of Asia," J. Gerontol., vol. 43, no. 4, pp. S99-113, 1988. https://doi.org/10.1093/geronj/43.4.S99.

[22] D. Gu, R. Gomez-Redondo, and M. E. Dupre, "Studying disability trends in aging populations," J. Cross. Cult. Gerontol., vol. 30, no. 1, pp. 21–49, 2015. https://doi.org/10.1007/s10823-014-9245-6.

[23] J. B. Abalos, Y. Saito, G. T. Cruz, and H. Booth, "Who cares? Provision of care and assistance among older persons in the Philippines," J. Aging Health, vol. 30, no. 10, pp. 1536–1555, 2018. https://doi.org/10.1177/0898264318799219.

[24] M. P. Tolentino and M. S. Kakihara, "The challenges of population ageing in the Philippines and Brazil," Journal of Asian Societies, pp. 59–66, 2021.

[25] J. P. A. Joel, R. J. S. Raj, C. A. D. Durai, and R. Vedaiyan, "RETRACTED ARTICLE: Human adaptive mechatronics system integrated with cybernetics loop using neuromuscular controller in occupational therapy for elderly person with disability," J. Ambient Intell. Humaniz. Comput., vol. 12, no. 7, pp. 7299–7308, 2021. https://doi.org/10.1007/s12652-020-02405-0.

[26] D. Garlan and M. Shaw, "An Introduction to Software Architecture," https://www.cs.cmu.edu/afs/cs/project/able/www/paper_abstracts/intro_softarch.html, 15-Oct-2001.

[27] H. Choi and K. Yeom, "An approach to software architecture evaluation with the 4+1 view model of architecture," in Ninth Asia-Pacific Software Engineering Conference, 2002, 2003. doi: 10.1109/APSEC.2002.1182998.

[28] P. B. Kruchten, "The 4+1 view model of architecture," IEEE Softw., vol. 12, no. 6, pp. 42–50, 1995. doi: 10.1109/52.469759.

[29] T. Robal, V. Viies, and M. Kruus, The Rational Unified Process with the "4+1" View Model of Software Architecture - a Way for Modeling Web Applications. BalticDB&IS, 2002.

[30] K. S. Bond et al., "Considerations when offering mental health first aid to a person with an intellectual disability: a Delphi study," BMC Psychol., vol. 9, no. 1, p. 28, 2021. https://doi.org/10.1186/s40359-021-00518-5.

[31] S. Meng et al., "The "4+1" view model on safe home system architecture," in 2010 IEEE International Conference on Software Engineering and Service Sciences, 2010. doi: 10.1109/ICSESS.2010.5552450.

[32] M. D. Adu, U. H. Malabu, A. E. O. Malau-Aduli, and B. S. Malau-Aduli, "The development of My Care Hub mobile-phone app to support self-management in Australians with type 1 or type 2 diabetes," Sci. Rep., vol. 10, no. 1, p. 7, 2020. https://doi.org/10.1038/s41598-019-56411-0.

[33] L. A. Jibb, B. J. Stevens, P. C. Nathan, E. Seto, J. A. Cafazzo, and J. N. Stinson, "A smartphone-based pain management app for adolescents with cancer: establishing system requirements and a pain care algorithm based on literature review, interviews, and consensus," JMIR Res. Protoc., vol. 3, no. 1, p. e15, 2014. doi: 10.2196/resprot.3041.

[34] D. V. Gunasekeran, Y.-C. Tham, D. S. W. Ting, G. S. W. Tan, and T. Y. Wong, "Digital health during COVID-19: lessons from operationalising new models of care in ophthalmology," Lancet Digit. Health, vol. 3, no. 2, pp. e124–e134, 2021. https://doi.org/10.1016/S2589-7500(20)30287-9.

[35] B. Unhelkar and S. Murugesan, "The enterprise mobile applications development framework," IT Prof., vol. 12, no. 3, pp. 33–39, 2010. doi: 10.1109/MITP.2010.45.

[36] D. Sambasivan, N. John, S. Udayakumar, and R. Gupta, "Generic framework for mobile application development," in 2011 Second Asian Himalayas International Conference on Internet (AH-ICI), 2011. doi: 10.1109/AHICI.2011.6113938.

Implementation of Cybersecurity Situation Awareness Model in Saudi SMEs

Monerah Faisal Almoaigel, Ali Abuabid
Informa Technology Department, College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia

***Abstract*—Saudi Small and Medium-sized Enterprises (SMEs) are witnessing rapid growth in technology and innovation. However, this growth is accompanied by increased cybersecurity threats, which pose significant challenges for SMEs. Cyber threats are becoming more complex and sophisticated, with SMEs becoming prime targets due to their weaker cybersecurity defenses. Hence, there exists a rich literature on critical challenges facing SMEs. Existing literature on these challenges addresses many research issues (e.g., finance, technology adoption, and management) associated with SMEs. However, one critical issue that has so far received no rigorous attention is cybersecurity situation awareness for research in the SME context. Thus, this study used a quantitative approach aiming to empirically test a model of cybersecurity situational awareness that can support SMEs in Saudi Arabia to implement cybersecurity measures and precautions with efficacy. An online survey of 350 participants was conducted to collect the research data. The study identified a significant positive relationship between Cyber Situational Awareness (Csa) and Implementation of Cybersecurity Controls (Icsc), suggesting that enhancing awareness can contribute to better control implementation. The study identified a significant positive relationship between Cyber Situational Awareness (Csa) and Implementation of Cybersecurity Controls (Icsc), suggesting that enhancing awareness can contribute to better control implementation. Finally, the paper provides several interesting findings and outlines future research directions.**

***Keywords*—Cyber situation awareness; cybersecurity control and precaution; Saudi; SMEs**

I. Introduction

The Kingdom of Saudi Arabia is rapidly growing in technology and innovation, with Small and Medium-sized Enterprises (SMEs) playing a significant role in this growth. However, this growth comes with increased cybersecurity threats, which pose significant challenges for SMEs. Cyber threats are growing in complexity and sophistication, and SMEs are becoming a primary target for cyber attackers due to their weaker cybersecurity posture. Therefore, SMEs might adopt a robust cybersecurity strategy that includes cyber situation awareness (CSA) to combat cyber threats effectively. SMEs are increasingly becoming targets for cyber attackers. SMEs in the UK alone are targeted by cyber-attacks more than seven million times yearly [1]. Similarly, in Saudi Arabia, a study by [2] found that SMEs were particularly vulnerable to cyber-attacks due to their limited resources and expertise.

One of the main reasons for the high rate of cyber-attacks on SMEs is their perceived vulnerability. Hackers often assume that SMEs have weaker security measures than larger organizations, making it easier to attack targets [3]. Furthermore, SMEs often lack the resources to invest in advanced cybersecurity technologies and hire dedicated staff [2]. As a result, they may be more susceptible to common attacks such as phishing, ransomware, and social engineering.

Cyber-attacks' impact on SMEs can be significant regarding financial losses and reputational damage. A study by [4] found that SMEs in the Netherlands lost an average of £65,000 per cyber-attack. Furthermore, the reputational damage caused by a cyber-attack can be particularly damaging for SMEs, as they may struggle to regain the trust of their customers and partners.

Researchers have proposed various solutions to address these challenges to improve SMEs' cybersecurity posture. These include investing in essential security measures such as firewalls and antivirus software, implementing employee training programs to improve cybersecurity awareness, and developing incident response plans to ensure businesses are prepared to respond to cyber-attacks [5].

Despite the serious consequences of cyber-attacks, SMEs often lack the resources and expertise to implement effective cyber security measures. Therefore, there is a need for a cyber situational awareness model that can assist SMEs in Saudi Arabia to implement cyber security controls and precautions effectively. Thus, this study aims to empirically test a model of cybersecurity situational awareness that can support SMEs in Saudi Arabia to implement cybersecurity measures and precautions with efficacy. Specifically, this study aims to validate the proposed model using empirical data collected from SMEs in Saudi Arabia to ensure its applicability and usefulness in the local context.

Overall, this paper is expected to contribute by providing a practical and effective framework for SMEs in Saudi Arabia to implement cyber security measures, thereby reducing their vulnerability to cyber-attacks. In other words, the extension of Endsley's theory of situation awareness to the cyber security domain can contribute to developing a more comprehensive understanding of how situational awareness can be leveraged to enhance cyber security in SMEs.

The remainder of this paper offers a comprehensive analysis of related literature in Section II. Then, the research approach used in this paper is discussed in Section III and followed by the findings that are presented and discussed in Section IV. Finally, the paper ends by summarizing the key findings, contributions, limitations, and future work in Section V.

II. Literature Analysis

A. Cybersecurity Situation Awareness Model

Situational awareness is a term that originated in the field of aviation in the 1940s and refers to a pilot's ability to accurately understand their current situation and the potential risks and opportunities in the environment around them [6]. Over time, situational awareness has been applied to various fields, including military operations, cybersecurity, and emergency response [7]. In the context of cybersecurity, situational awareness refers to the ability of an organization or individual to understand the current state of their cyber environment, including potential threats and vulnerabilities, and to use this understanding to make informed decisions about how to protect their assets and respond to potential incidents [1].

A cyber situation awareness model is a framework or methodology designed to help organizations improve their situational awareness in the cybersecurity domain. Such a model typically includes a set of processes and tools that enable an organization to monitor its network infrastructure and collect and analyze data about potential threats and vulnerabilities. It makes informed decisions about responding to incidents or improving its security posture [8]. Many different cyber situational awareness models have been developed, each with strengths and weaknesses. Some models are designed for specific industries or organizational contexts, while others are more general. Some models rely heavily on automated tools and data analytics, while others emphasize human expertise and decision-making.

One of the most widely used frameworks in the field of cybersecurity is the Kill Chain framework, which was first introduced by Lockheed Martin in 2011 [9]. The Kill Chain framework is designed to help organizations understand the different stages of a cyber-attack, from initial reconnaissance to final data exfiltration, and to develop appropriate defenses at each stage. The cybersecurity industry has widely adopted many organizations' frameworks to guide their cybersecurity strategies.

Another framework that has gained traction is the Diamond Model, introduced in 2014 by a group from the US Army Research Laboratory [10]. The Diamond Model is based on the premise that cyber-attacks are dynamic. Understanding the relationship between an attacker, a victim, an infrastructure, and an impact helps organizations better understand the threats they face and develop appropriate defenses. The US government and other organizations have used the framework to improve their situational awareness and incident response capabilities [11].

Moreover, the Cyber Security Situation Awareness Framework (CSSAF) was developed by researchers from the University of Plymouth in the UK and is designed to help organizations improve their situational awareness by integrating data from multiple sources, including network traffic, system logs, and threat intelligence feeds [12]. Several organizations have used the CSSAF to improve their cybersecurity posture, and the framework effectively detects and mitigates cyber-attacks.

The MITRE ATT&CK framework is a relatively new situational awareness framework introduced in 2015 [13]. The framework is designed to help organizations understand the tactics, techniques, and procedures (TTPs) that attackers commonly use and to develop appropriate defenses based on this knowledge. Many organizations use the MITRE ATT&CK framework to guide their cybersecurity strategies [14]. The following section will detail the cyber situation awareness models for SMEs in Saudi Arabia.

B. Cybersecurity Situation Awareness Model and Saudi SMEs

Saudi SMEs are businesses with less than 250 employees and annual revenue of less than 200 million SAR [15]. SMEs play a critical role in the economy of Saudi Arabia. Moreover, SMEs represent about 99% of all businesses in the country and employ around two-thirds of the private sector workforce [16].

Despite their significant contribution to the economy, SMEs in Saudi Arabia face several challenges, including access to financing, limited access to skilled labor, and a complex regulatory environment [17]. These challenges can be particularly acute in cybersecurity, where SMEs often lack the resources and expertise to effectively protect themselves against cyber threats. One of the main reasons for the high rate of cyber-attacks on SMEs is their perceived vulnerability. Hackers often assume that SMEs have weaker security measures than larger organizations, making it easier to attack targets [3]. Furthermore, SMEs often lack the resources to invest in advanced cybersecurity technologies and hire dedicated cybersecurity staff [2]. As a result, they may be more susceptible to common attacks such as phishing, ransomware, and social engineering. Cyber-attacks' impact on SMEs can be significant regarding financial losses and reputational damage. The research in [4] found that SMEs lost an average of £65,000 per cyberattack. Furthermore, the reputational damage caused by a cyberattack can be particularly damaging for SMEs, as they may struggle to regain the trust of their customers and partners.

Several recent studies have examined SMEs' challenges and opportunities in Saudi Arabia. For instance, the study in [18] found that accessing finance was one of the most significant challenges facing SMEs in Saudi Arabia. With struggling to secure funding, SMEs needed to grow and expand. The study also identified a lack of skilled labor and bureaucratic red tape as significant obstacles to growth. In addition, the study in [2] found that many SMEs in Saudi Arabia were not adequately prepared to protect themselves against cyber threats. This study also highlighted that only 30% of Saudi SMEs had implemented cybersecurity controls or precautions. Among those that had, many were using outdated or ineffective approaches, such as antivirus software and firewalls.

To address these challenges, researchers (e.g., [19], [20], and [1]) have proposed a range of solutions aimed at improving the cybersecurity posture of SMEs. These include investing in essential security measures such as firewalls and antivirus software, implementing employee training programs to improve cybersecurity awareness, and developing incident

response plans to ensure businesses are prepared to respond to cyberattacks. Overall, the high cyberattack rate on SMEs highlights these businesses' need for greater awareness and investment in cybersecurity measures. While there is no one-size-fits-all solution to these challenges, a proactive and comprehensive approach to cybersecurity can help SMEs mitigate the risks of cyberattacks and protect their businesses.

SMEs should adopt a robust cybersecurity strategy that includes cyber situation awareness (CSA) to combat cyber threats effectively and benefit from greater access to training and resources on cybersecurity best practices. For instance, [21] proposed a framework for improving cybersecurity awareness and education among SMEs. This might help to reduce the risk of cyberattacks and improve the overall security posture in Saudi SMEs.

Furthermore, the research in [22] proposed a model that includes the following components: data collection, data processing, threat identification, and response. They suggest that the model can be used to develop effective cybersecurity strategies for SMEs. Similarly, [23] proposed a cybersecurity awareness model that includes four components: data collection, data analysis, data dissemination, and response. The model can enhance cyber security control and precaution in SMEs.

To gain a comprehensive understanding of the actual cybersecurity practices in Saudi SMEs. The researchers use the proposed model by [1], an extension of Endsley's situational awareness theory [6], to portray SMEs' cybersecurity situational awareness. Fig. 1 portrays the proposed SME cyber security situational awareness model.

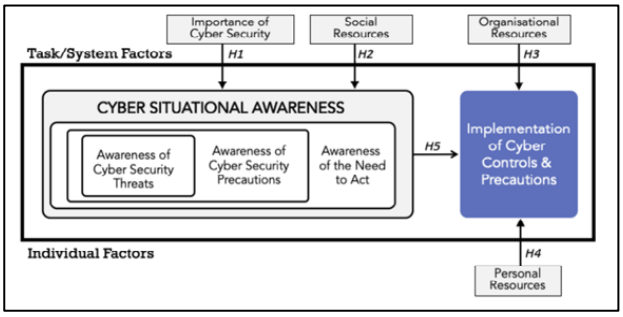


Fig. 1. Cyber security situational awareness in SMEs. (Karen et al., 2021).

The importance of cybersecurity helps SMEs administration understand that maintaining cyber security mechanisms enhances their business continuity. The availability of social resources enhances the awareness of SME administration about cybersecurity threats, precautions that should be applied to respond to these threats, and how they act accordingly. In contrast, the availability of organizational and personal resources facilitates the implementation of cyber controls and precautions. Table I summarizes the hypothesis that needs to be tested to validate the model.

- The first hypothesis (H1) suggests that the level of cyber situational awareness, which refers to understanding potential cybersecurity threats and vulnerabilities, is influenced by how much the SMEs understand the importance of cyber security in the context of Saudi SMEs [1]. This means that SMEs in Saudi Arabia are more likely to better understand potential cybersecurity threats and vulnerabilities if they recognize the importance of cybersecurity.
- The second hypothesis (H2) proposes that social resources influence the level of cyber situational awareness in Saudi SMEs [2] . This means that Saudi SMEs are more likely to be aware of potential cybersecurity threats and vulnerabilities if they access relevant social resources such as expert advice, training, or support from other SMEs.
- The third hypothesis (H3) contends that organizational resources influence the implementation of cyber security controls and precautions in Saudi SMEs [2]. This means that Saudi SMEs are more likely to implement cybersecurity controls and precautions effectively if they have the necessary organizational resources, such as funding, technology, and personnel.
- The fourth hypothesis (H4) presents that personal resources influence the implementation of cyber security controls and precautions in Saudi SMEs [1]. This means that individuals within Saudi SMEs are more likely to implement cybersecurity controls and precautions effectively if they have the necessary personal resources, such as technical skills and knowledge.
- The fifth hypothesis (H5) proposes that the level of cyber situational awareness influences the adoption of cyber security controls and precautions in Saudi SMEs [1]. This means that Saudi SMEs are more likely to adopt adequate cybersecurity controls and precautions to protect their information systems if they are aware of potential cybersecurity threats and vulnerabilities.

TABLE I. PROPOSED RESEARCH HYPOTHESIS

No.	Hypothesis Description
H1	The level of cyber situational awareness in SMEs is influenced by their understanding of the importance of cyber security.
H2	Their social resources influence the cyber situational awareness in SMEs.
H3	Implementing cyber security controls and precautions in SMEs is influenced by their organizational resources.
H4	Implementing cyber security controls and precautions in SMEs is influenced by their personal resources.
H5	Their level of cyber situational awareness influences the adoption of cyber security controls and precautions in SMEs.

III. RESEARCH APPROACH

This study used an online survey approach to understand SME perceptions of factors affecting cybersecurity situational awareness. An online survey was considered appropriate due to fast access to individuals, increased ability to reach difficult contact participants, and ease of having automated data collection, which reduced researchers' time and effort [24]. Furthermore, online surveys save researchers money due to the electronic data collection [25].

The study was conducted in the spirit of the positivist research tradition and followed four stages: literature analysis to develop the theoretical concepts, survey instrument development, administration of the survey, and empirical data analysis. The literature analysis identified a set of cybersecurity situational awareness factors. These served as the foundation for developing an initial survey instrument divided into four parts: profile of responding managers, characteristics of participating business, factors affecting cybersecurity situational awareness in SMEs, and implementation of cybersecurity controls and precautions.

The target online survey participants are SMEs in Saudi Arabia registered with the Small and Medium Enterprises General Authority (Monsha'at) (Small and Medium Enterprises General Authority, 2023). It develops several supporting programs and projects that advance a culture of self-employment, private enterprise, and innovation. It also provides funding sources and develops standards and policies for SMEs.

The online survey was sent (via email and SMEs' social media accounts) to the Saudi SMEs, and 350 responded to the online survey. The obtained low response was not a shock, [26] stated that online survey has often been plagued by low response.

IV. Results and Discussion

A. Importance of Cyber Security (Ics)

The research aimed to assess the importance of cyber security for SMEs, and the participants were asked two items to gather insights into the importance of cyber security in their SMEs. The results revealed that 54.7% of the participants acknowledged that cybersecurity is critical to their business. In contrast, only 5.1% believed it to be of very low importance. Fig. 2 shows descriptive statistics of the importance of cybersecurity 1 (Ics1).

To further understand the participants' approach to recognizing cyber security risks that may affect their business, they were asked about the measures they had taken in the previous year. The responses were diverse, with most participants (a significant portion) indicating that their companies had implemented 10 or more measures to mitigate risks. The second-highest response came from participants who were unsure about the measures taken by their organization. Surprisingly, the lowest response came from participants who felt that none of the options provided were suitable, suggesting a lack of proactive measures. Fig. 3 shows the descriptive statistics of the importance of cybersecurity 2 (Ics2).

Upon analyzing these results, it became evident that participants who perceived their businesses to be at a very high risk of cyber security threats were more inclined to implement more measures to safeguard their operations. Conversely, those who considered the risk shallow exhibited either a lack of awareness regarding the measures taken or a complete absence of proactive actions. Interestingly, despite acknowledging a high risk, some participants did not appear concerned or motivated to take appropriate measures.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	High	68	19.4	19.4	19.4
	Low	21	6.0	6.0	25.4
	Neutral	52	14.8	14.8	40.2
	Very High	192	54.7	54.7	94.9
	Very Low	18	5.1	5.1	100.0
	Total	351	100.0	100.0	

Fig. 2. Descriptive statistics of the importance of cybersecurity 1(Ics1).

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1-5 Times	64	18.2	18.2	18.2
	10+ Times	97	27.6	27.6	45.9
	6-10 Times	47	13.4	13.4	59.3
	I Don't Know	94	26.8	26.8	86.0
	None	49	14.0	14.0	100.0
	Total	351	100.0	100.0	

Fig. 3. Descriptive statistics of the importance of cybersecurity 2(Ics2).

These findings highlight the varying levels of awareness and response to cyber security risks among businesses. While some organizations take significant measures to protect their assets, others may not fully recognize the potential consequences or lack the necessary resources or knowledge to address these risks adequately. It emphasizes the importance of raising awareness, enhancing education, and promoting a proactive approach to cyber security across all businesses, regardless of their personal resource levels. Fig. 4 shows the correlation between Important of Cybersecurity (Ics1) and (Ics2).

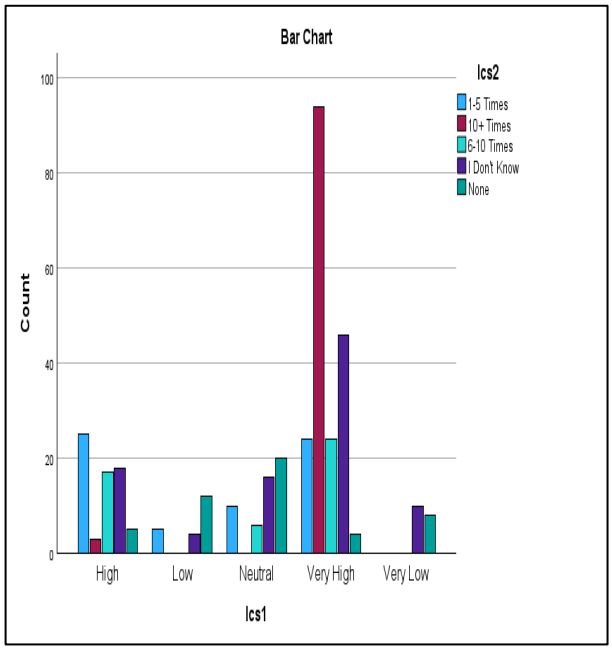


Fig. 4. Correlation between Importance of Cybersecurity (Ics1) and (Ics2).

B. Social Resources (Sr)

As part of the social analysis, the researchers formulated three items for the participants to gather insights into the perceptions surrounding cybersecurity practices in Saudi SMEs. The questions focused on competitor measures, customer data safety, and B2B suggestions regarding cyber security. The responses to these questions were consistent, with the highest number of participants strongly agreeing that cybersecurity practices are in high demand across all cases. This indicates a widespread recognition among participants that implementing robust cybersecurity measures is crucial in various aspects of business operations. Fig. 5 illustrates the correlation analysis of Social Resources (Sr1), (Sr2), and (Sr3).

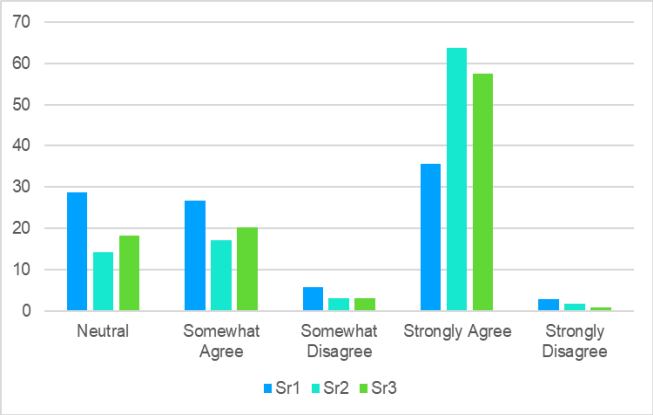


Fig. 5. Correlation analysis of social resources (Sr1), (Sr2) and (Sr3).

Firstly, when asked about their competitors' measures in implementing cybersecurity, many participants strongly agreed that such practices are highly sought after. This suggests that businesses in the same industry increasingly adopt cybersecurity measures to protect their valuable assets and sensitive information. Secondly, participants were asked about the safety of customer data. Once again, a considerable majority strongly agreed that ensuring customer data security is paramount. This underscores the growing awareness among businesses regarding the potential risks associated with data breaches and the need to safeguard customer information. Lastly, regarding B2B suggestions regarding cybersecurity aspects, participants strongly agreed that cybersecurity practices are in high demand. This indicates that Saudi SMEs engaging in B2B relationships are becoming more proactive in emphasizing the importance of cybersecurity and expecting their partners to implement robust measures to protect shared information and maintain a secure business environment.

Taken together, these findings paint a clear picture that in the context of Saudi SMEs, cyber security practices are highly regarded and in-demand from a social perspective. SMEs and their stakeholders recognize the significance of implementing effective security measures to mitigate the risks posed by cyber threats. This growing emphasis on cybersecurity highlights the need for continuous improvement and adaptation to address the evolving landscape of cyber threats in the SME's environment.

C. Organizational Resources (Or)

To perform organizational analysis, the researchers devised four items to gain insights into participants' perspectives on cybersecurity within their respective organizations. These questions focused on organizational-provided cybersecurity information, information-seeking behaviors, agreement that SMEs struggle to manage all the advice provided, and the impact of on-air cybersecurity advice on their work. Upon analyzing the responses, varied opinions across all question designs within this category were observed. Most participants tended to provide responses leaning towards the Neutral or Somewhat Agree spectrum. This indicates a lack of strong agreement or certainty regarding the organizational perspectives on cybersecurity or the scenarios presented.

Firstly, participants were asked about the availability of cybersecurity information provided by their organizations. The responses varied, with many participants expressing neutrality or a somewhat agreeable stance. This suggests that the participants may not view the organizational provision of cybersecurity information as highly reliable or effective. Secondly, participants were asked about their information-seeking behaviors regarding cybersecurity. Once again, the responses tended towards a neutral or somewhat agreeable position. This implies that participants may not actively seek out additional cybersecurity information beyond what is provided by their organization, or they may feel uncertain about the effectiveness of their information-seeking efforts. Fig. 6 displays the correlation analysis of Organizational Resources (Or1), (Or2), (Or3) and (Or4).

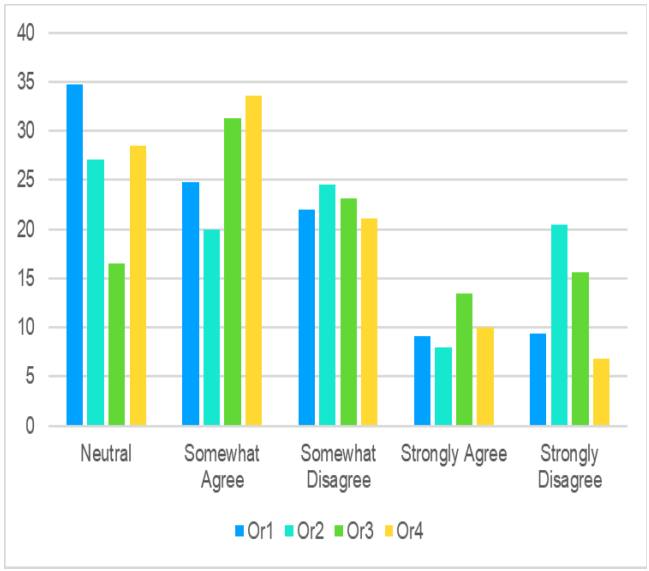


Fig. 6. Correlation analysis of organizational resources (Or1), (Or2), (Or3) and (Or4).

Furthermore, participants were presented with a statement regarding SMEs' ability to manage all the cybersecurity advice provided to them. The responses predominantly reflected a neutral or somewhat agreeable perspective. This suggests that participants may perceive SMEs to face challenges in effectively managing and implementing the abundance of

cybersecurity advice available. Lastly, participants were asked about the impact of on-air cybersecurity advice on their work. The responses once again leaned towards a neutral or somewhat agreeable standpoint. This indicates that participants may find it difficult to assess the relevance or applicability of on-air cybersecurity advice to their specific organizational context, potentially making their work more challenging.

The analysis of these independent responses reveals a lack of solid agreement or certainty among participants regarding their organizational perspectives on cybersecurity. The prevalence of neutral or somewhat agreeable responses suggests that participants may harbor reservations or uncertainties about the effectiveness, relevance, or practicality of their organizations' cybersecurity measures and advice. This highlights the importance of addressing these concerns and fostering more precise communication and understanding between organizations and employees regarding cybersecurity practices.

D. Personal Resources (Pr)

In personal resource analysis, the researchers posed five items to participants regarding their personal abilities in handling, managing, or implementing cybersecurity measures on their own. Additionally, they were inquired about their perceptions of personal information safety. The responses from participants yielded various results, mirroring the findings from the analysis of organizational resources. In many cases, participants expressed a somewhat agreeable or disagreeable stance toward the statements, indicating a potential lack of knowledge or familiarity with cybersecurity. Most participants may possess limited understanding or proficiency in this area, while a minority group demonstrated strong knowledge and capability to handle unforeseen cybersecurity situations.

Firstly, participants were asked about their personal abilities in managing cybersecurity. The responses revealed a range of opinions, with many participants leaning towards a somewhat agreeable or disagreeable standpoint. This suggests that many participants may lack the necessary skills or knowledge to handle cybersecurity matters on their own effectively. Fig. 7 illustrates the correlation analysis of Personal resources (Pr1), (Pr2), (Pr3), (Pr4) and (Pr5).

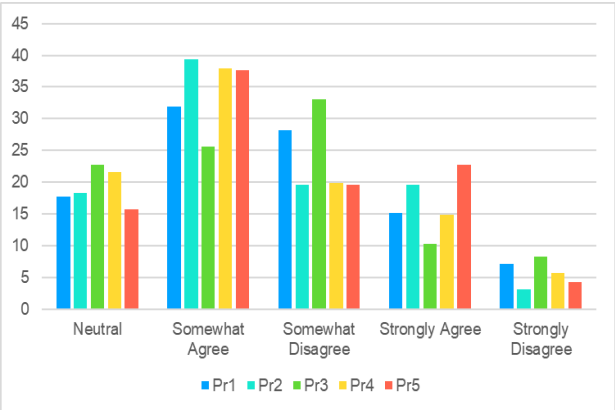


Fig. 7. Correlation analysis of personal resources (Pr1), (Pr2), (Pr3), (Pr4) and (Pr5).

Furthermore, participants were questioned about their personal information safety. The responses showcased a mixture of perspectives. Many participants tended to express a somewhat agreeable or disagreeable viewpoint, indicating that they may not have a strong confidence in the security of their personal information.

The analysis of these personal resource responses highlights a significant knowledge gap or lack of familiarity with cybersecurity practices among the participants. While a minority group demonstrated competence in handling cybersecurity matters, most appear to possess limited understanding or capability in this domain. This emphasizes the importance of raising awareness and providing education and resources to enhance individuals' personal cybersecurity skills and knowledge. We can collectively strengthen our overall cybersecurity posture by empowering individuals to protect their personal information better and effectively respond to cybersecurity threats.

E. Implementation of Cyber Controls and Precautions (Icsc)

To gain insight into the organizational setup, the researchers gathered information from the participants regarding the number of cybersecurity controls implemented by their SMEs and the extent to which their security policies covered various aspects. Upon analyzing the responses, the researchers observed a higher frequency of participants, indicating a lack of knowledge about these aspects. This suggests that many participants were unsure or unaware of the specific controls in place within their organizations. However, it is worth noting that the second and third most common responses indicated that their businesses had implemented between 1 and 5 controls and 10 or more controls, a positive indication of proactive security measures being taken. Fig. 8 displays the descriptive statistics of Implementing Cyber Controls and Precautions (Icsc1). Nevertheless, the fourth highest response rate was recorded for the option "none," implying that there may be certain aspects of cybersecurity that were overlooked or not adequately covered by our survey questions. This signifies a potential gap in assessing the participants' understanding of their organization's cybersecurity practices.

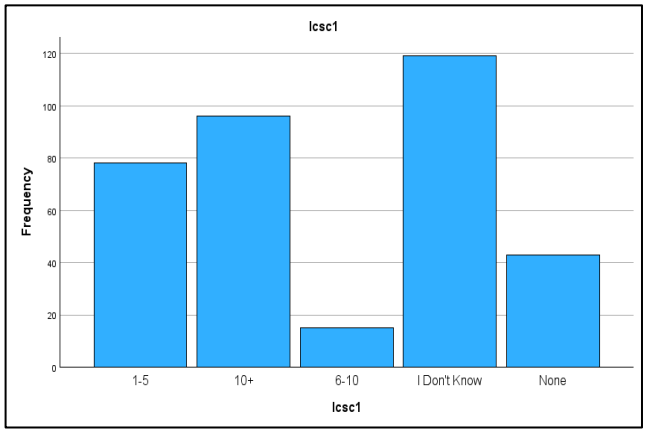


Fig. 8. Descriptive statistics of implementation of cyber controls and precautions (Icsc1).

It is crucial for the researchers to address this limitation and explore additional areas of concern that might have been missed in the initial survey design. Another significant finding is that most participants expressed a lack of awareness regarding their business's actions pertaining to cybersecurity. This highlights a concerning lack of knowledge or visibility among participants regarding the specific measures and initiatives their organizations undertake to mitigate cybersecurity risks. Fig. 9 shows the descriptive statistics of Implementing Cyber Controls and Precautions (Icsc2).

These findings underscore the importance of enhancing communication and awareness within organizations regarding cybersecurity practices. Providing employees with comprehensive information about the implemented controls is crucial, and ensuring they understand their roles and responsibilities in maintaining a secure environment is crucial.

Additionally, conducting more comprehensive assessments and addressing the areas of uncertainty can help organizations identify and rectify potential gaps in their cybersecurity strategies.

F. Reflective Measurement Model

In this research paper, the researchers analyzed several factors representing underlying factors presented in Table II. These factors include "Importance of Cybersecurity," "Social Resources," "Organizational Resources," "Personal Resources," and "Implementation of Cybersecurity Controls." The researchers also have a formative construct called "Cyber Situational Awareness."

To evaluate the reflective constructs, we examined the item loadings of their indicators. Item loadings indicate the strength of the relationship between each indicator and its respective factor.

Most of the item loadings were slightly below the recommended threshold of 0.708. (Hair Jr et al., 2021) suggested a relatively strong association between the indicators and their factors. However, five items, specifically Or2, Or3, Or4, Pr2, and Pr5, had item loadings below the threshold.

The researchers also assessed the internal consistency reliability of the reflective factors using composite reliability and Cronbach's alpha. Composite reliability values measure the reliability of the items in capturing the factors, while Cronbach's alpha estimates the internal consistency of the item set. Although most composite reliability values were below the satisfactory threshold of 0.70, none were problematically high (above 0.95) [27].

The researchers examined the average variance extracted (AVE) for each factor to evaluate convergent validity. AVE represents the proportion of variance in the items explained by the factor. Three factors had AVE values above the acceptable threshold of 0.50, indicating that they explain significant variance in the items [28]. However, "Organizational Resources" and "Personal Resources" had AVE values below 0.50, suggesting that they explain a relatively smaller proportion of variance in their items.

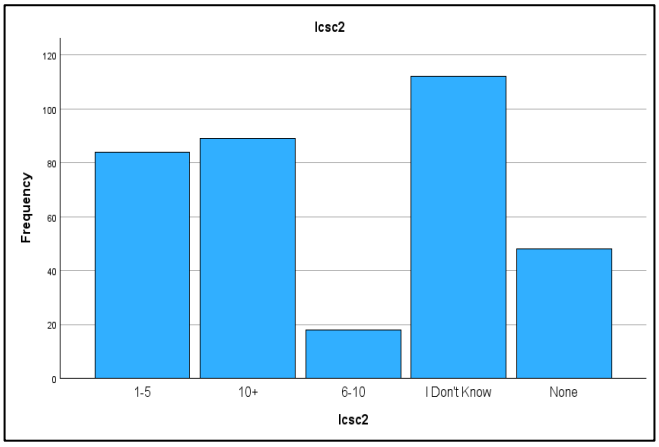


Fig. 9. Descriptive statistics of implementation of cyber controls and precautions (Icsc2).

TABLE II. RESULT SUMMARY FOR REFLECTIVE MEASUREMENT MODEL

Factor	Items	Loadings	Indicator reliability	AVE	Composite reliability	Cronbach's alpha	Discriminant validity
Importance of Cybersecurity	Ics1	.829	.702	.6408	0.69576	.415	Yes
	Ics2	.771	.731				
Social resources	Sr1	.718	.706	.5430	0.659471	.324	Yes
	Sr2	.765	.692				
	Sr3	.727	.694				
Organizational resources	Orl	.756	.694	.4608	0.611686	.299	Yes
	Or2	.648	.699				
	Or3	.680	.692				
	Or4	.624	.699				
Personal resources	Pr1	.719	.699	.4783	0.68684	.354	Yes
	Pr2	.669	.693				
	Pr3	.736	.706				
	Pr4	.748	.696				
	Pr5	.571	.690				
Implementation of cyber security controls	Icsc1	.898	.725	.8227	0.884216	.542	Yes
	Icsc2	.916	.724				

G. Result for Formative Construct Significance Testing

Table III displayed the results of a statistical analysis conducted on a formative factor named Cyber situational awareness "Csa category." This factor comprises four formative items: Csa1, Csa2, Csa3, and Csa4. Examining the outer weights column, it was observed that the relative importance of each item in shaping the overall formative factor. Higher values indicate a more significant influence of the respective item on the factor. Notably, Csa1 possesses the highest outer weight of 0.766, signifying its prominent role in defining the factor compared to the other items. Moving on to the T-value column, the statistical significance of the relationship between each item and the formative factor was assessed. Higher T-values indicate a more substantial relationship. All items exhibit considerable T-values, with Csa1 boasting the highest at 72.801. These results suggest that all items significantly contribute to the formative factor. The P-value column enables evaluating the level of significance associated with each item. P-value lower than the chosen significance level (typically 0.05) indicates statistical significance, implying that the indicator's relationship with the factor is unlikely to occur randomly [28]. In this analysis, all items demonstrate exceptionally low P-values (0.000), confirming their statistical significance.

In the 95% BCa confidence interval column, the researchers encountered a range that estimates the true relationship between each item and the formative factor. Narrower intervals indicate higher precision in estimating these relationships. Notably, all items exhibit relatively tight confidence intervals, suggesting high precision in capturing their relationships with the factor. The "Significance" column provides a succinct indication of the statistical significance of each item. In this case, all items are marked as "Yes," signifying their significant impact on the formative factor. These findings underscore the substantial importance of all four items (Csa1, Csa2, Csa3, and Csa4) in defining the formative factor. The high outer weights, significant T-values, low P-values, and narrow confidence intervals collectively provide strong evidence of the meaningful contribution of each item. Consequently, these results significantly enhance our understanding of the investigated phenomenon.

H. Results of the Structural Model Path Co-Efficient

In Table IV the hypothesis column represents the specific hypotheses being tested in the analysis. The hypotheses are labeled as H1, H2, H3, H4, and H5, indicating different relationships between the dependent and independent variables. Where D-I column specifies the dependent and independent variables involved in each hypothesis. It indicates the direction of the relationship being examined. For example, H1: Cyber situational awareness - Importance of cyber security (Csa-Ics) indicates that the independent variable Cyber situational awareness "Csa" is hypothesized to influence the dependent variable Importance of cyber security "Ics." The path coefficients in the hypothesis table indicate the strength and direction of the relationships between variables. Positive coefficients in H1 and H2 suggest that increasing Csa positively impacts Ics and social resources (Sr), respectively. The negative coefficient in H3 indicates that an increase in Icsc is associated with decreased organizational resources (Or). However, the relationships in H4 and H5 are not statistically significant, suggesting that the coefficients may not accurately represent the true impact.

Overall, the path coefficients provide valuable insights into the relationships between variables, highlighting significant associations and the need for further investigation in non-significant cases.

T-value provides information on the statistical significance of each path coefficient. The T-values indicate the degree to which the observed path coefficients deviate from zero. For H1, the T-value of 1.999 indicates a statistically significant relationship between Csa and Ics. Similarly, H2 shows a highly significant relationship between Csa and Sr, with a T-value of 4.837. In contrast, H3 reveals a statistically significant negative relationship between Icsc and Or, as indicated by the T-value of -2.492. However, H4 does not exhibit a significant relationship between Icsc and Pr, with a T-value of -1.010.

TABLE III. RESULT SUMMARY FOR FORMATIVE CONSTRUCT SIGNIFICANCE TESTING

Formative Factor	Formative items	Outer weights	T value	P value	95%BCa confidence interval	Significance P<.05
Cyber situational awareness (Csa_category)	Csa1	.766	72.801	.000	4.231 – 4.456	Yes
	Csa2	.651	34.550	.000	2.826 – 3.154	Yes
	Csa3	.725	44.251	.000	3.336 – 3.632	Yes
	Csa4	.736	43.310	.000	3.336 – 3.641	Yes

TABLE IV. SIGNIFICANCE TESTING RESULTS OF THE STRUCTURAL MODEL PATH CO-EFFICIENT

Hypothesis	D-I	Path coefficients	T value	P value	95% confidence intervals	Significance P<.005	R square
H1	Csa-Ics	.106	1.999	.046	.003 - .353	Yes	.011
H2	Csa-Sr	.251	4.837	.000	.143 - .340	Yes	.063
H3	Icsc-Or	-.132	-2.492	.013	-.171 - -.020	Yes	.017
H4	Icsc-Pr	-.054	-1.010	.313	-.107 - .034	No	.003
H5	Icsc-Csa	.079	1.488	.138	-.023 - .163	No	.006

The P-value column complements the T-values by providing the level of significance associated with each path coefficient. In H1, the P-value of 0.046 confirms the statistically significant relationship between Csa and Ics. H2 demonstrates a highly significant relationship between Csa and Sr, indicated by the P-value of 0.000. H3 exhibits a statistically significant relationship between Icsc and Or, with a P-value of 0.013. Conversely, H4 suggests no significant relationship between Icsc and Pr, with a P-value of 0.313.

The 95% confidence intervals offer a range within which the true population parameter is likely to fall. For H1, the confidence interval of 0.003 to 0.353 reinforces the statistical significance of the relationship between Csa and Ics. Similarly, the narrow interval of 0.143 to 0.340 for H2 supports the significance of the relationship between Csa and Sr. H3's confidence interval of -0.171 to -0.020 suggesting a significant negative relationship between Icsc and Or. However, caution is necessary when interpreting H4's confidence interval of -0.107 to 0.034, as the relationship between Icsc and Pr is not statistically significant. Fig. 10 shows the results of PLS-SEM analysis.

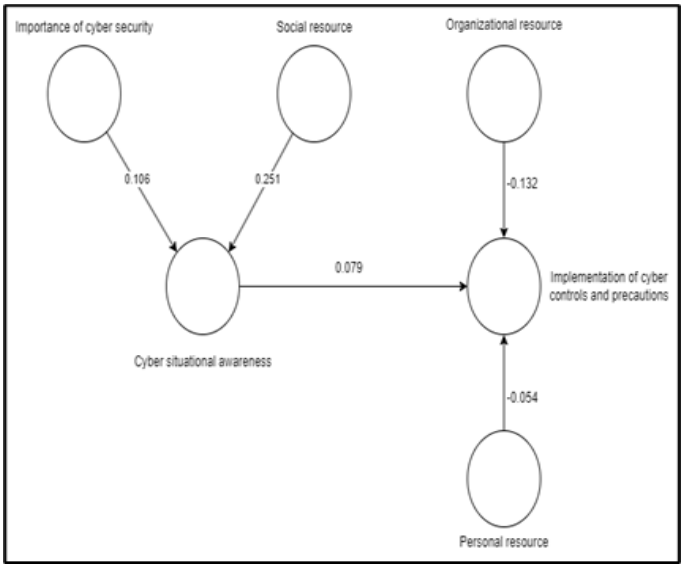


Fig. 10. Results of the PLS-SEM analysis.

I. *Impact of Negative Path Coefficient Values among Hypotheses*

Hypothesis H3 examines the relationship between the implementation of cyber security controls and precautions (Icsc) and organizational resources (Or), while Hypothesis H4 explores the relationship between Icsc and personal resources (Pr). Although both hypotheses have negative path coefficients, H3 (-0.132) is statistically significant, whereas H4 (-0.054) is not. Here is a detailed justification for these findings:

Hypothesis H3: The negative path coefficient (-0.132) in H3 suggests that as the implementation of cyber security controls and precautions (Icsc) increases, the availability or allocation of organizational resources (Or) decreases. Several reasons contribute to this relationship:

- Resource Reallocation: When organizations prioritize the implementation of cyber security controls and precautions, they often allocate resources, such as financial investments, personnel, and technology, to support these measures. This reallocation of resources can lead to reduced availability or allocation of resources for other organizational activities, resulting in a negative relationship between Icsc and Or.
- Trade-offs: Implementing cyber security controls and precautions often involves making trade-offs in resource allocation. Organizations may need to invest in security technologies, employee training, or hiring specialized personnel, which could lead to limited resources for other organizational needs. As a result, the negative path coefficient indicates that an increase in Icsc is associated with decreased available resources.
- Efficiency and Effectiveness: Implementing cyber security controls and precautions effectively requires utilizing resources efficiently. Organizations with limited resources may face challenges in implementing comprehensive security measures, leading to increased vulnerability and organizational risk. Consequently, the negative relationship between Icsc and or may arise from the difficulty in maintaining sufficient resources for cyber security and other organizational functions.

Hypothesis H4: Although H4 suggests a negative relationship between Icsc and personal resources (Pr), the non-significant p-value (0.313) indicates that this relationship is not statistically significant in the given sample. Several reasons may contribute to this result:

- Individual Factors: Personal resources encompass an individual's capabilities, knowledge, skills, and access to relevant information. The negative path coefficient implies that, in theory, as Icsc increases, personal resources (Pr) should decrease. However, in the context of the specific sample or survey respondents, other factors such as individual characteristics, experiences, and attitudes may overshadow the impact of Icsc on personal resources.
- Complexity of Personal Resources: Personal resources in cybersecurity could include individual knowledge, expertise, awareness, and adherence to security practices. Assessing personal resources accurately can be challenging, as it involves subjective factors and self-perception. Measurement limitations or insufficient sensitivity in capturing personal resource variations may contribute to the non-significant findings.
- Indirect Relationship: The relationship between Icsc and personal resources may be indirect, mediated by other variables not considered in the hypothesis. Factors such as organizational culture, training programs, or information-sharing practices might influence personal resources, moderating the relationship between Icsc and Pr. Not accounting for these mediating factors could result in a non-significant direct relationship between Icsc and Pr.

By considering a more extensive and diverse sample, conducting qualitative investigations, or refining the measurement instruments, researchers can better understand the relationship between Icsc and personal resources in the specific context of cyber security.

V. Conclusion and Future Work

Based on the research findings, several fruitful insights can be drawn that shed light on the relationships between cybersecurity items within organizations. The study identified a significant positive relationship between Cyber Situational Awareness (Csa) and Implementation of Cybersecurity Controls (Icsc), suggesting that enhancing awareness can contribute to better control implementation. In simple terms, Saudi Arabia should focus on raising cyber situational awareness among SMEs so that the SMEs have implementation of cyber security controls in their companies. Additionally, a significant positive relationship was found between Csa and Social Resources (Sr), emphasizing the importance of promoting awareness to enhance organizational resilience. Saudi Arabia should promote awareness among the SMEs so that the SMEs can have better organizational resilience to different cyber-attacks. Furthermore, a significant negative relationship between Icsc and Organizational Resources (Or) indicates that robust control implementation can potentially reduce organizational risk. SMEs in Saudi Arabia should implement robust cyber security mechanisms in their companies to reduce organizational risk. However, no significant relationships were found between Icsc and Personal Resources (Pr). In the context of Saudi Arabia, implementing cyber security does not rely on the personal resources of the employees of the SMEs. These findings provide valuable insights for organizations seeking to improve their cybersecurity practices.

While the study provides valuable insights, some limitations should be acknowledged. Firstly, the effect sizes observed were relatively small, suggesting that there may be other factors at play that were not considered in the analysis. Additionally, the study relied on self-reported data, which may introduce biases and inaccuracies. Data was collected from different sectors. The sample size and composition could also affect the generalizability of the findings. Furthermore, the study focused on specific variables and did not consider potential mediating or moderating factors. These limitations should be taken into account when interpreting the results.

Future work can focus on several areas to address the limitations and extend the research. Firstly, qualitative research can provide a deeper understanding of the underlying mechanisms behind the identified relationships. This qualitative exploration can help identify specific aspects of Cyber Situational Awareness and Personal Resources that impact control implementation most. Additionally, future studies can delve deeper into the dimensions of Cyber Situational Awareness, such as knowledge of threats and risk assessment capabilities, to develop targeted interventions and comprehensive training programs. Exploring additional contributors to Organizational Resources, such as human factors, supply chain vulnerabilities, and emerging technological threats, can also provide a more holistic understanding of risk management. Furthermore, investigating the role of communication strategies, organizational culture, and individual cognitive biases can assist in designing effective risk communication campaigns and tailored training programs. Integrating technical controls with organizational and behavioral aspects in comprehensive frameworks can enhance the effectiveness of cybersecurity initiatives. These future directions can further advance the understanding and practice of cybersecurity within organizations.

Acknowledgment

Thanks to the Saudi Electronic University for all support provided for this work.

References

[1] Karen Renaud and Jacques Ophoff, “A cyber situational awareness model to predict the implementation of cyber security controls and precautions by SMEs,” Organizational Cybersecurity Journal: Practice, Process and People, pp. 24–46, 2021.

[2] F. Alharbi et al., “The impact of cybersecurity practices on cyberattack damage: The perspective of small enterprises in Saudi Arabia,” Sensors, vol. 21, no. 20, p. 6901, 2021.

[3] K. Luan, R. Halvorsrud, and C. Boletsis, “Evaluation of a Tool to Increase Cybersecurity Awareness Among Non-experts (SME Employees),” in Proceedings of the International Conference on Information Systems Security and Privacy (ICISSP), 2023, pp. 509–518.

[4] J. M. Archibald and K. Renaud, “Refining the pointer ‘human firewall’ pentesting framework,” Information & Computer Security, vol. 27, no. 4, pp. 575–600, 2019.

[5] T. Kokkonen, “Anomaly-based online intrusion detection system as a sensor for cyber security situational awareness system,” Jyväskylä studies in computing, no. 251, 2016.

[6] M. R. Endsley, Designing for situation awareness: An approach to user-centered design. CRC press, 2016.

[7] M. Husák, T. Jirsík, and S. J. Yang, “SoK: Contemporary issues and challenges to enable cyber situational awareness for network security,” in Proceedings of the 15th International Conference on Availability, Reliability and Security, 2020, pp. 1–10.

[8] G. Erdogan, R. Halvorsrud, C. Boletsis, S. Tverdal, and J. B. Pickering, “Cybersecurity Awareness and Capacities of SMEs,” 2023.

[9] N. Naik, P. Jenkins, P. Grace, and J. Song, “Comparing attack models for it systems: Lockheed martin’s cyber kill chain, mitre att&ck framework and diamond model,” in 2022 IEEE International Symposium on Systems Engineering (ISSE), IEEE, 2022, pp. 1–7.

[10] M. K. Ahn, Y. H. Kim, and J.-R. Lee, “Hierarchical multi-stage cyber attack scenario modeling based on G&E model for cyber risk simulation analysis,” Applied Sciences, vol. 10, no. 4, p. 1426, 2020.

[11] D. S. Rodriguez-Bermejo, R. D. Medenou, R. P. de Riquelme, J. M. Vidal, F. Torelli, and S. L. Sánchez, “Evaluation methodology for mission-centric cyber situational awareness capabilities,” in Proceedings of the 15th International Conference on Availability, Reliability and Security, 2020, pp. 1–9.

[12] N. Neshenko, C. Nader, E. Bou-Harb, and B. Furht, “A survey of methods supporting cyber situational awareness in the context of smart cities,” J Big Data, vol. 7, no. 1, pp. 1–41, 2020.

[13] M. N. Nafees, N. Saxena, A. Cardenas, S. Grijalva, and P. Burnap, “Smart grid cyber-physical situational awareness of complex operational technology attacks: A review,” ACM Comput Surv, vol. 55, no. 10, pp. 1–36, 2023.

[14] A. Georgiadou, S. Mouzakitis, and D. Askounis, “Assessing mitre att&ck risk using a cyber-security culture framework,” Sensors, vol. 21, no. 9, p. 3267, 2021.

[15] Small and Medium Enterprises General Authority, “Saudi SMEs official definition,” https://www.my.gov.sa.

[16] General Authority for Statistics, “Importance of small and medium enterprises (SMEs) to Saudi economy,” https://www.stats.gov.sa/en.

[17] A. Al-Tit, A. Omri, and J. Euchi, "Critical success factors of small and medium-sized enterprises in Saudi Arabia: Insights from sustainability perspective," Adm Sci, vol. 9, no. 2, p. 32, 2019.

[18] O. M. Elhassan, "Obstacles and problems facing the financing of small and medium enterprises in KSA," Journal of Finance and Accounting, vol. 7, no. 5, pp. 168–183, 2019.

[19] F. Alharbi et al., "on cyberattack damage: The perspective of small enterprises in Saudi Arabia," Sensors, vol. 21, no. 20, p. 6901, 2021.

[20] V. V. Muthuswamy, "Cyber Security Challenges Faced by Employees in the Digital Workplace of Saudi Arabia's Digital Nature Organization," International Journal of Cyber Criminology, vol. 17, no. 1, pp. 40–53, 2023.

[21] M. , Bada and J. R. Nurse, "The social and psychological impact of cyberattacks," Emerging cyber threats and cognitive vulnerabilities, pp. 73–92, 2020.

[22] M. S. Satar and G. Alarifi, "Factors of E-business adoption in small and medium enterprises: evidence from Saudi Arabia," Hum Behav Emerg Technol, vol. 2022, 2022.

[23] M. Hassan, K. Saeedi, H. Almagwashi, and S. Alarifi, "Information Security Risk Awareness Survey of Non-governmental Organization in Saudi Arabia," in The International Research & Innovation Forum, Springer, 2022, pp. 39–71.

[24] A. AbuAbid, M. Rahim, and H. Scheepers, "Experienced Benefits and Barriers of e-Business Technology Adoption by SME suppliers," vol. 2011, p. 11, 2011, doi: 10.5171/2011.7917780.

[25] J. Pallant, SPSS survival manual: A step by step guide to data analysis using IBM SPSS. McGraw-hill education (UK), 2020.

[26] F. Ridzuan and W. M. N. W. Zainon, "A review on data cleansing methods for big data," Procedia Comput Sci, vol. 161, pp. 731–738, 2019.

[27] A. Diamantopoulos, M. Sarstedt, C. Fuchs, P. Wilczynski, and S. Kaiser, "Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective," J Acad Mark Sci, vol. 40, pp. 434–449, 2012.

[28] A. Purwanto, "Partial least squares structural squation modeling (PLS-SEM) analysis for social and management research: a literature review," Journal of Industrial Engineering & Management Research, 2021.

Network Security Detection Method Based on Abnormal Traffic Detection

Tao Xiao*, Yang Ke, Hu YiWen, Wang HongYa
State Grid Jiangxi Electric Power Co, Ltd. Training Center, Nanchang 330013, China

***Abstract*—To discover potential risks and vulnerabilities in the network in time and ensure the safe operation of the network, a network security detection method based on abnormal traffic detection is studied. Construct network security detection architecture from several aspects, including the front-end interface module, control center module, network status extraction module, anomaly detection module, alarm module, and database module. Use NetFlow technology to capture network traffic from the network in the form of flow, and use the KNN algorithm in the traffic filtering submodule to filter network traffic packets and eliminate duplicate traffic data. After filtering traffic, the traffic data is transmitted to the feature selection sub-module. PCA-TS algorithm is used to reduce the dimension of the network traffic data and select the network traffic characteristics, and then it is input into the SVM classifier. The improved SVM multi-classification algorithm is used to classify normal and abnormal traffic, complete abnormal traffic detection, and achieve network security detection. Experimental results show that the time for feature selection of this method does not exceed 3.0s, and the G score in the detection process also remains above 0.70, indicating that this method has strong network security detection capability.**

***Keywords*—Abnormal traffic; network security detection; data dimensionality reduction; flow characteristics; traffic capture; alarm module**

I. INTRODUCTION

Network security detection refers to a comprehensive security assessment and inspection of the computer network system to find potential security risks and vulnerabilities and take corresponding measures to protect the security and integrity of the network system [1]. Its significance lies in preventing potential threats, protecting important data, maintaining business continuity, improving user trust and complying with regulatory requirements. By preventing potential threats, vulnerabilities and weaknesses in the network system can be found and repaired in time to avoid security attacks [2]. At the same time, security incidents such as hacker intrusion, data leakage and malware infection can be avoided by timely finding and solving security problems [3]. Network security detection can help enterprises and individuals protect important business and personal data. Data security and confidentiality can be ensured by detecting vulnerabilities and risks in the network system [4]. It can also ensure the normal operation of the network, reduce business interruption and loss caused by security vulnerabilities and attacks, find and repair the vulnerabilities in the network system in time, and ensure the continuity and stability of the business [5]. By improving network security through network security detection, users can use online services more confidently without worrying about personal information leakage or account theft [6]. Therefore, network security detection is of great significance.

With the popularization of computer network applications and services, the number of Internet users continues to increase, and the demand for Internet information sharing continues to expand [7]. The threat of network security attacks has become more serious, and network anomaly detection has become an increasingly important task in network security research [8]. There are many reasons for network exceptions, such as network overload, worm network intrusion, routing policy modification, and distributed denial of service attacks. Network traffic anomaly is the most common threat in network anomaly. Abnormal network traffic may reduce the central network speed or even cause network paralysis, which will cause serious damage to the network environment [9]. Weihai caused by abnormal network traffic is generally characterized by bandwidth occupation, network blocking, failure to send normal information on time, network packet loss, etc. [10]. For computer systems, servers, and clients, the harm caused by abnormal network traffic is shown as occupying a large amount of memory space, and data responses are transmitted to the server normally with different responses [11]. Many scholars have studied network security detection methods to solve these threats to network security and achieve timely hazard warnings and other functions.

Wozniak M et al. [12] studied the cyclic neural network model for threat detection of the Internet of Things and network malware. This method classifies the information in the network through the cyclic neural network to detect malicious threat information, but this method cannot achieve security alarm in the detection process, and the detection results of multiple attacks are not clear enough; Steno P et al. [13] studied uses deep learning to detect threat objects in security screening. This method uses deep learning network and cross-entropy loss calculation to realize risk screening in network objects, but this method cannot detect the degree of flow fluctuation in the network, resulting in a small detection range; Gaber T et al. [14] studied an injection attack detection method for intelligent Internet of Things applications using machine learning. This method detects network attack traffic in smart cities, uses constant removal and recursive feature elimination methods to achieve feature selection, and uses machine learning classifiers to classify attack traffic. Although the accuracy of this method is as high as 99%, this method needs a lot of time in feature extraction and detection.

The abnormal traffic detection method identifies and detects abnormal situations inconsistent with normal network traffic behavior by analyzing the network traffic changes. This

detection method can help find abnormal activities in the network, such as network attacks, malware propagation, data leakage, etc., to ensure the network's security and stability [15]. There are many methods to detect abnormal network traffic, such as traffic analysis methods, machine learning methods, etc. The traffic analysis method determines whether the traffic is abnormal by capturing and parsing the network data packets, analyzing the source address, destination address, protocol type and other information of the data packets, as well as the size, frequency and other characteristics of the data packets. For example, an exception may exist if an IP address sends many data packets quickly or a port receives abnormally large data traffic. The machine learning method uses algorithms to analyze and model network traffic to identify abnormal traffic. Machine learning can automatically identify abnormal data packet size, abnormal connection behavior, etc. [16] by learning the characteristics of normal traffic behavior. Therefore, this paper proposes an abnormal traffic detection method based on support vector machine (SVM). The innovation lies in the construction of a network security detection architecture, which uses NetFlow technology to capture traffic data from the network and uses KNN algorithm to remove duplicate data. In the feature selection submodule, PCA-TS algorithm is used to reduce the dimensionality of network traffic and select features. An improved SVM multi classification algorithm is used to classify normal and abnormal traffic, achieving efficient abnormal traffic detection and network security detection.

II. Design of Network Security Detection Method

A. Construction of Network Security Detection Architecture

To improve the network security and operation status, this paper studies the network security detection architecture, shown in Fig. 1. The architecture is divided into a foreground interface, control center, network status extraction, anomaly detection, alarm, and database modules. Feature selection and anomaly detection modules are the architecture's main parts.

The specific contents of the network security detection architecture are as follows:

1) Front-end interface: The main functions of the front-end interface module include network topology node display function, user information display and user login function, system working status display function, log information display function and abnormal flow alarm display function. The main function of the foreground interface is to provide users with a good and beautiful interface display function and provide users with a good interaction function. The network topology node display function can view the current network's working nodes and the current network's topology. The alarm display function can more intuitively see the attacked node. The log information display function can view the system log, including user login information and running status record information on the architecture.

2) Control center: The main functions of the control center module include user authority management function, configuration function, alarm function, task allocation function, etc. The user authority management function can ensure that the authority is not abused. Only super users have administrator authority to modify sensitive content such as network configuration information. The alarm function makes it possible to quickly give feedback to the foreground interface to prompt network exceptions when the detection architecture is abnormal and even locate the attacked node in time. The task allocation function mainly includes switching between the network status extraction function, normal detection status, and log viewing function.

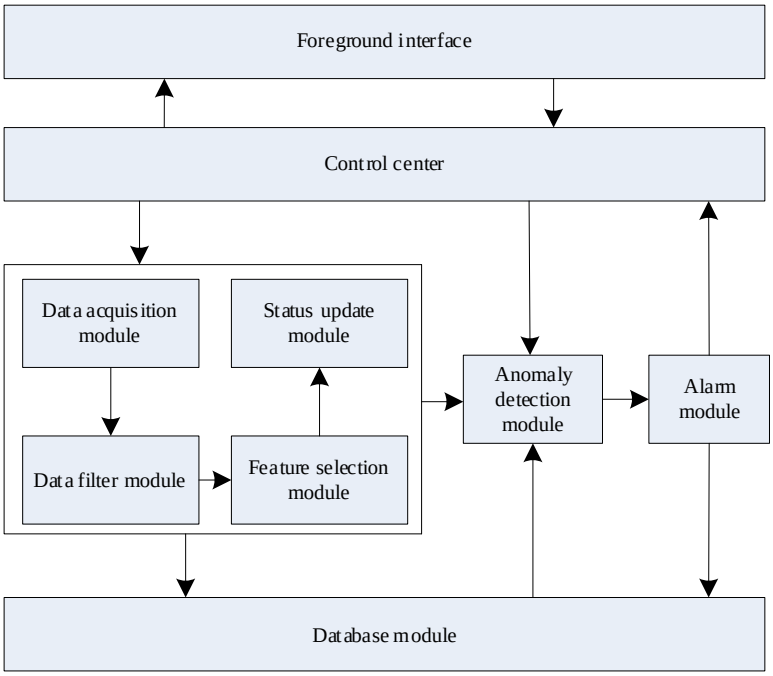


Fig. 1. Overall design of network security detection architecture.

3) Feature selection module: The main function of the feature selection module is to capture network traffic through the function of collecting data and selecting the features of the captured traffic to achieve feature extraction. In the safe working mode, the module saves the extracted feature information to the database. In the normal working mode, the extracted features are transmitted to the abnormal flow detection module to judge the abnormal flow of the network. In the update mode, the extracted features are updated to the database. Feature selection module is one of the most important functional modules, including four sub-modules: traffic capture module, traffic filtering module, feature selection module and network status update module.

The traffic capture submodule's main function is to capture each node's data packets and then save the data packets to the packet queue. The main function of the traffic filtering sub-module is to filter the redundant content and duplicate content captured and then wait for the feature selection sub-module to extract features. The feature selection sub-module is the core module of the architecture. It receives the data package of the filter sub-module, selects its features, and saves the extracted state information to the database. The status update sub-module mainly updates the recorded standard status information.

4) Anomaly detection module: The anomaly detection module is mainly responsible for judging whether the network has been attacked by abnormal traffic and passing the detection results to the alarm module. At the same time, the detection results are saved in the database as log management, which can facilitate viewing historical record information. The algorithm implements this module. At the same time, this module is one of the most important functional modules and is the key judge of whether an intrusion occurs.

5) Alarm module: The main function of the alarm module is to send an alarm to the user when the abnormal flow detection module judges that an intrusion has occurred. The alarm module receives the result of the abnormal flow detection module. If the result is true, the alarm information will be written into the database and sent to the control center simultaneously. The alarm can be sent out an alarm tone or pop up a window light on the foreground interface.

6) Database module: The main function of the database module is to record various types of information, including log information, user information records, alarm information records, etc. The database module is the information center of the entire architecture. All data extraction, information exchange and record-keeping between modules are completed in this module. The database module requires the physical support of the database software, or it can be deployed to an independent server separately.

B. Network Traffic Capture

Based on NetFlow technology, the traffic capture sub-module under the feature selection module in the network security detection architecture captures network traffic in flow. Use the V9 NetFlow technology as a sniffer or probe in the network to transmit the traffic records to a data collector with a specified IP address. The output package format of NetFlow V9 is shown in Fig. 2. FlowSet represents the collection of traffic records. The Template Flowset in Fig. 2 is the template for subsequent data records, enhancing network traffic records' flexibility.

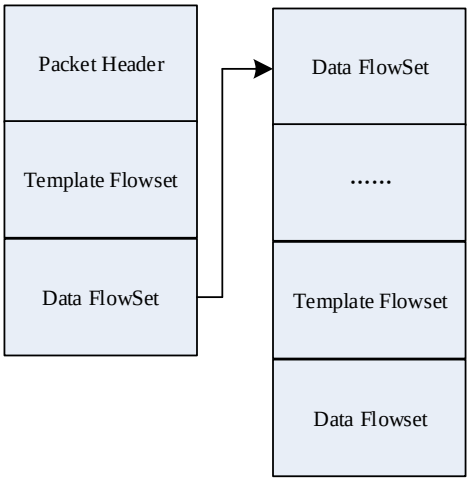


Fig. 2. Analysis of the output package format of NetFlowV9.

Encapsulates the network traffic records into UDP packets and transmits them to the collector with UDP protocol to ensure high efficiency when transmitting a large number of traffic records. To prevent NetFlow from generating a large amount of data and causing network congestion, a dedicated link is designed for the traffic record output to the collector in the congestion-sensitive network. When the collector cannot be placed at the router's next hop or the transmission link cannot be exclusive to NetFlow, a special link needs to be designed to handle the large amount of data NetFlow generates.

C. Traffic Filtering of KNN Classification Algorithm

After capturing the network traffic data, due to the huge amount of network traffic data [17], and some botnet traffic and duplicate content, effective measures must be taken to filter the traffic. Through reasonable filtering means, the traffic data can be more conducive to subsequent network security detection [18]. This paper uses the KNN algorithm to filter traffic in the traffic filtering sub-module under the feature selection module.

1) Filtering analysis: KNN algorithm realizes classification by measuring the distance between different eigenvalues to achieve a data filtering effect. If most of the k, that the most similar samples belong to a category, the samples also belong to that category. The most similar definition here is that the eigenvalue of the sample is the nearest in the feature space. Among K is an integer less than or equal to 20. In the KNN algorithm, the most similar samples are correctly divided into corresponding classes [19]. When this method is used for classification, the possible classification of samples is determined according to the classification of the nearest one or several samples. Fig. 3 shows the operation mode of the algorithm.

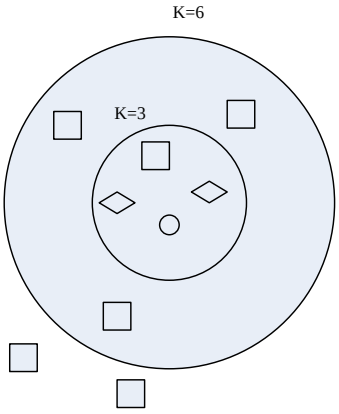


Fig. 3. Analysis of KNN algorithm operation mode.

It can be seen from Fig. 3 that when $K = 3$, the grey filled circle shall belong to the hollow triangle type. When $K = 6$, the grey filled circle shall be classified as a hollow square. Therefore, it can be concluded that the selection of values K affects the filtering results; the optimal value can make the filtering effect the best.

2) Flow filtration: Due to the large difference between normal traffic and abnormal traffic caused by attacks, the external parameters of abnormal traffic can be filtered out from the external parameters of mixed traffic, including normal and abnormal traffic, using the KNN method to achieve faster traffic data processing [20]. The implementation process is:

The obtained labelled data (packet length, URL length) and unlabeled data (packet length, URL length) are regarded as vectors, and their Euclidean distances are calculated. European distance can be calculated by Formula (1):

$$dist(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (1)$$

In Formula (1), $dist(X,Y)$ represents a data set between European distance of X and Y; x_i and y_i respectively represent i of the labeled data and unlabeled data; n indicates the number of traffic data. This formula can measure the absolute distance between points in multidimensional space.

3) The first one closest to the unlabeled data is counted K the labelled data with the most occurrences among the data will be marked with the same label as the unlabeled data. Obtain a batch of labelled data, which can support subsequent work in terms of data quantity.

D. Design of Network Traffic Feature Selection Method

Real network traffic contains many feature attributes, and the existing anomaly traffic detection methods based on feature analysis cannot meet the real-time requirements of high-dimensional feature analysis [21], [22]. Therefore, when this paper filters the traffic data through the KNN algorithm, the feature selection submodule uses the traffic feature selection algorithm based on principal component analysis (PCA) and tabu search (TS) to conduct feature reduction and near-optimal feature subset selection for high-dimensional features through PCA-TS, providing reliable feature data for subsequent abnormal traffic detection.

1) Dimension reduction of traffic data: The dimensionality of the traffic data filtered by the KNN algorithm is reduced to facilitate the subsequent feature selection [23]. Principal component analysis is an effective method of analyzing data in statistics, mainly used for feature extraction and data dimension reduction. The idea is to reduce the dimension of a data set with high dimension and correlation by using the feature space transformation of statistical properties of the data set [24]. PCA transforms the original space into a new principal component space, and the principal components are unrelated.

Assume that the network traffic data set contains N samples $N = \{x_1, x_2, \ldots, x_m\} \in R^n$, where,$R^n$is the feature space,$m$ is the characteristic dimension. Find variable space $Z = \{z_1, z_2, \ldots, z_k\}$, satisfied $k < m$ and $cov(z_i, z_j) = 0$, through transformation k new variables Z can represent most of the information of m original variables X, as shown in Formula (2):

$$Z = \kappa N dist(X,Y) \qquad (2)$$

In Formula (2),κ is a one $m \times m$ orthogonal matrix and is the covariance matrix of the eigenvalue matrix of data samples $C = \frac{1}{N}\sum_{i=1}^{N}(x_i - u)(x_i - u)^T$, where, $u = \frac{1}{N}\sum_{i=1}^{N} x_i$. Therefore, it is transformed into solving the eigenproblem as shown in Formula (3):

$$\lambda_i = CPZ \qquad (3)$$

In Formula (3), λ_i is characteristic value of C, P is the corresponding eigenvector. Principal component analysis selects several characteristic values with a high contribution rate $\lambda_i$corresponding eigenvector Pas the principal component, to achieve the purpose of dimension reduction. The characteristic contribution rate is shown in Formula (4):

$$\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i} = \frac{\sum_{i=1}^{m}\lambda_i}{n} \geq R \qquad (4)$$

In Formula (4), R is the threshold value of the feature contribution rate, feature dimension m is selected according to R to determine the general choice R85%~95%. When using PCA for analysis, different variables in the data often have different dimensions, leading to a large difference in the dispersion of the values of each variable, thus affecting the calculation accuracy. To eliminate the possible impact of different dimensions, the variables need to be standardized first, and then the dimension can be reduced by PCA.

2) Feature selection based on the tabu search algorithm: After the dimensionality reduction of traffic data through the PCA algorithm, feature selection can be done. Tabu Search (TS) algorithm is a heuristic global optimization search method, which obtains the global optimal solution by marking the searched local optimal solution and avoiding repeated search in the iterative calculation [25]. The main idea of the algorithm is first to determine an initial effective solution z, for each solution z define a neighborhood $Y(z)$, determine

several candidate solutions from the neighborhood of the current solution, and select the best candidate solution from them. Selecting the best candidate solution is a search process. To avoid the search process being limited to cycles, TS avoids the local optimization of the search algorithm by constructing a tabu table and defining stop rules. Tabu list before saving n the second taboo length avoids returning to the original solution, thus improving the search ability of the solution space; Stop rule defines that when the optimal solution cannot be improved within several iterations, the algorithm stops. In addition, neighborhood, tabu list, tabu length, amnesty rule and initial solution in the tabu search algorithm will directly affect the search optimization results.

Feature selection based on tabu search is an optimization problem constrained by the objective function, and the appropriate objective function improves the quality of search and optimal feature selection. A good feature solution should guarantee as much classification information as possible on the minimum number of features. In information theory, the greater the information gains of an attribute, the greater the amount of information it contains [26]. Based on the information gained, the classification information of feature vectors can be effectively evaluated. Therefore, this paper selects information gain as the objective function and defines the objective function as shown in Formula (5):

$$G_T = R\sum_{i=1}^{m} C(i) \times \frac{\sum_{j=1}^{n} G(A_j)}{n} \quad (5)$$

In Formula (5), $C(i)$ represents sample i whether it is correctly classified, m is the number of samples; $G(A_j)$ is information gain of features i. Ensure that the maximum classification information is guaranteed with a small number of features through Formula (5), and select divided by n that can ensure faster tabu search speed and avoid overfitting.

The selection of the initial solution in tabu search greatly impacts the effect of tabu search. In the calculation process of other optimal feature selection algorithms, due to the large feature dimension of actual network traffic, it will affect the efficiency of the tabu search algorithm, and feature redundancy will also affect the selection of the optimal feature set. Therefore, the initial solution of tabu search has an important impact on search efficiency and quality.

Generally, the larger the feature, the higher the accuracy of the analysis is. However, in practice, too large a feature space will cause two problems: (1) The huge feature space not only needs higher storage space but also increases the measurement time, which is difficult to apply to real-time traffic analysis; (2) In some applications, such as anomaly detection, service classification, etc., the characterization of different network services requires different feature attribute vectors. If all features are used to represent different service flows, not only the learning effect is reduced, but also the learning time is increased. So, feature selection is to mine the best feature set to describe network traffic; best and tabu search provides a near-optimal solution.

3) Feature selection design based on PCA-TS algorithm: The statistical characteristics of network traffic refer to the characteristics of extracting ports and protocols from the attributes of packets or flows. Such as message length, arrival interval, number of messages, flow duration, number of messages in the flow, etc. Feature vectors represent these statistical characteristics. Such as a network flow F, the characteristic description based on the flow can be expressed as $F = \{y_1, y_2, \ldots, y_n\}$, where y_i represents the value of the feature. The feature set of a flow may contain as many as hundreds of features. Finding a small number of optimal feature subsets to describe the flow is important to improve learning efficiency.

Therefore, this paper makes full use of the feature that PCA can perform fast and effective feature reduction on high-dimensional data, and improves the efficiency of solving the optimal solution of the tabu search method by eliminating feature redundancy and reducing the dimension space. To this end, this paper selects network traffic characteristics by combining PCA and TS algorithms. The flow chart of the PCA-TS feature selection method is shown in Fig. 4.

In Fig. 4, the specific implementation steps of the PCA-TS algorithm are as follows:

1) The tabu table is empty, and initialization parameters are set: tabu length $L_J = 13$, maximum iterations $D_{max} = 600$, maximum improvement times $\bar{D}_{max} = 100$.

2) Use PCA to reduce the original network traffic characteristics and obtain the reduced feature collection $G'_T = \{T_1, T_2, \ldots, T_p\}$, p is the number of feature sets after reduction.

3) To feature set G'_T perform binary coding to obtain the initial solution R_{init}.

4) Set termination conditions, when getting $\bar{D}_{max}$, the search stops; When the best solution cannot be improved by passing R_{init}, stop searching.

5) Judge whether the termination conditions are met. If the termination conditions are met, end the operation and output the optimal flow feature subset. Otherwise, go to the next step.

6) Initial solution R_{init} brings into the neighborhood structure to calculate the neighborhood solution, and the best candidate solution is selected through the objective function.

7) Judge whether the candidate solution meets the amnesty rule. If yes, update the optimal solution in the tabu list and go to step (4), otherwise go to the next step.

8) Calculate the tabu attribute of the candidate solution, select the initial value of the optimal replacement tabu table for non-tabu objects, and go to step (4).

9) End, output the optimal flow characteristic subset G'_R.

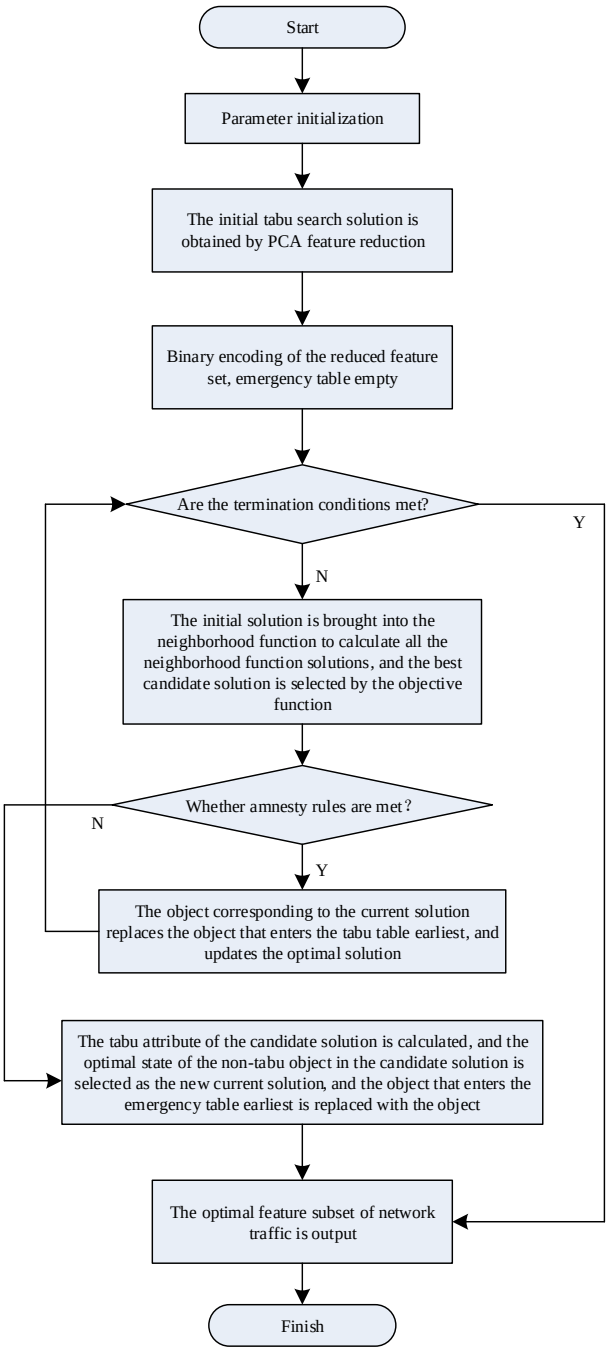


Fig. 4. Process of feature selection method based on PCA-TS.

In Fig. 4, the specific implementation steps of the PCA-TS algorithm are as follows:

1) The tabu table is empty, and initialization parameters are set: tabu length $L_J = 13$, maximum iterations $D_{max} = 600$, maximum improvement times $\bar{D}_{max} = 100$.

2) Use PCA to reduce the original network traffic characteristics and obtain the reduced feature collection $G'_T = \{T_1, T_2, \ldots, T_p\}$, p is the number of feature sets after reduction.

3) To feature set G'_T perform binary coding to obtain the initial solution R_{init}.

4) Set termination conditions, when getting $\bar{D}_{max}$, the search stops; When the best solution cannot be improved by passing R_{init}, stop searching.

5) Judge whether the termination conditions are met. If the termination conditions are met, end the operation and output the optimal flow feature subset. Otherwise, go to the next step.

6) Initial solution R_{init} brings into the neighborhood structure to calculate the neighborhood solution, and the best candidate solution is selected through the objective function.

7) Judge whether the candidate solution meets the amnesty rule. If yes, update the optimal solution in the tabu list and go to step (4), otherwise go to the next step.

8) Calculate the tabu attribute of the candidate solution, select the initial value of the optimal replacement tabu table for non-tabu objects, and go to step (4).

9) End, output the optimal flow characteristic subset G'_R.

E. Abnormal Flow Detection

After selecting network traffic characteristics, you can use the selected traffic characteristics to detect abnormal traffic through the network security architecture abnormal traffic detection module. Improve the detection efficiency of abnormal flow and improve the detection accuracy. This paper uses the SVM algorithm to detect abnormal traffic, assuming there is k type of samples, and then it is necessary to construct k two class classifiers. Each classifier is used to separate one class from the rest. During training, please take one of them as positive, and the rest $k-1$ class is negative. When judging, the sequence of the tested samples passes through k, the total of two class classifiers k output values is $f_i(x) = sgn(g_i(x)), i = 1,2,\ldots,k$. If the decision result contains only one+1, the corresponding classifier's sample class to be detected is the positive class. Suppose there is more than one+1 in the decision result, that is, classification overlap. In that case, it is also necessary to compare the decision function value of the classifier whose output is+1, and the positive class of the classifier with the largest value represents the class of the sample to be detected. If the judgment result is -1, the sample is considered to be indivisible. Therefore, this paper proposes an improved SVM multi-classification algorithm. The idea of class distance in clustering analysis is used as the basis for sorting the second-class classifiers in the detection model.

About k class flow characteristic samples, calculate the center distance from each class to other classes, and then calculate the average distance from each class to other classes. The class with the largest average distance is the class with the most obvious specificity, and such class is preferred as the positive class of the second-class classifier ranking first. The relevant definition of distance is:

Definition 1: Center distance. The center distance of the flow characteristic samples of class i and j is defined as the Euclidean distance in the space of the spherical center that can contain all class i traffic characteristic samples to the spherical center that can contain all samples of class j traffic characteristics, recorded as d_{ij}.

Definition 2: Average distance. The mean distance between the class $i$traffic characteristics and the remaining categories is defined as the mean of the center distance from the class i traffic characteristics to the other samples of the traffic characteristics, recorded as γ_i, and meet:

$$\gamma_i = \left(\frac{G'_R}{k-1}\right)\sum_{j=1}^{k} d_{ij}(i \neq j).$$

The specific implementation steps are as follows:

1) Calculate the center distance of $d_{ij}(i, j = 1,2,\ldots,k, i \neq j)$between various flow characteristic samples and other flow characteristic samples according to definition 1.

2) Calculate the average distance of $\gamma_i(i = 1,2,\ldots,k)$ between various flow characteristic samples and other flow characteristic samples according to definition 2.

3) Compare the size of step $2\gamma_i$, and then follow the categories that numbered in descending orderγ_i.

4) Construct one by one according to the sample number sequence obtained in step (3) the k two class classifiers. The sample with the first number is the positive class of the first second-class classifier, the sample with the second number is the positive class of the second-class classifier, and so on.

During abnormal flow detection, let the sample to be tested pass through each two-class classifier in turn. If the decision result of the sample to be tested in a classifier is+1, then it is determined that the sample is the positive class of the corresponding classifier, and the detection of this sample is terminated. If the output result of samples passing through all the second-class classifiers in turn is - 1, it is determined as unknown flow, added to the set to be verified, and waiting for re training. The sample decision process is shown in Fig. 5.

The distance first SVM multi-classification algorithm can improve the classifier's detection accuracy. Therefore, although k class two classifiers were constructed when building model sets. However, in the improved SVM multi-classification algorithm, when a classifier is judged as+1, the judgment is terminated to shorten the sample detection time.

At this time, according to the specific process of abnormal network traffic detection method, abnormal traffic detection can be realized by the following methods:

1) NetFlow technology captures network traffic packets through flow.

2) The captured data packets are used in the KNN algorithm to filter traffic and eliminate duplicate traffic packets.

3) After data filtering, the traffic packets are input into the PCA-TS algorithm for data dimensionality reduction and feature selection.

4) After obtaining the traffic characteristics, the structure contains N sample a set of traffic, including normal traffic and $N-1$ abnormal flow with obvious difference in the distribution characteristics of three kinds of flow.

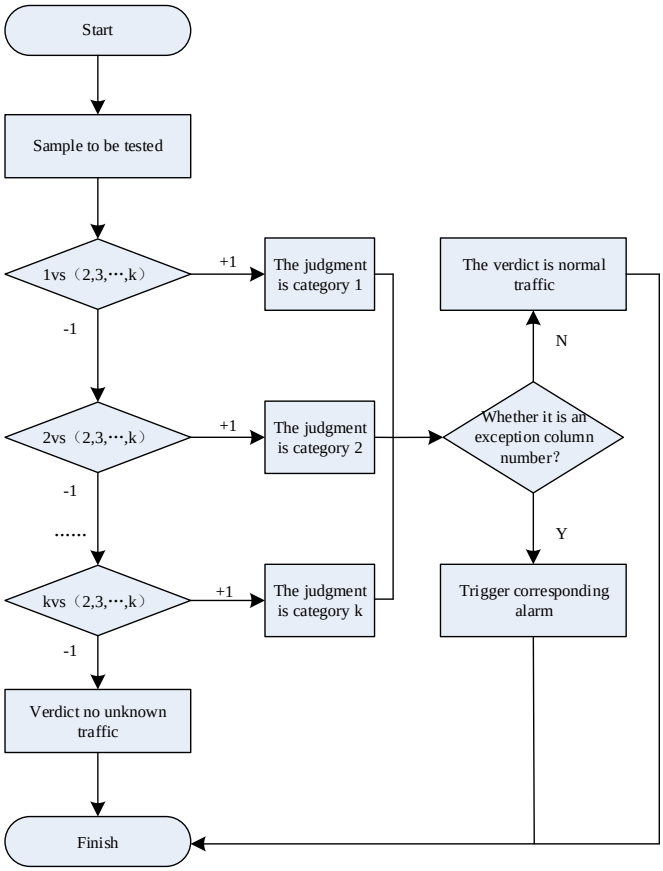


Fig. 5. Sample decision process.

5) The improved SVM multi - classification method is used to input the characteristic samples of the traffic to be measured into the SVM multi-class classifier. If the second-class classifier can recognize the characteristic samples of the traffic to be measured, it is determined that the traffic of the corresponding category is detected. If the detected flow is normal, continue; if it is abnormal, send an alarm. Repeat step (5). If the flow characteristic sample to be detected is determined to be an unknown flow, perform step (6).

6) Add unknown traffic to the collection to be verified. If the traffic in the set to be verified can be clustered and is significantly different from the normal traffic, it can be considered that a new anomaly has occurred.

7) Add new exceptions to the training samples in step (1) for re training to obtain a new model set, and repeat step (5) to achieve network security detection.

III. Experimental Analyses

To verify the effectiveness of the network security detection method in this paper, this paper constructs a simulation experiment through the NS2 simulation platform and shows the network topology in Fig. 6.

In Fig. 6, R1, R2 and R3 are routers, of which R2 is the "key router". The link between R2 and R3 is the bottleneck link, with a bandwidth of 10Mbps and a delay of 30ms. All other links have a bandwidth of 100Mbps and a delay of 15ms. The network contains 25 legitimate TCP connections, 10 of which are background traffic.

Meanwhile, attack parameters are designed: attack cycle is 1s, attack pulse duration is 150ms or 200ms or 250ms, and attack pulse intensity is 30Mbps or 40Mbps. The observation time window WS duration is 90s, and the attack packet types are UDP, ICMP, and invalid TCP.

The experimental data set uses the Honeynet Challenges data set provided by the HoneyPot Project. HoneyPot Project is a non-profit network security research institution committed to studying the latest network attacks and developing open-source security tools to improve the network environment. The organization has volunteers from all over the world. On its official website, many open-source security tools are developed to improve the network security environment. Table I shows some network traffic attributes of the experiment.

Analyze this method's abnormal feature selection ability before abnormal flow detection, and analyze the time required for different abnormal flow feature selections. The analysis results are shown in Table II.

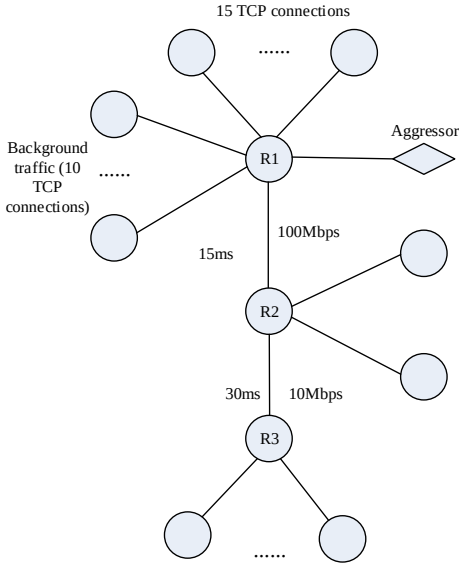


Fig. 6. Topology structure of the experimental network.

TABLE I. Attribute List of Network Traffic Status

Serial number	Symbolic representation	Feature description
1	outPckets	Total number of egress packets
2	outBytes	Total size of the egress packet
3	inPckets	Total number of incoming packets
4	inBytes	Indicates the total packet size of the entry
5	minOutpktLen	Indicates the minimum Byte of the egress packet
6	maxOutpktLen	Indicates the maximum Byte of the egress packet
7	meanLenOutsm	Average Byte of the egress packet
8	minInpktLen	Indicates the maximum Byte of the incoming packet
9	maxInpktLen	Indicates the maximum Byte of the incoming packet
10	meanLenInsm	Average Byte of an incoming packet
11	Outtotalduration	The total duration of the exit flow
12	Outavgduration	The average duration of the exit flow
13	Intotalduration	The total duration of the inlet stream
14	Inavgduration	The average duration of the inlet stream
15	OutavePktPerSecond	Average outbound packet size per second
16	InavelPktPerSecond	The average size of an entry packet in seconds
17	InnerOneToMulty	Indicates the number of one-to-many internal IP addresses
18	OutOneToMulty	Indicates the number of one-to-many external IP addresses
19	stdLenOutqsm	Mean square error of the egress message
20	stdLenInqsm	Mean square error of the incoming message

TABLE II. TIME REQUIRED FORABNORMAL FEATURE SELECTION

Feature sequence number	Feature type	Specific description	Time required for feature selection /s
1	duration	Connection duration	2.4
2	service	The network service type of the target host	2.6
3	src_bytes	Number of bytes of data from the source host to the destination host	2.1
4	land	Determine whether the connection is coming from the same host or port	1.5
5	urgent	Number of urgent packets	1.7
6	num_failed_logins	The number of failed login attempts	1.5
7	num_compromised	Compromised frequency	1.7
8	num_access_files	The number of times the control file was accessed	1.5
9	is_hot_login	Whether the login belongs to the hot list	1.6
10	protocol_tyoe	Protocol type	2.6
11	flag	The connection status is normal or incorrect	1.4
12	dst_bytes	The number of bytes of data from the destination host to the source host	2.1
13	hot	The number of times to access system-sensitive files and directories	2.5
14	wrong_fragment	The number of incorrect segments	1.9
15	dst_host_diff_srv_rate	Among the top 100 connections, the proportion of connections that have different services to the same destination host as the current connection	1.5
16	dst_host_srv_diff_host_rate	Among the top 100 connections, the number of connections that have the same destination host as the current connection and the number of connections that have a different source host from the current connection	1.5
17	srv_count	Number of connections that have the same service as the current connection in the last two seconds	1.6

According to Table II, this method can effectively select multiple features during feature selection to provide reliable data for subsequent network security detection. At the same time, in the feature selection process, the time for this method to realize feature selection does not exceed 3.0s. Therefore, this method has a strong feature selection ability, which provides a reliable guarantee for subsequent abnormal traffic detection.

This paper uses the G score to evaluate this method's abnormal traffic detection effect. G score is defined as:

$$G = \sqrt{precision \times recall} \quad (6)$$

In Formula (6), $precision$ and $recall$ indicate the accuracy rate and recall rate in turn. The higher the G score, the stronger the method's ability to detect abnormal traffic is. Test the G scores of different numbers of packets when they are attacked by different traffic and show the experimental data results through Fig. 7.

As shown in Fig. 7, with the increase in the number of test packets, the G scores obtained by this method begin to decline under attacks of different attack data types. Among them, when being attacked by UDP, the G score obtained by the test is the lowest among the three attack types, but not less than 0.70. It shows that this method can maintain high performance in the detection process. This paper can get a higher G score for detecting ICMP and invalid TCP attacks. It can be seen that this detection method can effectively detect multiple types of attacks.

This study selected three different scenarios for verifying abnormal traffic detection. In scenario B1, there are no attacks in the network, or the attacks present in the network have no direct impact on TCP data traffic. In the B2 scenario, the attacks present in the network have a direct impact on TCP data traffic, but there are no serious attacks. In the C3 scenario, there are serious attacks in the network. Verify the detection effectiveness of our method by analyzing the degree of traffic fluctuations in different scenarios. The analysis results are shown in Fig. 8.

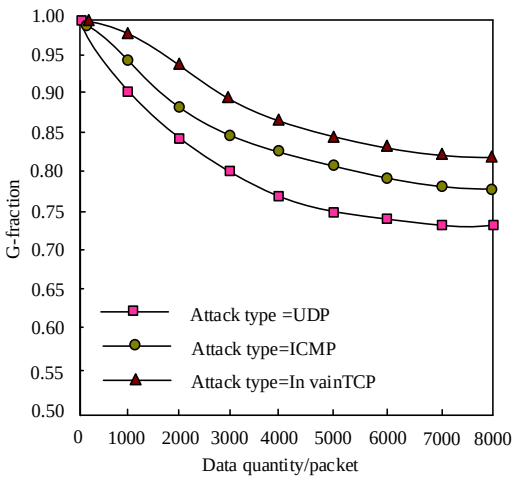


Fig. 7. Analysis of G score test results.

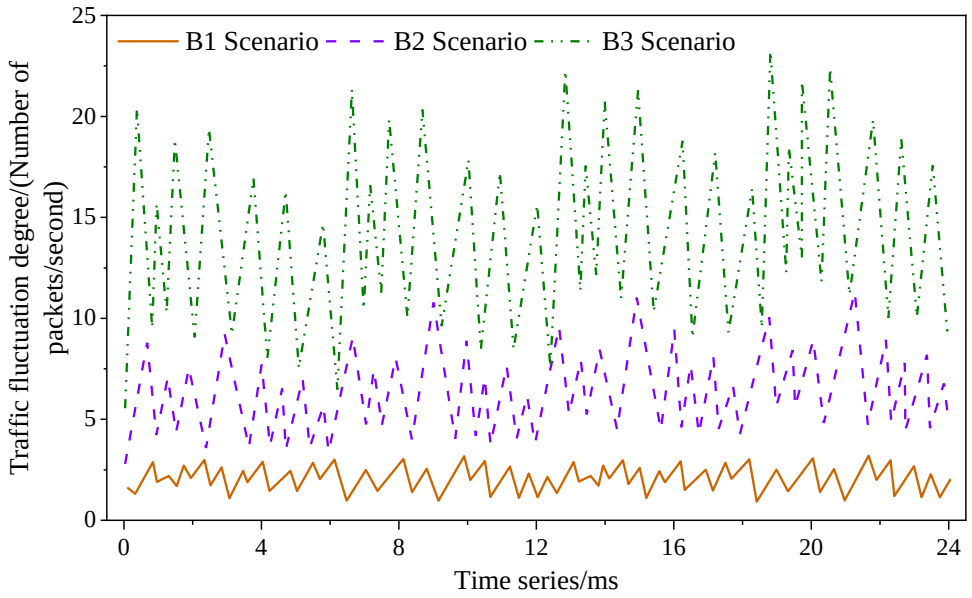
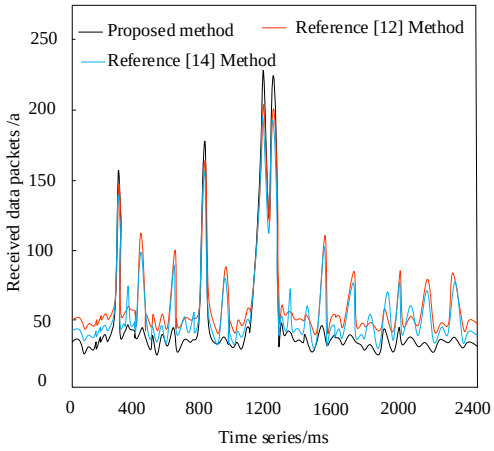


Fig. 8. Analysis of traffic fluctuation degree.

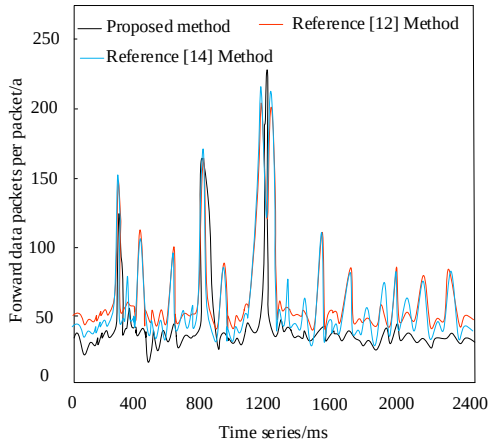
According to Fig. 8, when there is no attack in the network or the attack has no direct impact on TCP data traffic, the traffic always fluctuates below five packets/second, with a small fluctuation degree. When attacks in the network directly impact TCP data traffic, the fluctuation of network traffic increases. When serious attacks exist in the network, the traffic amplitude exceeds 20 packets/second fluctuations. From the above analysis, it can be seen that this method can effectively detect flow fluctuation.

Injecting attack traffic at different times, namely 400s, 800s, and 1200s, analyze the changes in the number of received and forwarded packets detected by the method proposed in this paper, the method proposed in reference [12], and the method proposed in reference [14]. The analysis results are shown in Fig. 9.

As shown in Fig. 9, after applying the method proposed in this article, under normal circumstances, the number of received packets is approximately equal to the number of forwarded packets; After injecting an attack, it will restrict the forwarding of abnormal packets, resulting in a lower number of forwarded packets than received packets. At this point, the number of forwarded packets shows a significant decrease. In response to persistent attacks, the number of packets forwarded by the port gradually decreases until the normal number of forwards is restored. From this, it can be seen that the method proposed in this article has strong ability to detect abnormal traffic and can achieve network security detection. By comparing the methods in reference [12] and reference [14], it can be seen that although the overall trend of the two methods is similar to that of the method in this paper, both methods show abnormal increase or decrease, indicating that the two comparison methods are affected by attacks and have misidentification phenomena..



(a) Analysis of changes in the number of received data packets.



(b) Analysis of changes in the number of forwarded packets.

Fig. 9. Analysis of changes in the number of packets during injection attacks.

IV. CONCLUSION

This paper studies the network security detection method based on abnormal traffic detection, uses this method, and applies this method to the experimental detection process. Experiments show that this method has a good detection effect on the common abnormal traffic and attacks in the network. Given the shortcomings of the current research, the following aspects can be improved in the future research work:

1) Find or build appropriate data sets. The existing real data sets have some shortcomings, lacking real attack data. Most researchers use traditional network data sets for experiments. However, due to the network environment's limitations, the simulation data cannot fully reflect the real network conditions.

2) Accurately identify the types of network attacks and make reasonable solutions. At present, this method can only achieve the detection and early warning of abnormal traffic and cannot achieve the processing of abnormal traffic. In the future, some effective abnormal traffic processing methods can be designed to improve network security.

AVAILABILITY OF DATA AND MATERIALS

The datasets used in this paper are available from the corresponding author upon request.

CONFLICTS OF INTEREST

The authors declared that they have no conflicts of interest regarding this work.

AUTHORSHIP CONTRIBUTION STATEMENT

Tao Xiao: Writing-Original draft preparation

Conceptualization, Supervision, Project administration.

Yang Ke: Language review

Hu YiWen: Methodology

Wang HongYa: Software

REFERENCES

[1] E. Balamurugan, A. Mehbodniya, E. Kariri, K. Yadav, A. Kumar, and M. A. Haq, "Network optimization using defender system in cloud computing security-based intrusion detection system withgame theory deep neural network (IDSGT-DNN)," Pattern Recognit Lett, vol. 156, pp. 142–151, 2022.

[2] C. Ding, Y. Chen, Z. Liu, A. M. Alshehri, and T. Liu, "Fractal characteristics of network traffic and its correlation with network security," Fractals, vol. 30, no. 02, p. 2240067, 2022.

[3] M. A. Ferrag, L. Shu, O. Friha, and X. Yang, "Cyber security intrusion detection for agriculture 4.0: Machine learning-based solutions, datasets, and future directions," IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 3, pp. 407–436, 2021.

[4] S. Pande, A. Khamparia, and D. Gupta, "An intrusion detection system for health-care system using machine and deep learning," World Journal of Engineering, vol. 19, no. 2, pp. 166–174, 2022.

[5] A. K. Bediya and R. Kumar, "A novel intrusion detection system for internet of things network security," in Research Anthology on Convergence of Blockchain, Internet of Things, and Security, IGI Global, 2023, pp. 330–348.

[6] J. Chen and Y. Miao, "Study on network security intrusion target detection method in big data environment," International journal of internet protocol technology, vol. 14, no. 4, pp. 240–247, 2021.

[7] Y. Wang et al., "Exhaustive research on the application of intrusion detection technology in computer network security in sensor networks," J Sens, vol. 2021, pp. 1–11, 2021.

[8] Q. Ding and J. Li, "AnoGLA: An efficient scheme to improve network anomaly detection," Journal of Information Security and Applications, vol. 66, p. 103149, 2022.

[9] C. H. Nwokoye, I. I. Umeh, N. N. Mbeledogu, and V. O. S. Okeke, "Scan-Based Worms: The Impact of IPV4 Address Space on Epidemic Computer Network Models.," Engineering Letters, vol. 29, no. 2, 2021.

[10] C. Do Xuan, "Detecting APT attacks based on network traffic using machine learning," Journal of Web Engineering, pp. 171–190, 2021.

[11] L. Duan, J. Zhou, Y. Wu, and W. Xu, "A novel and highly efficient botnet detection algorithm based on network traffic analysis of smart systems," Int J Distrib Sens Netw, vol. 18, no. 3, p. 15501477211049910, 2022.

[12] M. Woźniak, J. Siłka, M. Wieczorek, and M. Alrashoud, "Recurrent neural network model for IoT and networking malware threat detection," IEEE Trans Industr Inform, vol. 17, no. 8, pp. 5583–5594, 2020.

[13] P. Steno, A. Alsadoon, P. W. C. Prasad, T. Al-Dala'in, and O. H. Alsadoon, "A novel enhanced region proposal network and modified loss function: threat object detection in secure screening using deep learning," J Supercomput, vol. 77, pp. 3840–3869, 2021.

[14] T. Gaber, A. El-Ghamry, and A. E. Hassanien, "Injection attack detection using machine learning for smart IoT applications," Physical Communication, vol. 52, p. 101685, 2022.

[15] K. Lin, X. Xu, and F. Xiao, "MFFusion: A multi-level features fusion model for malicious traffic detection based on deep learning," Computer Networks, vol. 202, p. 108658, 2022.

[16] G. Xie, Q. Li, and Y. Jiang, "Self-attentive deep learning method for online traffic classification and its interpretability," Computer Networks, vol. 196, p. 108267, 2021.

[17] A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé, "Distributed detection/localization of change-points in high-dimensional network traffic data," Stat Comput, vol. 22, no. 2, pp. 485–496, 2012.

[18] V. Mic and P. Zezula, "Data-dependent metric filtering," Inf Syst, vol. 108, p. 101980, 2022.

[19] B. H. Meyer, A. T. R. Pozo, and W. M. N. Zola, "Improving barnes-hut t-sne algorithm in modern gpu architectures with random forest knn and simulated wide-warp," ACM Journal on Emerging Technologies in Computing Systems (JETC), vol. 17, no. 4, pp. 1–26, 2021.

[20] L. Guo, S. Wang, J. Yin, Y. Wang, J. Yang, and G. Gui, "Federated user activity analysis via network traffic and deep neural network in mobile wireless networks," Physical Communication, vol. 48, p. 101438, 2021.

[21] S. Yang, L. Ning, X. Cai, and M. Liu, "Dynamic spatiotemporal causality analysis for network traffic flow based on transfer entropy and sliding window approach," J Adv Transp, vol. 2021, pp. 1–17, 2021.

[22] M. Li, Y. Liu, Q. Zheng, W. Qin, and X. Ren, "Stable feature selection based on brain storm optimisation for high-dimensional data," Electron Lett, vol. 58, no. 1, pp. 10–12, 2022.

[23] L. Luo et al., "Adaptive data dimensionality reduction for chemical process modeling based on the information criterion related to data association and redundancy," Ind Eng Chem Res, vol. 61, no. 2, pp. 1148–1166, 2022.

[24] J. Ren, H. Wang, K. Luo, and J. Fan, "A Priori Modeling of NO Formation with Principal Component Analysis and the Convolutional Neural Network in the Context of Large Eddy Simulation," Energy & Fuels, vol. 35, no. 24, pp. 20272–20283, 2021.

[25] W. Chang et al., "Prediction of hypertension outcomes based on gain sequence forward tabu search feature selection and xgboost," Diagnostics, vol. 11, no. 5, p. 792, 2021.

[26] Z. Zhang, H. Tang, and Z. Xu, "Fatigue database of complex metallic alloys," Sci Data, vol. 10, no. 1, p. 447, 2023.

ODFM: Abnormal Traffic Detection Based on Optimization of Data Feature and Mining

Xianzong Wu
College of Intelligence and Computing, Tianjin University, Tianjin, China

***Abstract*—The booming of computer networks and software applications has led to an explosive growth in the potential damage caused by network attacks. Efficient detection of abnormal traffic in networks is appealing for facilely mastering the traffic tracking and locating for network usage at low resource cost. High quality abnormal traffic detection of Internet becomes particularly relevant during the automated services of multiple application situations. This paper proposes a novel abnormal traffic detection algorithm called ODFM based on the optimization of data feature and mining. Specially, we develop a feature selection strategy to reduce the feature analysis dimension, and set a peer-to-peer (P2P) traffic identification module to filter and mine the related service traffic to reduce the amount of data detection and facilitate the abnormal traffic detection. Experimental results demonstrate that the proposed algorithm greatly improves the detection accuracy, which verifies its effectiveness and competitiveness in the general tasks of abnormal network traffic detection.**

***Keywords*—Abnormal traffic; detection; data mining; feature dimension optimization; network security**

I. Introduction

With the increasing complexity and volume of network traffic, the need for effective anomaly traffic detection algorithms is essential to ensure the security and reliability of networks. Anomaly traffic, which deviates significantly from normal patterns, can indicate potential cyber threats, network performance issues, or abnormal user behavior [1]. Anomaly traffic detection involves the identification of abnormal patterns or behaviors within network traffic that deviate significantly from the expected norm. The study has drawn great attention in recent decades due to the strong security demands on the network communications. However, traffic detection is deemed to be a developing and challenging issue since it needs to deal with various difficulties coming from imbalanced data, increasing network traffic volume, evolving and sophisticated attacks, as well as dynamic and variable network environments [2-5]. Various techniques have been developed to detect and analyze such anomalies, ranging from statistical-based methods [6-8], time series analysis [9-11], machine-learning based methods [12-14], deep-learning based methods [15-17], ensemble methods [18-20], flow-based analysis [21-23], hybrid methods [24, 25], to unsupervised clustering methods [26, 27].

Statistical based methods analyze network traffic data statistically to model normal behavior and detect anomalies [8]. These methods include mean-based models, standard deviation-based models, moving average models, and Gaussian distribution modeling. Time series analysis techniques model network traffic as a time-ordered series of data points, capturing temporal dependencies and patterns. Techniques such as Autoregressive Integrated Moving Average (ARIMA) [28], Hidden Markov Models (HMM) [29], and Wavelet Transform [30] is commonly used for time series analysis in anomaly detection. On the other hand, machine learning algorithms leverage historical network traffic data to distinguish between normal and anomalous patterns. Supervised learning algorithms like Support Vector Machines (SVM) [31], Random Forests [32], and Neural Networks classify traffic based on labeled datasets. Unsupervised learning algorithms, such as clustering and density estimation, detect anomalies without prior labeling. Deep learning algorithms, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), automatically learn hierarchical and temporal patterns from raw network traffic data, excelling in detecting complex and evolving anomalies. However, they require significant computational resources and large labeled datasets. Ensemble methods combine multiple anomaly detection models to improve overall detection performance. Techniques like bagging, boosting, and stacking reduce false positives and enhance accuracy. Flow-based analysis aggregates network traffic into flows, representing communication between specific source and destination IP addresses. Flow-based anomaly detection uses flow features such as traffic volume, duration, or byte count to identify deviations from normal flow behavior. Besides, hybrid methods integrate multiple detection techniques, such as statistical, machine learning, and rule-based methods, to enhance accuracy. These approaches leverage the strengths of each technique to effectively identify a wide range of anomalies. Unsupervised clustering methods group network traffic instances into clusters, identifying anomalies as instances that deviate or form separate clusters. Clustering algorithms like k-means [33], DBSCAN [34], and self-organizing maps (SOM) [35] are commonly employed for anomaly traffic detection.

These mentioned methods have been faced with various difficulties and challenges. For instance, many continually develop sophisticated techniques to evade detection, employing tactics like traffic encryption, obfuscation, and mimicry. This poses a challenge for anomaly detection systems to identify and classify these evolving attacks accurately. Network traffic

often exhibits an imbalanced distribution, where normal traffic significantly outweighs anomalous traffic. Imbalanced data can lead to biased models that exhibit a higher false positive rate or overlook certain types of anomalies. Handling imbalanced data and addressing the bias towards the majority class is critical for effective anomaly detection. Network environments are dynamic, constantly evolving with changing user behavior, new applications, and network configurations. Anomaly detection algorithms need to adapt and remain capable of detecting novel or emerging anomalies. This requires continuous monitoring, model updates, and the ability to adapt to evolving network dynamics. Besides, real-time anomaly detection places stringent requirements on processing time and scalability. Efficient and scalable algorithms are necessary to handle large-scale network traffic, detect anomalies within strict time constraints, and ensure real-time analysis. Moreover, deep learning and complex machine learning models, while powerful in detecting anomalies, often lack interpretability and explain ability. Understanding the rationale behind anomaly detection outcomes is essential for effective response and decision-making by analysts. Ensuring transparent and interpretable anomaly detection models is a challenging task [36]. These challenges are actively being explored and addressed through ongoing research and development efforts. Advancing anomaly traffic detection techniques requires focusing on handling evolving attacks, addressing imbalanced and limited labeled data, enhancing adaptability to dynamic environments, reducing false positives, and improving interpretability. By addressing these challenges, the future advancement of anomaly detection systems in network traffic analysis can be achieved.

Developing appropriate anomaly detection method need to consider several factors, including the specific network environment, the characteristics of the anomalies, computational resources, and available data. Ongoing research and studies are actively exploring novel algorithms, advancements in machine learning techniques, and the integration of domain knowledge to enhance the accuracy and efficiency of anomaly detection in network traffic. These efforts aim to improve the detection capabilities in diverse network environments and adapt to evolving threat landscapes. In this paper, we propose a new approach called ODFM to anomaly traffic detection that jointly optimizes the data volume and feature dimension to achieve high-accuracy detection results. The main contributions of the paper include:

1) The paper propose a novel method of combining feature dimension reduction preprocessing with data mining optimization to general tasks of anomaly traffic detection, which takes into account the P2P traffic identification and the related-service traffic filter.

2) We introduce a feature selection strategy to achieve the feature dimension reduction, which can automatically locate key traffic information, speed up the anomaly detection and then propel the data optimization and engine extraction. It effectively addresses the issue of feature vector construction of data mining model.

3) An elaborate system design is conducted to verify the effectiveness of the proposed ODFM algorithm. Experimental evaluations have demonstrated its efficacy and great potential under various traffic anomaly detection conditions.

II. The Proposed ODFM Method

A. Foundation Statement

The proposed ODFM method is designed to address the issue of anomaly traffic detection. To achieve the goal of ODFM, we develop an ODFM based anomaly network traffic detection system that can effectively identify and classify abnormal patterns within network traffic data. The system aims to enhance network security, mitigate potential threats, and optimize network performance. The system mainly involves data mining mechanism to complete the network anomaly traffic detection.

Data mining mechanism aims to facilitate advanced anomaly detection and prevention in network traffic through the application of association mechanisms. When employing association mechanisms, it becomes feasible to effectively detect patterns and hidden knowledge between diverse and interconnected data items. By combining various data attribute values, these mechanisms can predict attribute values for a certain class of data, which provides significant advantages in acquiring and utilizing patterns from massive datasets. Within computer information systems, the utilization of association mechanisms enables precise analysis of network anomalies, fault information, and user network data. The formation of fault factor sets and the integration and analysis of related information categories enhance the scrutiny of traffic data and management processes. Moreover, potential rules between different network information and data are derived, thereby lessening the likelihood of network risks and achieving efficient early warning and handling of abnormal network traffic.

The primary objective of this anomaly network traffic detection system is to reduce the data volume and feature dimensions processed by the anomaly detection module, thereby enhancing the detection accuracy. This system focuses specifically on detecting malicious scanning and denial of service (DoS)/ distributed denial of service (DDoS) anomalies under the transmission control protocol, making it particularly applicable in transmission control protocol environments with a certain scale of hosts. Traditional anomaly traffic detection systems typically involve dealing with massive input data and implementing high-dimensional feature models for training. In contrast, the design of this system develops a feature selection strategy to reduce the dimensionality of feature analysis, and includes a P2P traffic identification module to filter relevant business traffic, thus reducing the data processing load and improving the accuracy of the detection process.

B. The Workflow of ODFM System

Fig. 1 depicts the detailed workflow of the proposed ODFM system architecture, which is designed with two primary phases, off-line training and online classification. In the off-line training phase, real-time network traffic and the corresponding known traffic labels are processed to create labeled training data. Through the application of hybrid feature extraction techniques in data preprocessing, an offline traffic

classification model is developed. During the online classification stage, the system performs real-time network traffic collection and an off-line data file analysis within the traffic collection and analysis module. All essential field information related to the traffic is stored in the database for further analysis. In the data preprocessing module, the feature selection algorithm and feature extraction engine are utilized. The former is employed in an offline state for comprehensive analysis, while the latter conducts secondary statistical analysis, reorganization, and calculation of the original traffic data from the database. This facilitates the construction of feature vectors for subsequent data mining models.

The data preprocessing module encompasses some vital components such as data storage and access, key table structural design, storage design, and trigger design. These elements ensure an efficient handling of the data during processing. Within the abnormal traffic classification detection module, the primary objective is to identify and filter P2P traffic, construct a weak classifier, and enhance its performance. By following this system design, it is possible to improve the accuracy and effectiveness of online traffic classification, enabling efficient detection and analysis of abnormal network traffic.

C. The Architecture Design

To achieve comprehensive maturity and facilitate modular construction in developing the proposed ODFM system for general anomaly network traffic detection, it is imperative to design and establish a corresponding software architectural framework. Maturity, in this context, refers to universally recognized technology that guarantees superior system performance, ultimately enhancing the system's long-term viability. Modular construction involves methodically dividing software functionality into autonomous modules, thereby streamlining software maintenance and facilitating seamless upgrades.

The ODFM system is meticulously constructed upon the robust Model-View-Controller (MVC) design pattern, adopting a multi-layered approach. The design ensures an efficient segregation of crucial elements, encompassing business logic, interface display, and data models. As a result, the ODFM system achieves parallel operations, which can bolster its computational capabilities, and thereby optimizing the overall performance of network anomaly traffic detection.

The software technology implemented in the proposed system predominantly encompasses three distinctive layers. Fig. 2 illustrates the three-layer technical architecture. Firstly, Presentation Layer unveils an intuitive and user-friendly interface for seamless system operation and discernible result display. It adeptly handles the input data, promptly forwarding it to the Business Processing Layer or, in turn, receiving pertinent data communicated from the Business Parallel Processing Layer. Secondly, Business Processing Layer encompasses specific business logic, which is further partitioned into three sub-layers, including data acquisition, business applications and external interfaces. Notably, the external interfaces facilitate seamless data exchange and robust sharing capabilities between the ODFM system and other compatible systems, fostering optimal data application. Web Services has been judiciously chosen as the technology of choice for shaping these external interfaces. Lastly, Data Layer assumes the responsibility for efficient storage and meticulous management of indispensable network traffic information. It reliably provides indispensable data that fuels the construction of intricately designed classification models within the system. By harnessing the power of this meticulously crafted software technology architecture, the ODFM system confidently attains exemplary levels of maturity, modularity, and operational efficiency. With this robust framework in place, the system flawlessly accomplishes reliable network traffic analysis and impeccable management.

D. Implementation of Modules in ODFM

The ODFM design implementation adopts the Java programming language. For the proper functioning of the windows packet capture (Winpcap) in Java, it relies on utilizing Winpcap at the data link layer control level. Winpcap provides the required network underlying resource access interface for Java, offering extensive capabilities to interact with the network resources. This implementation ensures the advantage of maintaining system independence during the operational process.

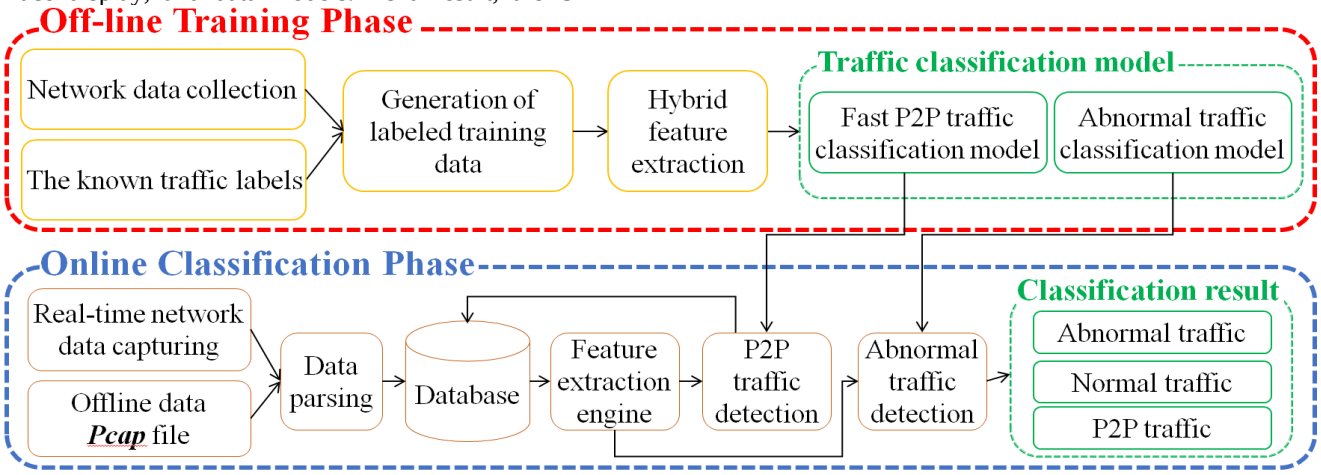


Fig. 1. The detailed flow of the proposed ODFM, which contains an off-line training phase and an online classification phase.

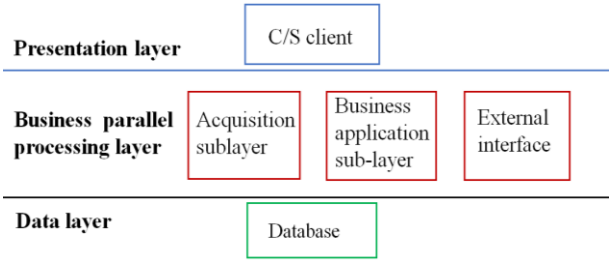


Fig. 2. The three-layer technical architecture of ODFM system.

In the module of traffic collection and parsing, Winpcap serves as a network data capturing framework, which offers excellent application performance in practical scenarios. It consists of filters, wpcap.dll, and packet.dll components. During the data capturing, the JpcapCaptor class provided by Jpcap enables the capture of network traffic. Typically, an instance object of this class is suited for specific network adapter devices, allowing various specified operations to be carried out. Firstly, network device selection is performed using the JpcapCaptor.getDeviceList() function of Java, which returns an array of NetworkInterface objects, representing the available network devices. Opening the network interfaces is done in a static mode through the openDevice() method, which requires four parameters, including enabling promiscuous mode. Secondly, packet capture is conducted using the processPacket() and loopPacket() callback methods within the JpcapCaptor class. The former is commonly used and supports non-blocking and timeout policies, while the latter does not provide these features. Thirdly, filtering rules can be set using Jpcap to achieve the filtration of unwanted packets. For instance, using an IP filter expression would only retain IP packets. Applying such filter rules not only reduces the system's data processing load but also improves application performance significantly. During the system operation, once network traffic capture is realized, the data packets need to be analyzed and processed. By extracting key header fields and conducting comprehensive data and feature extraction, the groundwork is laid for future anomaly traffic detection.

In the module of data storage, traffic storage mainly lays the data foundation for the data mining model. In the construction of the ODFM system, the open-source MySQL database server is selected as the database management system. To enhance data storage efficiency and application performance, the c3p0 connection pool technology is also utilized. This involves submitting the opening and closing of database connections to the connection pool, which significantly improves the efficiency of data access at the application level. For the data parsing, the c3p0-config.xml is configured with the necessary database parameters. The inclusion of jar packages such as c3p0-0.9.1.2.jar, mysql-connector-java-5.08-bin.jar (the MySQL database driver package), and commons-dbutils-1.4.jar (JDBC encapsulation library) enables simplified development using the dbutils framework. This framework streamlines development by encapsulating result sets, managing resource releases, and facilitating database transactions. The database storage module design primarily focuses on table, trigger, and stored procedure implementations to fulfill the requirements of business needs. Trigger design plays a pivotal role in connecting the P2P traffic identification module with the abnormal traffic detection module. It enables the effective filtering of P2P traffic within transmission control protocol (TCP), thereby enhancing the accuracy of abnormal traffic data processing and reducing false positives. This is accomplished through a trigger that automatically deletes P2P traffic in the tcp_table based on IP addresses when a result is inserted into the p2p_result_table, thereby facilitating data filtering. These implementations are critical in meeting the functional requirements of the ODFM system while optimizing database storage efficiency and supporting accurate traffic analysis.

In the module of feature extraction engine, effective feature extraction provides corresponding feature samples for the traffic classification model. Fig. 3 provides a complete transition diagram of a TCP connection state. During the implementation of this module, statistical analysis of traffic information is conducted, and certain TCP sessions are maintained. For instance, non-three-way handshake RST packets are eliminated. The module's core revolves around the utilization of different network connection states and the sequence of state transformations. This allows for the description of traffic information and enables the incorporation of state statistics mechanisms for analyzing packets in various states. In a TCP connection, if a TCP flow is initiated by the SYN initiator, the first forward starting point is determined. All directions are indicated as forward, denoted by "+", while the direction of packets transmitted by the counterpart is represented by "-". In a complete TCP connection process, both nodes involved in the connection can initiate the connection, resulting in a symmetrical transformation graph. However, statistical analysis is typically performed based on one of the states. A complete TCP connection usually comprises three stages: the three-way handshake, data exchange, and connection release. Generally, only the relevant states are iterated in a self-looping manner. The application of this module allows for a comprehensive analysis of TCP traffic, ensuring accurate tracking of connection states and statistical analysis of corresponding packets. By considering the unique characteristics of each state, the module enables effective traffic analysis and facilitates the identification of potential anomalies or irregularities within TCP connections. During the operation of data feature extraction, this study focuses on analyzing packets that fall within {SYN, SYN/ACK, ACK, RST/ACK, RST, FIN, FIN/ACK, Data} categories. Any malformed packets that do not adhere to the RFC specification are promptly flagged for immediate alarm. In cases where the TCP connection is normal or RST packets are detected, the stream recording is halted, and the packet information is recorded and the connection is cleared. To effectively maintain the same flow information, the TCP packet sequence number is leveraged. Suppose the sequence number is denoted as S, the load size of the current packet is represented as L, and the sequence number of the subsequent TCP packet is $S+L$. This approach allows for accurate tracking and management of the flow information. On the other hand, the feature extraction of P2P traffic identification is relatively straightforward. It involves that extracting the payload length information from each UDP stream and realizing packet statistics using a List set. By focusing on these specific packet categories and applying appropriate techniques for maintaining flow information and

P2P traffic feature extraction, the ODFM streamlines the traffic analysis process and ensures the identification of abnormal or irregular packets while adhering to standardized protocols.

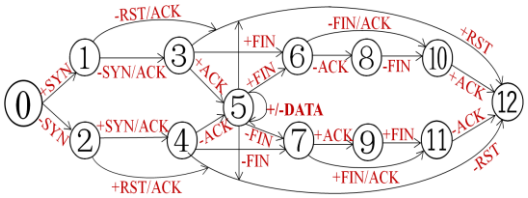


Fig. 3. The transition diagram of a TCP connection state.

In the module of abnormal traffic classification module, we analyze the relevant feature samples extracted by the feature extraction engine module in the traffic classification module to obtain classification results. This allows us to construct a decision tree classifier and apply the AdaBoost algorithm [37] in the anomaly traffic classification module. With this operation, the classifier can be transformed into a strong classifier, which further enhances the accuracy of P2P traffic identification without the need for additional boosting.

In the anomaly traffic classification process, we primarily focus on the off-line training and online classification calculations. For practical classification, we need to develop additional functionality based on the related *jar* packages of *Weka*. Therefore, understanding the underlying algorithms encapsulated in Weka is crucial. Once the data mining process is complete, the anomalous network traffic is passed to the alert module, which generates alert information. The files containing non-anomalous traffic information are then deleted, completing the entire anomaly network traffic detection process. The decision tree training model is stored in memory. By loading the training samples and applying the "*buildClassifier* (Instances instances)" method, the trained classifier model is loaded into memory. Through these procedures, we can effectively analyze and classify abnormal network traffic, which ensures accurate detection and response to potential threats.

III. Experimental Evaluation

In this section, comprehensive experiments are conducted to validate the effectiveness of our proposed ODFM method in the context of anomaly network traffic detection tasks. Stress tests are carried out on each module to analyze the system's traffic processing capabilities. The average processing speed of each module is evaluated and presented in Table I, which highlights the competitive speed of the abnormal traffic classification detection module and indicates favor consumed time when processing a mass of data packets in the modules of traffic collection, parse and store. Although the system design incorporates the AdaBoost modeling strategy [37], which may initially require a longer training time, real-time classification becomes feasible after successfully completing the model building process. The trained model can be serialized in memory, enabling efficient offline training. During real-time classification, only the feature samples need to be directly input for processing.

Furthermore, the system employs multi-threaded parallel computing to pursue efficient processing times. For instance, the processing of 500,000 network data is completed in less than 180 seconds. However, in practical applications, traffic capture and statistics are typically implemented at minute intervals. In general, the statistical interval time above 180s can be considered to meet the real-time requirements of the system. The experimental results indicate the robustness and efficiency of the proposed ODFM method for anomaly network traffic detection, indicating its suitability and effectiveness for real-time applications.

TABLE I. Results of Data Processing of each Module

Module\Parameter	Packets (ten-thousand)	Time (s)
Collection, Parse, Store	19.6	60
Feature extraction engine	51.2	60
Traffic classification detection	100	5.8

IV. Conclusion

In this paper, we propose a novel anomaly traffic detection approach called ODFM that incorporates the optimization of feature dimension reduction and data mining. Different from the traditional methods that take massive data as input and implement the complex design of high-dimensional feature sample to achieve the model training, this paper first adopts the feature selection mechanism to reduce the feature analysis dimension, and sets the P2P traffic identification module to filter the related service traffic, so as to reduce the amount of data detection and improve the detection accuracy. The motivation behind ODFM is that an optimization of approximate feature dimension and normal traffic mining can complete fast anomaly detection while ensuring high detection accuracy. Experiments indicate that the whole pipeline can produce competitive detection results, and it can be able to address various challenging anomaly traffic situations, showing its obvious efficacy in the task of network anomaly traffic detection.

References

[1] Z. Minjie and Z. Yilian, "Abnormal Traffic Detection Technology of Power IOT Terminal Based on PCA and OCSVM," 2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 2023, pp. 549-553.

[2] H. Gong, C. Liu, W. Gao, L. Wang and X. Wang, "MSTP Network Data Traffic Anomaly Optimization Detection Algorithm," 2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS), Chengdu, China, 2023, pp. 32-35.

[3] M. Cao, D. -n. Cheng, X. Wu and B. Wang, "Research on Auto-adaptive Traffic-Aware Abnormal Detection Method," 2009 International Symposium on Information Engineering and Electronic Commerce, Ternopil, Ukraine, 2009, pp. 445-449.

[4] G. Chen and X. Tan, "Network abnormal traffic detection method based on multi kernel KPCA-PSO-ELM," 2021 2nd International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), Shenyang, China, 2021, pp. 183-187.

[5] H. Li, E. He, C. Kuang, X. Yang, X. Wu and Z. Jia, "An Abnormal Traffic Detection Based on Attention-Guided Bidirectional GRU," IEEE 22nd International Conference on Communication Technology (ICCT), Nanjing, China, 2022, pp. 1300-1305.

[6] G. Kaur, V. Saxena and J. P. Gupta, "Anomaly Detection in network traffic and role of wavelets," International Conference on Computer Engineering and Technology, Chengdu, China, 2010, pp. V7-46-V7-51.

[7] P. Kromkowski, S. Li, W. Zhao, B. Abraham, A. Osborne and D. E. Brown, "Evaluating Statistical Models for Network Traffic Anomaly Detection," 2019 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2019, pp. 1-6.

[8] I. C. Paschalidis and G. Smaragdakis, "A Large Deviations Approach to Statistical Traffic Anomaly Detection," Proceedings of the 45th IEEE Conference on Decision and Control, San Diego, CA, USA, 2006, pp. 1900-1905.

[9] H. Zhao et al., "A Fourier Series-Based Anomaly Extraction Approach to Access Network Traffic in Power Telecommunications," 2017 International Conference on Computer Systems, Electronics and Control (ICCSEC), Dalian, China, 2017, pp. 550-553.

[10] S. M. A. Karim, N. Ranjan and D. Shah, "A Scalable Approach to Time Series Anomaly Detection & Failure Analysis for Industrial Systems," Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2020, pp. 0678-0683.

[11] R. Sharma, N. Singh and S. Birla, "An Experimental Study for Comparing Different Method for Time Series Forecasting Prediction & Anomaly Detection," International Conference on Electrical, Computer and Communication Technologies (ICECCT), Erode, India, 2021, pp. 1-4.

[12] A. Guezzaz, Y. Asimi, M. Azrour and A. Asimi, "Mathematical validation of proposed machine learning classifier for heterogeneous traffic and anomaly detection," in Big Data Mining and Analytics, vol. 4, no. 1, pp. 18-24, March 2021.

[13] A. Khudoyarova, M. Burlakov and M. Kupriyashin, "Using Machine Learning to Analyze Network Traffic Anomalies," IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), St. Petersburg, Moscow, Russia, 2021, pp. 2344-2348.

[14] R. Singh, N. Srivastava and A. Kumar, "Machine Learning Techniques for Anomaly Detection in Network Traffic," International Conference on Image Information Processing (ICIIP), Shimla, India, 2021, pp. 261-266.

[15] Y. Sun, H. Ochiai and H. Esaki, "Deep Learning-Based Anomaly Detection in LAN from Raw Network Traffic Measurement," Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, USA, 2021, pp. 1-5.

[16] Y. Li et al., "NIN-DSC: A Network Traffic Anomaly Detection Method Based on Deep Learning," International Conference on Signal and Image Processing (ICSIP), Suzhou, China, 2022, pp. 390-394.

[17] H. Tong, "Research on Multiple Classification Detection for Network Traffic Anomaly Based on Deep Learning," International Symposium on Computer Science and Intelligent Control (ISCSIC), Beijing, China, 2022, pp. 12-16.

[18] D. Yang and M. Hwang, "Unsupervised and Ensemble-based Anomaly Detection Method for Network Security," International Conference on Knowledge and Smart Technology (KST), Chon buri, Thailand, 2022, pp. 75-79.

[19] W. Huan, H. Lin, H. Li, Y. Zhou and Y. Wang, "Anomaly Detection Method Based on Clustering Undersampling and Ensemble Learning," IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 2020, pp. 980-984.

[20] A. S. Varal and S. K. Wagh, "Misuse and Anomaly Intrusion Detection System using Ensemble Learning Model," 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), Bhubaneswar, India, 2018, pp. 1722-1727.

[21] C. F. T. Pontes, M. M. C. de Souza, J. J. C. Gondim, M. Bishop and M. A. Marotta, "A New Method for Flow-Based Network Intrusion Detection Using the Inverse Potts Model," IEEE Transactions on Network and Service Management, vol. 18, no. 2, pp. 1125-1136, June 2021.

[22] M. S. E. Sayed, N. -A. Le-Khac, M. A. Azer and A. D. Jurcut, "A Flow-Based Anomaly Detection Approach with Feature Selection Method Against DDoS Attacks in SDNs," IEEE Transactions on Cognitive Communications and Networking, vol. 8, no. 4, pp. 1862-1880, Dec. 2022.

[23] Z. Jadidi, V. Muthukkumarasamy, E. Sithirasenan and K. Singh, "Flow-based anomaly detection using semisupervised learning," 2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS), Cairns, QLD, Australia, 2015, pp. 1-5.

[24] V. Timcenko and S. Gajin, "Hybrid Machine Learning Traffic Flows Analysis for Network Attacks Detection," Telecommunications Forum (TELFOR), Belgrade, Serbia, 2022, pp. 1-8.

[25] D. Vinod and M. Prasad, "A novel hybrid automatic intrusion detection system using machine learning technique for anomalous detection based on traffic prediction," International Conference on Networking and Communications (ICNWC), Chennai, India, 2023, pp. 1-7.

[26] A. G. Roselin, P. Nanda, S. Nepal and X. He, "Intelligent Anomaly Detection for Large Network Traffic with Optimized Deep Clustering (ODC) Algorithm," in IEEE Access, vol. 9, pp. 47243-47251, 2021.

[27] G. Ping, S. Feng, Y. Li and X. Ye, "Unsupervised Anomalous Traffic Detection Based on Cascading Representation and Multiple-Clustering," IEEE 8th International Conference on Computer and Communications (ICCC), Chengdu, China, 2022, pp. 2303-2307.

[28] A. H. Yaacob, I. K. T. Tan, S. F. Chien and H. K. Tan, "ARIMA Based Network Anomaly Detection," International Conference on Communication Software and Networks, Singapore, 2010, pp. 205-209.

[29] K. M. León-López, F. Mouret, H. Arguello and J. -Y. Tourneret, "Anomaly Detection and Classification in Multispectral Time Series Based on Hidden Markov Models," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-11, 2022.

[30] V. Geppener and B. Mandrikova, "Combination of wavelet transform and Autoencoder for complex data analysis and anomaly detection," 2021 International Conference on Information Technology and Nanotechnology (ITNT), Samara, Russian Federation, 2021, pp. 1-4.

[31] S. Kumar, S. Nandi and S. Biswas, "Research and application of One-class small hypersphere support vector machine for network anomaly detection," International Conference on Communication Systems and Networks (COMSNETS 2011), Bangalore, India, 2011, pp. 1-4.

[32] D. Yao, M. Yin, J. Luo and S. Zhang, "Network Anomaly Detection Using Random Forests and Entropy of Traffic Features," International Conference on Multimedia Information Networking and Security, Nanjing, China, 2012, pp. 926-929.

[33] Z. Zhu, Y. Xie, X. Yang and W. Hu, "A fast anomaly network traffic detection method based on the constrained k-nearest neighbor," 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2023, pp. 318-323.

[34] Z. G. Prodanoff, A. Penkunas and P. Kreidl, "Anomaly Detection in RFID Networks Using Bayesian Blocks and DBSCAN," SoutheastCon, Raleigh, NC, USA, 2020, pp. 1-7.

[35] K. Saha, M. M. Rahman Fakir and M. M. A. Hashem, "An Unsupervised Self-Organizing Map Assisted Deep Autoencoder Gaussian Mixture Model for IoT Anomaly Detection," International Conference on Electrical Information and Communication Technology (EICT), Khulna, Bangladesh, 2021, pp. 1-6.

[36] M. Shen et al., "Machine Learning-Powered Encrypted Network Traffic Analysis: A Comprehensive Survey," IEEE Communications Surveys & Tutorials, vol. 25, no. 1, pp. 791-824, Firstquarter 2023.

[37] W. Li and Q. Li, "Using Naive Bayes with AdaBoost to Enhance Network Anomaly Intrusion Detection," International Conference on Intelligent Networks and Intelligent Systems, Shenyang, China, 2010, pp. 486-489.

Automatic Bangla Image Captioning Based on Transformer Model in Deep Learning

Md. Anwar Hossain[1*], Mirza AFM Rashidul Hasan[2], Ebrahim Hossen[3],
Md Asraful[4], Md. Omar Faruk[5], AFM Zainul Abadin[6], Md. Suhag Ali[7]
Dept. of Information and Communication Engineering, Pabna University of Science and Technology, Pabna, Bangladesh[1,3,4,5,6]
Dept. of Information and Communication Engineering, University of Rajshahi, Rajshahi, Bangladesh[2]
Dept. of Software Engineering, Daffodil International University, Dhaka, Bangladesh[7]

***Abstract*—Indeed, Image Captioning has become a crucial aspect of contemporary artificial intelligence because it has tackled two crucial parts of the AI field: Computer Vision and Natural Language Processing. Currently, Bangla stands as the seventh most widely spoken language globally. Due to this, image captioning has gained recognition for its significant research accomplishments. Many established datasets are found in English but no standard datasets in Bangla. For our research, we have used the BAN-Cap dataset which contains 8091 images with 40455 sentences. Many effective encoder-decoder and Visual Attention approaches are used for image captioning where CNN is utilized for the encoder and RNN is used for the decoder. However, we suggested a transformer-based image captioning model in this study with different pre-train image feature extraction models like Resnet50, InceptionV3, and VGG16 using the BAN-Cap dataset and find out its effective efficiency and accuracy based on many performances measured methods like BLEU, METEOR, ROUGE, CIDEr and also find out the drawbacks of others model.**

Keywords*—Bangla image captioning; image processing; natural language processing; attention mechanism; transformer model

I. Introduction

For image captioning, humans first look at the image and detect the object. Each human brain has a huge local language vocabulary. After detecting the object, it finds out the vocabulary of this respective object and generates a description of this image. This process is easier for humans but it needs some steps for our machine. For machines, it integrates two fundamental components of artificial intelligence, namely Computer Vision (CV) [1] and NLP [2]. When it comes to computer vision, various pre-trained CNN models [3] are employed to extort image features. These appearances are subsequently fed through an RNN [4] to generate captions utilizing the LSTM mechanism [5]. These days, there are several uses for picture captioning, including self-driving cars, social media, security and surveillance, travel and tourism, healthcare, robotics, and many more. Nowadays, the seventh most used language worldwide is Bangla [6]. For the huge number of populations, it is recognizable for significant research.

However, this research can find out the lack of a present days' model where existing model RNN is performed for generating word of sequences and discover the absence of context awareness where the existing model is trained using tokenizer word format without any relative positioning and attention mechanism. In this paper we try to take the following objectives:

- Build up an image captioning sculpt lying on transformers.
- Attention on the context.
- Compare the suggested model performance with other popular images captioning model.

The proposed model's performance undergoes assessment through various Natural Language Processing evaluation methods that rely on both machine-generated captions and reference captions. These methods include BLEU [7], METEOR [8], ROUGE [9], and CIDEr [10].

II. Literature Review

We discuss several approaches and cutting-edge techniques used in Bangla picture captioning in this part. This section is separated into two sections: The visual attention-based method and the CNN-LSTM-based approach. Ultimately, we endeavour to identify the limitations of the prevailing models.

A. CNN-LSTM-based Approach

The initial Bangla image captioning model is constructed based on the CNN-LSTM architecture, as outlined by [11]. These are separated into two parts :1. Image features extraction 2. Language is generated based on the features. The image feature is extracted by the VGG16 model [12]. The model operates with two inputs: the image and the tokens' order(which represent unique words in the dictionary). It employs an embedding layer to derive the respective word embeddings from the tokens. The word embedding layer's output is subsequently compressed to 512 dimensions using a dense layer. This result is reciprocated to the sequence produced by the embedding layer as its output data that is stacked on the LSTM. The stacked LSTM sequence data generate the n-length caption. Here is a blocked diagram Fig. 1.

B. Visual Attention-based Approach

The visual attention-based approach [13] is described in three parts: 1. Image feature extracted by CNN [3] 2. Attention mechanism for getting weighted image features [14]

3. GRU [13] for generating the caption. Here is a block diagram Fig. 2 of this model.

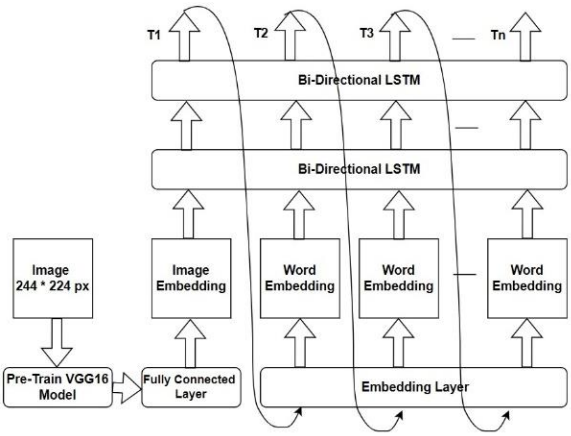


Fig. 1. Image captioning CNN-LSTM model [11].

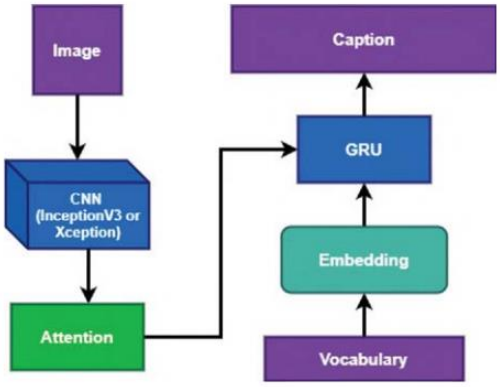


Fig. 2. Visual attention model for image captioning [13].

In this model image features are extracted by the CNN to focus on an important portion image. the image feature vector is directed to the attention mechanism, which generates a context vector specific to this image. In this attention mechanism there are three parts: 1. calculate the alignment score 2. calculate the weight 3. calculate the context vector. For the alignment score, the Eq. (1) is:

$$A_{t,i} = a(s_{t-1}, h_i) \quad (1)$$

Where $A_{t,\,i}$ is the alignment score, s_{t-1} is the previous decoder output, h_i is the encoded hidden state and a() is a function of attention. Calculating the weight is only the output of the softmax function.

$$w_{t,i} = softmax(A_{t,i}) \quad (2)$$

To calculate the context, vector the equation is as follows:

$$C_t = \sum_i^T w_{t,i}\, h_i \quad (3)$$

Where C_t is the context vector. In the Embedding section that creates the vector of words. The GRU [15] received inputs of the context vector along with the word vector from the Embedding layer, subsequently producing the caption.

After understanding on popular model, we can specify some major areas in which we can improve our Bangla image captioning model.

TABLE I. DRAWBACKS OF PREVIOUS MODEL

Model Name	Drawbacks
CNN-LSTM Model [11]	1. Limited Parallelization: Sequential execution to generate caption. 2. Context Understanding: It has no attention mechanism to keep the context of the image. 3. Complexity and Training Efficiency: Require more time to execute.
Visual Attention Model [13]	1. Position Tracking: No track on the position in caption words. 2. Sequential Computation: it has no capability of parallel processing. 3. Scalability and Generalization: GRU-based models may encounter difficulties in scaling to larger and more diverse datasets.

Using our proposed model Fig. 4 we solve the drawbacks which are indicated in Table I. In our proposed model where multi-head self-attention layer has multiple numbers of layers that help to parallel execution and the self-attention mechanism helps to focus on the contextual information in the image which eliminates the limitation of parallelization and context Understanding in the CNN-LSTM model approach. Secondly, In our proposed model (see Fig. 4) the positional embedding layer tracks the position of the input vector and the masked self-attention layer filters the key point of the Bangla caption sentences which eliminates the limitation of position tracking and scalability in the Visual attention model approach. In the word embedding layer that tokenizes the Bangla sentence and turns to convert the respective word vector which reduces the complexity of the model. All the layers are fully described in the methodology section.

III. METHODOLOGY

This section provides an impression of the operational construction of our intended model. That's are divided into two parts: Data Collection and Transformer model. In the data collection section, we describe the procedure of data collection and data pre-processing. In proposed transformer model section, we describe the model architecture.

A. Data Collection

We have used two data sets first for the image dataset Flickr8k [16] which contains 8091 images and second for the Bangla caption using the BAN-Cap dataset [17] which contains 40455 captions. Each image has five captions. Here is a Fig. 3 of the most frequent words in this dataset.

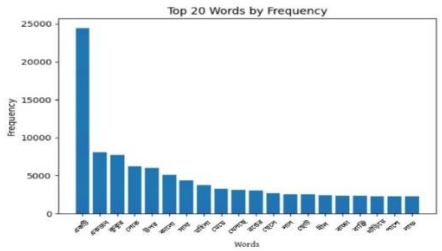


Fig. 3. Top most 20 frequently used word.

B. Transformer Model

The term "Transformers" was first used in the article "Attention is all you need." [18]. It serves as the foundation for some popular versions like GPT-2 [19] and BERT [20]. Language translation models and question-and-answer-based models are two examples of transformer models' many applications. Transformers can be utilized for a variety of application cases because of their versatile architecture**.** For Bangla image captioning, our proposed model, depicted in Fig. 4, is formulated based on the transformer model. We describe our proposed model into two parts: Encoder and Decoder.

C. Encoder

The encoder is tasked with handling the input data. The inputs may be a sequence of words in a natural language sentence, Image data, or any other sequential data. The Encoder is the main work of extracting meaningful representations from the input data that are known as "contextual embeddings". In this section, we describe several parts which are Image Feature Extraction, Positional Embedding, Multi-head self-attention, Add and normalization, and Feed Forward Layer.

1) Image feature extraction: Image feature extraction is a procedure that transforms raw data into a numerical form. Which helps us to identify and capture the relevant patterns of the image. For feature extraction we have used different pre-train models ResNet50[21], Inception V3 [22], and VGG16 models [12]. After the feature extraction, the features are represented with vectors that are shown in Fig. 4.

In our proposed model Fig. 4 we try to solve the drawbacks of the previous model. In this model first, image and captions sequential data are converted into vectors respectively patch embedding and word embedding. To track the position of this sequential data we add a positional encoder that solves the drawbacks of the previous model which are fully described below. After positional encoding the multi-head self-attention it handles every element of the input sequence in parallel, making them more suitable for parallelization during both training and inference. This capacity to handle data in parallel improves training effectiveness and lessens the vanishing gradient issue that helps the limitation of parallelization in the previous model. In the Word embedding layer, we used word tokenizers that easily tokenize the word and convert it into word vectors that store synthetic and semantic information about those words, and all the layers are described briefly in the below section.

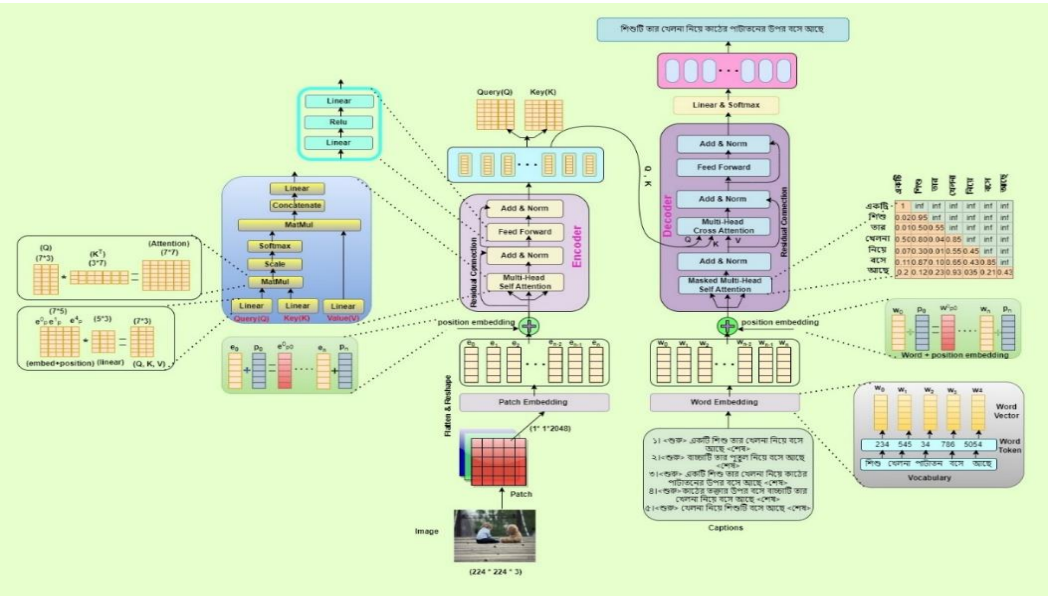


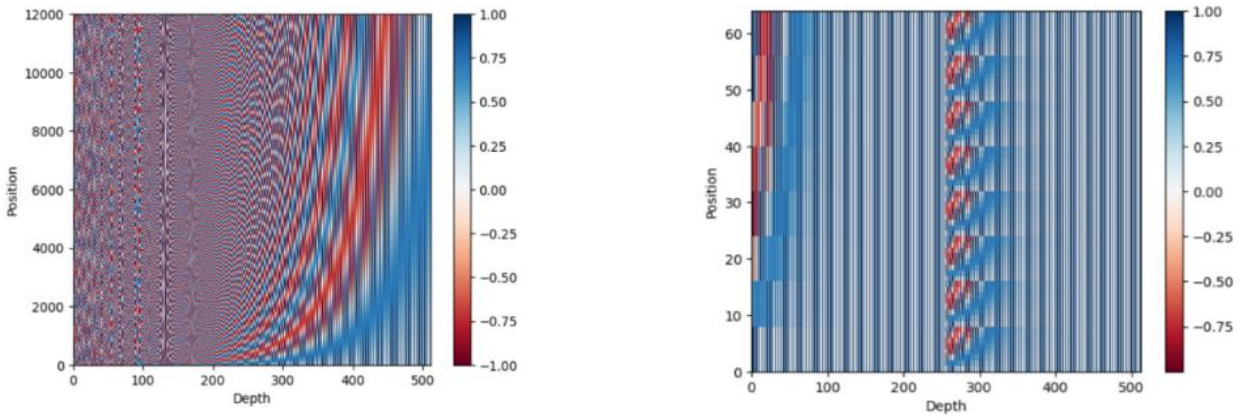
Fig. 4. Diagram of our proposed model.

2) Position embedding: Since ours is devoid of repetition and convolution, to ensure the model effectively utilizes the sequence order, it is vital to furnish precise information about the token positions within the sequence. To achieve this, we incorporate "position encodings" into the input embeddings of both the encoder and decoder layers. For better understanding here an example "রহিমের দুটি ঘর রয়েছে, সে দীর্ঘ দশ বছর ধরে রহিমার সাথে ঘর করছে" In this sentence the first word "ঘর" meaning the number of House but the second word "ঘর" meaning the Family. For this reason, we need to track the word that's called position embedding. In Fig. 4 we see how the position value is added with an image vector in the encoder and a word vector in the decoder to track them. There are a variety of methods of position encoding systems. We use the sine and cosine functions for position encoding. The sine function is used for odd positions that are:

$$P_{(pos,2k)} = sin\left(\frac{pos}{10000^{2k/d}}\right) \quad (4)$$

The cosine function is used for even positions that are a mathematical form:

$$P_{(pos,2k+1)} = cos\left(\frac{pos}{10000^{2k/d}}\right) \quad (5)$$



(a) (b)

Fig. 5. (a) Caption position encoding (b) Image position encoding.

From Fig. 5 we show that the value of each position is represented between 1 to -1. On the graph, Position is shown by the Y axis and the associated position value by the X axis for both image and caption encoding.

3) Multi-head self attention: As the name suggests, Attention means to focus on the input data that are relatively close to the features and therefore establishing a relationship with them. Fig. 6 shows that the main three focusing points of this image are the baby, doll, and wood. There is a caption generated based on those points.

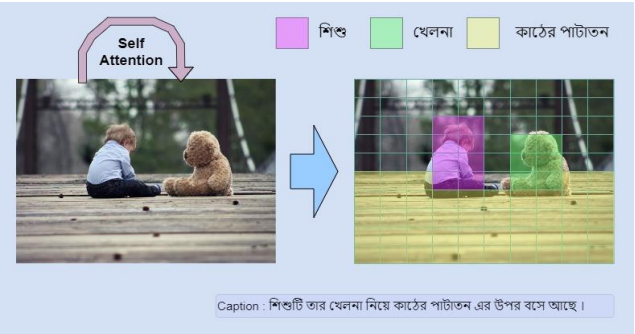


Fig. 6. Self-Attention in an image.

From Fig. 4 we show the mechanism of the multi-head attention layer with a better example. Firstly, the Linear section receives the positional embedding's output. The linear layer functions as a fully connected neural network that performs multiplication with positional encoding, thereby generating a matrix. This matrix is duplicated into the Query, Key, and Value matrices.

- Query(Q) - Represents the whose value needs to be determined
- Key(K) - Represents the features
- Value(V) - Represents the actual value of the input

Secondly, the MatMul section creates an attention-value matrix that is shown in Fig. 4. The mathematical representation is:

$$A_{(Q,K)} = Q.k^T \tag{6}$$

Thirdly, to normalize the attention values within the attention matrix, each value is divided by $\sqrt{dk}$ where (dk) is the dimension of Key metrics. The mathematical representation is:

$$s_i = \frac{A_{(Q,k)}}{\sqrt{d_k}} \tag{7}$$

Fourthly, the S_i value passed into the softmax() and the output vector of the activation function is a matrix multiplied by the Value(V) matrix in the MatMul section and creates an attention-weight matrix. Which attention weight value is high which represents the focus point. Here the final mathematical equation is:

$$A_{(Q,K,V)} = softmax\left(\frac{A_{(Q,K)}}{\sqrt{d_k}}\right).V \tag{8}$$

This attention equation is for one head. For the name multi-head, the concatenate layer Fig. 4 is added to those heads. The mathematical equation is:

$$MultiHead_{(Q,K,V)}$$
$$= concat(head_1, head_2, head_3 \ldots\ldots head_n) \tag{9}$$

n denotes the number of heads.

4) Add & Norm: This Layer does two activities, as its name indicates. The 'Add' portion of the process, which controls flow via residual connections, is the initial phase. 'Norm', the next step, accomplishes layer normalization.

5) Feed forward: This Layer incorporates a fully connected point-wise feed-forward network, employing the ReLU activation function to conduct two linear transformations, as illustrated in Fig. 4. This layer determines the weights used during exercise. The mathematical representation is:

$$FF(x) = ReLU(xw_1 + b_1)w_2 + b_2) \tag{10}$$

Weight matrices denote by W_1 and W_2, and bias denoted by b_1 and b_2, where the ReLU [23] function is :

$$ReLU(x) = max(0, x) \tag{11}$$

D. Decoder

For tasks like language translation, text generator, and text summarization, the decoder in a Transformer is in charge of producing an output sequence. It works in conjunction with the encoder, which analyzes the input sequence. In Fig. 4 Decoder is comprised of: 1. word embedding layer 2. Position embedding 3. Masked multi-head self-attention layer 4. Addition & Normalization 5. Multi-head cross attention 6. Feed Forward layer 7. Linear & SoftMax layer. In the previous Encoder section, we have already explained those layers which similar to both the encoder and decoder. So, in this section, we describe the below layer.

1) Word embedding: It is an essential component of the transformer concept. That is responsible for converting input tokens of words and generating a word vector. This vector records the word's semantic and grammatical details thereby assisting this model in acquiring a meaningful representation of the input text. Fig. 4 shows an example of Bangla words converted into word vectors by tokenization which are detonated by $w_{0.....}$ w_n.

2) Masked multi-head self-attention: This level is essential in the Transformer model because it prevents the model from attending to future locations and maintains the autoregressive characteristic while enabling it to focus on different input sequence segments. After receiving the previous decoder output stack, the first sublayer adds positional information to it and applies self-attention to it. Decoders are altered to focus exclusively on the words that come before them, whilst the encoder is made to pay attention to every word respective to the input sequence of where it appears in the sequence. Consequently, the forecast for a word at a specific position in the sequence can only rely on the known outputs for the preceding words. This is accomplished in the multi-head attention mechanism. Through the application of a mask to the outcomes of the scaled matrix multiplication, the values that would otherwise correspond to prohibited values are suppressed to achieve this masking. Fig. 4 gives a pattern of a mask filter on the words. In Fig. 4 infinity(inf) means that has no probability with the next words and maximum value means the high probability of the next word. Here is a simple representation of the mask filter.

$$mask(Q, K^T) = mask\left(\begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \dots & \dots & \dots & \dots \\ w_{k1} & w_{k2} & \dots & w_{kn} \end{bmatrix}\right) = \begin{bmatrix} w_{11} & inf & \dots & inf \\ w_{21} & w_{22} & \dots & inf \\ \dots & \dots & \dots & \dots \\ w_{k1} & w_{k2} & \dots & w_{kn} \end{bmatrix} \quad (12)$$

E. Model Parameters

We trained our proposed model with different hyper-parameter value that's are indicated on Table II. Those internal variables are adjusted in our model during the training process to minimize the error between predicted and target captions. Those model parameters perfectly capture the features, relation, and pattern of the data.

TABLE II. EXPERIMENTAL PARAMETERS

Parameters name	Value
Vocabulary size	12000
Batch size	64
Buffer size	1000
Dropout rate	0.001
Number of Layer	8
Dimension of model	512
Number of head	8
Number of Epoch	40
Maximum length of sentence	25

IV. RESULT AND DISCUSSION

This part describes the functionality and visualization of the model we've suggested, and then we compare it to other models. Here in Table III, we show our proposed model's performance with different pre-train CNN models. The performance is measured by BLEU [7], METEOR [8], ROUGE [9], and CIDEr [10]. In Table III, the ResNet50+Proposed model gives the maximum output. Ok, Table IV compares our model to several model methodologies.

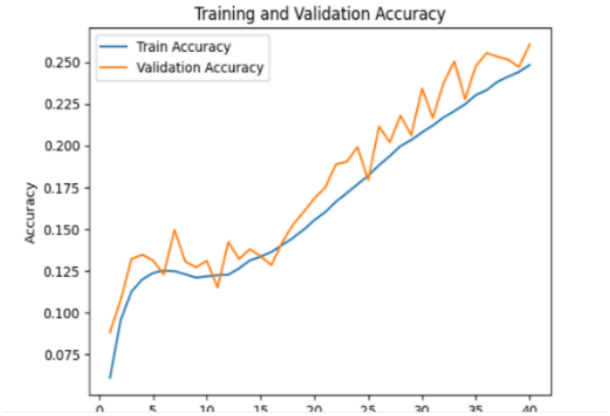
Now, here below in Fig. 7, 8, and 9 show the accuracy and loss curve of our model. The validation curve is represented by a yellow color line and the training curve is represented by a blue color line. The learning curve is the identifier of model overfitting and underfitting evaluation. From Those figures, we see that the accuracy curve of our model is closely together and gradually increasing. Other side, the loss curve is closely together and gradually decreasing. For this, the model is free from overfitting and underfitting issues and called it is called a good fit model. We hope that our model are more balanced fit for large datasets like Flick30k [25].

TABLE III. PERFORMANCE TABLE FOR OUR PROPOSED MODEL

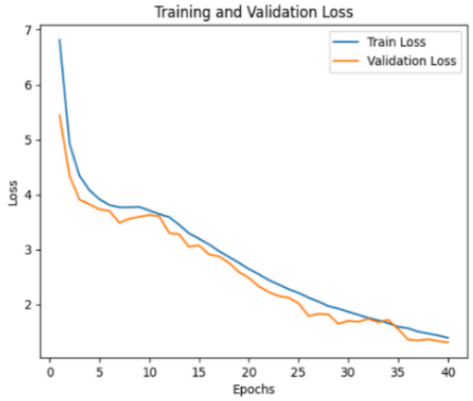
Model	Dataset	BLEU1	BLEU2	BLEU3	BLEU 4	METEOR	ROUGE	CIDEr
Resnet50+ proposed model	Flick8k+ BAN-Cap_captiondata	**64.38**	**58.58**	**43.40**	**24.30**	**0.31**	**0.39**	**0.28**
InceptionV3+ proposed model	Flick8k+ BAN-Cap_captiondata	61.01	56.22	41.03	23.93	0.31	0.41	0.29
VGG16+proposed model	Flick8k+ BAN-Cap_captiondata	60.38	55.44	39.94	21.42	0.29	0.38	0.26

TABLE IV. COMPARE PERFORMANCE WITH OTHER MODEL

Model name	Dataset	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr
CNN-Merge based [24]	Flick8k+ BAN-Cap_captiondata	56.5	35.5	22.1	13.1	0.281	0.290	0.178
Visual-Attention based [13]	Flick8k+ BAN-Cap_captiondata	58.7	36.8	25.4	14.4	0.293	0.288	0.199
Resnet50+ proposed model	Flick8k+ BAN-Cap_captiondata	**64.38**	**58.58**	**43.40**	**24.30**	**0.31**	**0.39**	**0.28**
InceptionV3+ proposed model	Flick8k+ BAN-Cap_captiondata	**61.01**	**56.22**	**41.03**	**23.93**	**0.31**	**0.41**	**0.29**
VGG16+ proposed model	Flick8k+ BAN-Cap_captiondata	**60.38**	**55.44**	**39.94**	**21.42**	**0.29**	**0.38**	**0.26**

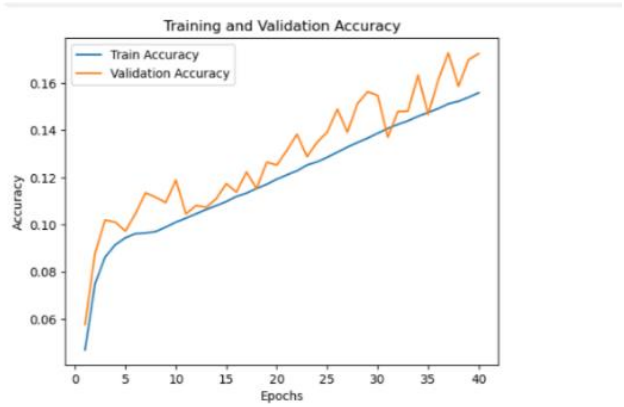


(a)

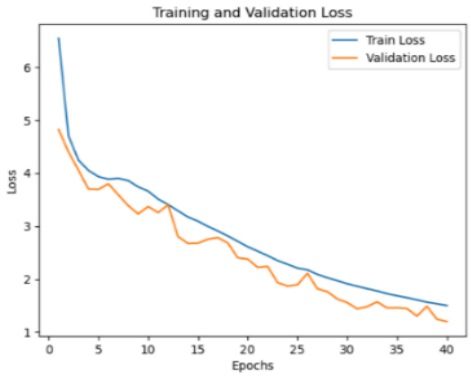


(b)

Fig. 7. (a) Accuracy scheme and (b) Loss scheme for ResNet50+Proposed-model.

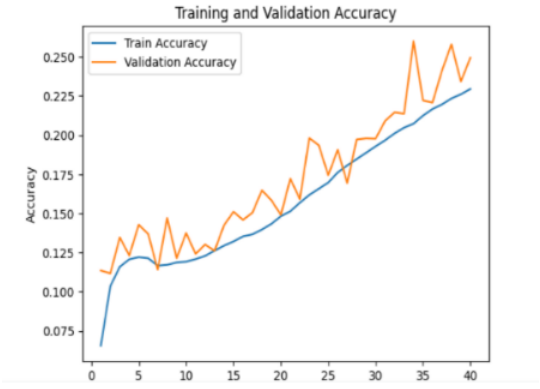


(a)

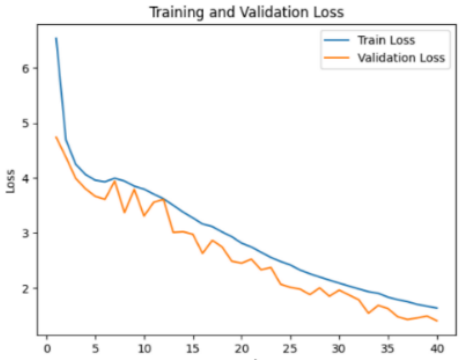


(b)

Fig. 8. (a) Accuracy scheme (b) Loss scheme for Inception V3+Proposed-model.



(a)



(b)

Fig. 9. (a) Accuracy scheme and (b) Loss scheme for VGG16+Proposed-model.

Now, below Fig. 10, Fig. 11, and Fig. 12 are shown some examples of our model prediction with attention mechanism.

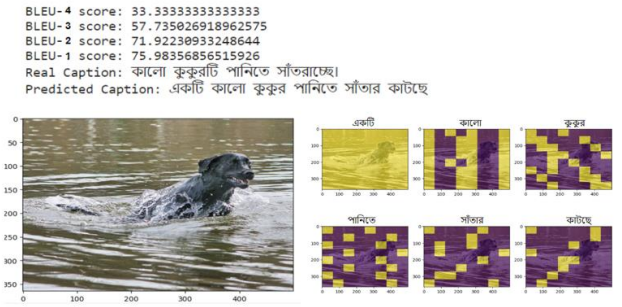


Fig. 10: Performance ResNet50+ Proposed model with Attention Plot

Fig. 11: Performance Inception V3+Proposed model with Attention Plot

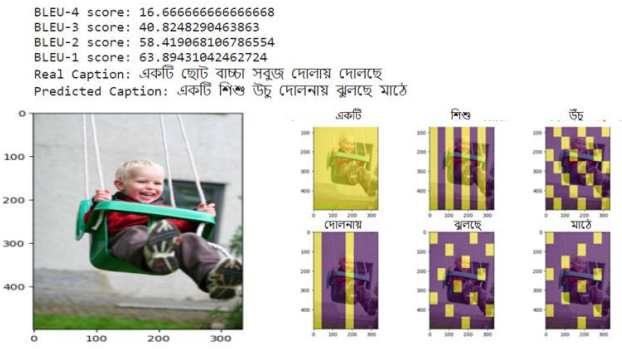


Fig. 12: Performance VGG16+ Proposed model with Attention Plot

V. Conclusions and Future Work

This study presents a transformer-based paradigm for captioning images in Bangla. This proposed model is turned up with better efficiency and performance based on different evaluation methods. It also proves that this model is a good fit for Bangla image captioning on the Flick8k+BAN_Cap dataset. we expect that this paper will help to encourage to others develop more efficient transformer-based models in different NLP and Computer Vision tasks. We also expected that it would be helpful for Image Captioning on higher datasets like Flick30k and others datasets and in the future, we also try to do this.

Acknowledgment

We appreciate the Department of Information and Communication Engineering at Pabna University of Science and Technology for supporting this investigation.

References

[1] Fang, W., Ding, L., Love, P. E., Luo, H., Li, H., Pena-Mora, F., Zhong, B., & Zhou, C. “Computer vision applications in construction safety assurance,” Automation in Construction, 110, 103013. 2020, https://doi.org/10.1016/j.autcon.2019.103013.

[2] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I., “Improving language understanding by generative pre-training,” 2018,

[3] Asraful, M., Hossain, M. A., & Hossen, E. “Handwritten Bengali Alphabets, Compound Characters and Numerals Recognition Using CNN-based Approach,” Annals of Emerging Technologies in Computing (AETiC), 7(3), 60–77. 2023, https://doi.org/10.33166/aetic.2023.03.003.

[4] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. “Learning representations by back-propagating errors,” nature, 323(6088), 533-536, 1986, https://doi.org/10.1038/323533a0.

[5] Hochreiter, S., & Schmidhuber, J. “Long short-term memory. Neural computation,” 9(8), 1735-1780. 1997, https://doi.org/10.1162/neco.1997.9.8.1735.

[6] Szmigiera, M. “Most spoken languages in the world,” Statista. Retrieved Oct. 01, 2023 from https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/.

[7] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. “Bleu: a method for automatic evaluation of machine translation,” Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

[8] Denkowski, M., & Lavie, A. “Meteor universal: Language specific translation evaluation for any target language,” Proceedings of the ninth workshop on statistical machine translation. 2014.

[9] Lin, C.-Y. “Rouge: A package for automatic evaluation of summaries,” Text summarization branches out. 2004.

[10] Vedantam, R., Lawrence Zitnick, C., & Parikh, D. “Cider: Consensus-based image description evaluation,” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[11] Rahman, M., Mohammed, N., Mansoor, N., & Momen, S. “Chittron: An automatic bangla image captioning system,” Procedia Computer Science, 154, 636-642, 2019. https://doi.org/10.1016/j.procs.2019.06.100.

[12] Sutskever, I., Vinyals, O., & Le, Q. V. “Sequence to sequence learning with neural networks,” Advances in neural information processing systems, 27, 2014.

[13] Ami, A. S., Humaira, M., Jim, M. A. R. K., Paul, S., & Shah, F. M. “Bengali image captioning with visual attention,” 2020 23rd International Conference on Computer and Information Technology (ICCIT).

[14] Bahdanau, D., Cho, K., & Bengio, Y. “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv:1409.0473. 2014. https://doi.org/10.48550/arxiv.1409.0473.

[15] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. “Empirical evaluation of gated recurrent neural networks on sequence modeling,” arXiv preprint arXiv:1412.3555. 2014. https://doi.org/ 10.48550/arXiv.1412.3555.

[16] Hodosh, M., Young, P., & Hockenmaier, J. “Framing image description as a ranking task: Data, models and evaluation metrics,” Journal of Artificial Intelligence Research, 47, 853-899. 2013. https://doi.org/10.1613/jair.3994.

[17] Khan, M. F., Shifath, S., & Islam, M. S. “BAN-cap: a multi-purpose English-Bangla image descriptions dataset,” arXiv preprint arXiv:2205.14462.2022. https://doi.org/10.48550/arxiv.2205.14462.

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. “Attention is all you need,” Advances in neural information processing systems, 30. 2017.

[19] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. "Language models are unsupervised multitask learners," OpenAI blog, 1(8), 9, 2019.

[20] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805. 2018 https://doi.org/10.48550/arXiv.1810.04805.

[21] He, K., Zhang, X., Ren, S., & Sun, J. "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition. 2016

[22] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. "Rethinking the inception architecture for computer vision," Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[23] LeCun, Y., Bottou, L., Orr, G. B., & Müller, K.-R. "Efficient backprop," In Neural networks: Tricks of the trade (pp. 9-50). 2002. Springer. https://doi.org/10.1007/3-540-49430-8_2.

[24] Faiyaz Khan, M., Sadiq-Ur-Rahman, S., & Saiful Islam, M. "Improved bengali image captioning via deep convolutional neural network based encoder-decoder model," Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020.

[25] Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," Transactions of the Association for Computational Linguistics, 2, 67-78. 2014 https://doi.org/10.1162/tacl_a_00166.

The Hybrid Jaro-Winkler and Manhattan Distance using Dissimilarity Measure for Test Case Prioritization Approach

Siti Hawa Mohamed Shareef, Rabatul Aduni Sulaiman*, Abd Samad Hasan Basari
Department of Software Engineering, Faculty of Computer Science and Information Technology,
University Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

Abstract*—Software product line (SPL) is a concept that has revolutionized the software development industry. It refers to a set of related software products that are developed from a common set of core assets but can be customized to meet specific customer requirements. Integrating SPL techniques into test case prioritization (TCP) can greatly enhance its effectiveness. By considering variability across different products within an SPL, it becomes possible to prioritize test cases based on their relevance to specific product configurations. However, the concept itself still has certain issues, such as in finding the highest rate of early failure detection. Various solutions have been proposed to mitigate this problem, among them is to improve the calculation of string distance using hybrid technique to achieve a high degree for similarity. Dissimilarity-based Technique (DBP) is the basis for our ranking method. The objective is to identify further weaknesses in the product lines as well as the differences between the experiment and real-world applications. Our focus is to enhance hybrid techniques that produce the highest rate of early failure detection. In this paper, early fault detection is selected as the performance goal. In order to choose the optimal methods for DBP for TCP, a comparison between several string distance measures was conducted. This study proposed hybrid techniques that combined Jaro-Winkler and Manhattan string distance namely New Enhanced Hybrid Technique 1 (NEHT1), New Enhanced Hybrid Technique 2 (NEHT2) and New Enhanced Hybrid Technique 3 (NEHT3). The case study was generated using the PLEDGE tool based on a Feature Model (FM). Six test cases were used in the experiment. Result shows the effectiveness of the combination where it achieved higher degree of similarity for T1 vs. T4, T2 vs. T3, T2 vs. T6, and T3 vs. T6, as well as perfect degree of similarity for NEHT1 (100.00%). The result proves that the combination of both techniques improve SPL testing effectiveness compared to existing techniques.

Keywords*—Test case prioritization; software product line; dissimilarity-based technique; string distance; new enhanced hybrid

I. INTRODUCTION

Software product line (SPL) is a collection of related software products that share a common set of core assets while also offering variations to address diverse customer needs [1]. The characteristics may be constant throughout all SPL-derived products, or they may be varied and present in only some of them [2]. Instead of building each product from scratch, SPL approach emphasizes systematic reuse, enabling efficient development and maintenance of multiple products. SPL streamlines development, reduces redundancy, and enhances consistency across products [3]. Many industries implement SPL due to its ability to handle different phases of development using the commonality and variability concepts [4].

Software product line testing (SPLT) involves testing the shared components and individual product variants within an SPL [5]. It ensures the quality, compatibility, and correctness of both the common core assets and the unique features of each product. This type of testing addresses the challenges posed by varying configurations, shared components, and differing features, while maintaining overall product line quality [6]. Similar to testing in non-configurable code, testing in SPL experiences the coincidental correctness phenomena, which makes it more challenging to detect errors in these systems [7]. However, the testing of a single software system is a highly difficult and expensive stage of the software development process, according to the author [8].

Test case prioritization (TCP) is the process of ordering test cases based on certain criteria to optimize testing efforts [9]. Even though a few trailing test cases are not exercised, these test suites uncover bugs at the earliest possible time [10]. It presents a significant difficulty for software testing [11]. In the context of SPL, prioritization becomes complex due to the diversity of features and configurations [12]. Researchers suggested TCP procedures, where test cases were restructured and carried out in accordance with a given objective, to boost the efficacy and efficiency of testing [13]. A hybrid approach that considers both similarity and dissimilarity should be adopted for effective TCP in SPL development. Similarity refers to the degree to which two or more test cases share common characteristics or requirements while dissimilarity refers to the differences between test cases [14]. Techniques like Jaro-Winkler distance and Manhattan distance can be used to compare test cases for similarity, dependencies, and impact. Prioritizing test cases ensures that critical defects are identified early and that testing resources are allocated efficiently.

SPL has gained prominence in modern software development by allowing the creation of multiple products with shared features and components. TCP plays a vital role in ensuring the quality and reliability of these products. Hybrid string distance, a combination of various string similarity metrics, presents a promising approach to enhancing TCP in

SPL [15]. Hybrid string distance for TCP in SPL faces challenges such as diverse feature sets, low scalability, requires careful consideration of appropriate metrics, and low adaptability to dynamic changes [16]. Managing multiple product variants, selecting appropriate metrics, and ensuring the hybrid approach remains effective and adaptable to evolving requirements are essential for successful implementation. To optimize outcomes in TCP using hybrid string distance in SPL, several objectives should be pursued including comprehensive metric selection, feature-driven prioritization, and scalable algorithm design, adaptability to changes, and empirical validation and evaluation. The hybrid approach should consider specific product variant features, create a mechanism for prioritizing test cases based on these features, and ensure scalability without compromising performance. Finally, the approach should be tested on real-world SPLs to demonstrate improvements in TCP accuracy, coverage, and overall software quality.

This paper addresses the limitations of current TCP methods that struggle to accurately gauge the semantic similarity between test cases, resulting in less than optimal prioritization outcomes. To tackle this issue, the study poses a research question: "Which new hybrid technique can offer the highest early failure detection rate in TCP?"

The research introduces an innovative TCP approach that combines Hybrid Jaro-Winkler and Manhattan distance, integrating a dissimilarity measure. The main contribution lies in enhancing the precision and efficacy of TCP. This technique is used to overcoming the difficulties linked to precisely measuring semantic similarity. This method aims to advance software testing by significantly enhancing prioritization outcomes by providing a more robust and dependable approach for early failure detection in TCP.

The following section outlines the relevant literature. Section III present the proposed approach in detailed, incorporating the experimental settings and a combination of string distance measures whereas Section IV discuss on the results and discussion are presented, leading to the conclusions, in Section V.

II. Related Work

Incorporating TCP techniques like reordering test cases based on fault detection rate can significantly enhance the effectiveness of software testing by enabling early fault detection [17]. In the context of TCP for SPL, the term string distance refers to the measurement of the similarity or dissimilarity of various strings that stand in for test cases [8]. With this method, test cases are ranked according to how distinctive or diverse they are from one another in terms of the testing functionality, or the areas of the code covered. String distance allows for better resource allocation by identifying redundant or overlapping test cases [18]. However, according to Halim et al. [1], neglecting string distance would result in inefficient testing processes and delayed bug fixes. Therefore, incorporating string distance in TCP is essential for efficient software development [19, 20].

Similarity-based prioritization (SBP) focuses on identifying test cases that are similar to each other based on certain criteria, such as code coverage or functionality [1, 21]. The idea behind SBP is that if one test case covers a particular aspect of the software, then similar test cases are likely to cover the same aspect as well [22]. This approach is effective in reducing redundancy in testing efforts by selecting a representative subset of test cases [23]. However, dissimilarity-based prioritization (DBP) considers the diversity among test cases. DBP aims to select a diverse set of test cases that covers different aspects of the software under test [23, 24]. By considering dissimilarities between test cases, this approach ensures comprehensive coverage and reduces the risk of missing critical defects [25].

Sulaiman et al. [25] suggested a measurement based on maximal distance of dissimilarity measure for SPL, which assures thorough coverage and lowers the possibility of overlooking important faults. The study is based on the test case generated from a statechart in comparison to current work, which is based on the FM in the context of the SPL domain. By increasing string distance and prioritizing based on similarity, Halim et al. [1] suggested rearranging test cases to increase the rate of problem identification. The work compared various string distance measures and prioritization algorithms in order to determine the best methods for similarity-based on hybridization of Jaro-Winkler and Hamming distance equation.

Fault detection has been improved in existing studies via the use of new and enhanced hybrid techniques for string distance equations. Recent work by Pospisil et al. [26] aimed to enhance adaptive random TCP for model-based test suites using original technique for Jaccard, Manhattan distance and similarity functions. All of the examined systems achieved improved fault detection performance as a result of the proposed improvement. Another study by Kumar et al. [9] employed Item-based Collaborative Filtering (ICF) to prioritize and decrease the number of products before testing. Hamming string distance was used to calculate the degree of similarity between products. Results of the study show that this approach was able to reduce test suite size. Compared to the works by Pospisil et al. [26] and Kumar et al. [9], the current study concentrated more on using a hybrid string distance method to determine the degree of dissimilarity and then locate the distance with the greatest similarity reading.

III. Proposed Approach

The ranking method we use is based on dissimilarity. Our objective is to find further weaknesses in the product lines being evaluated as well as the point of difference between test case and real world. The study concentrates on the following research question:

RQ1: Which new enhanced hybrid technique produces the best early failure detection rate?

We start by outlining the conditions of our experiment before going on to describe the findings.

A. Experimental Settings

The experiment was carried out on Windows 11 with an AMD Ryzen 5 5625U processor running at 2.30 GHz and 8GB of RAM. The authors developed a New Enhanced Hybrid Techniques (NEHT) by improving string distance using three

hybrid techniques to evaluate the comparability of similarity and dissimilarity measures. For the purpose of generating configuration and prioritizing processes, this technique's similarity and dissimilarity measures will be assessed using current Feature Model (FM), Software Product Line Online Tool (SPLOT) and Product Line EDitor and tests GEneration (PLEDGE) tools. In SPL, FM allows for the systematic representation and management of features, their dependencies, and variations across different products [27]. SPLOT is a web-based tool that allows users to create incredibly dynamic Ajax-based setup and reasoning user interfaces [28], while PLEDGE is an open-source tool that selects and prioritizes product configurations, maximizing the feature interactions covered [29]. In order to test the SPL, the author selects an FM for machine learning based on the Global Positioning System (GPS) created by Saini et al. 2023 [8] as in Fig. 1. Due to the fact that not all possible feature combinations are viable, feature diagrams are used to limit the variety of a product line. Based on the FM in Fig. 1, the *.xml* files will be produced using SPLOT. The *.xml* file will be used to generate the six test cases displayed in Table I after being run using PLEDGE. An ordered list of configurations is often the outcome of a sampling method.

B. Hybrid of String Distance

The purpose of the proposed approach is to find dissimilarity between two test cases. Two strings distances were chosen to develop the proposed approach which is Jaro-Winkler and Manhattan distances. Jaro-Winkler distance is a string distance algorithm that measures the similarity between two strings [30]. It has been widely used in various fields, including TCP. Meanwhile, Manhattan distance is a popular metric used in TCP and works by first creating a matrix of all possible pairs of test cases [15]. It is used to measure the distance between two points on a grid-like system, where the distance is calculated by adding the absolute differences of those coordinates. In software testing, this metric helps prioritize test cases based on their proximity to each other.

The selection of the Hybrid Jaro-Winkler and Manhattan Distance Using Dissimilarity Measure for TCP Approach is grounded in its distinctive ability to address the challenges prevalent in existing TCP approaches. The hybrid nature of the chosen method combines the strengths of Jaro-Winkler and Manhattan string distance, offering a comprehensive solution for accurately capturing semantic similarity between test cases. The integration of a dissimilarity measure further enriches the approach, enhancing the precision of TCP. The decision to adopt this method is motivated by its potential to significantly improve prioritization results and contribute to more effective early failure detection in TCP.

By improving two string distance techniques, this method will produce a new hybrid technique that is precise in obtaining faster early failure detection rate. Fig. 2 describes the combinations of two string distance to develop the three new enhanced hybrid techniques. New enhanced hybrid technique 1 (NEHT1) modifies existing Jaro equation, and Manhattan equation replaces value of m and t with value of test cases *(T1, T2)*. New enhanced hybrid technique 2 (NEHT2) combines Jaro-Winkler and Manhattan equations, replaces m value with n value and t with value of test cases *(T1, T2)*, adds value of test cases, divides with n value and multiply with 1-dj. New enhanced hybrid technique 3 (NEHT3) combines Jaro and Manhattan equations where the formula replaces value of m with n and t with value of test cases *(T1, T2)*.

TABLE I. CONFIGURATIONS OF GPS FEATURE MODEL

Test Case	Configuration
T1	{GPS, Routing, Traffic Avoiding, Interface, Auto-rerouting, Screen, Touch}
T2	{GPS, Routing, Radio, Interface, 3D Map, AM, FM, Digital, Keyboard, Screen, LCD}
T3	{GPS, Routing, Traffic Avoiding, Radio, Interface, Auto-rerouting, AM, FM, Digital, Screen, LCD}
T4	{GPS, Routing, Interface, 3D Map, Keyboard, Screen, LCD}
T5	{GPS, Routing, Traffic Avoiding, Interface, 3D Map, Auto-rerouting, Screen, Touch}
T6	{GPS, Routing, Radio, Interface, Auto-rerouting, AM, FM, Digital, Keyboard, Screen, Touch}

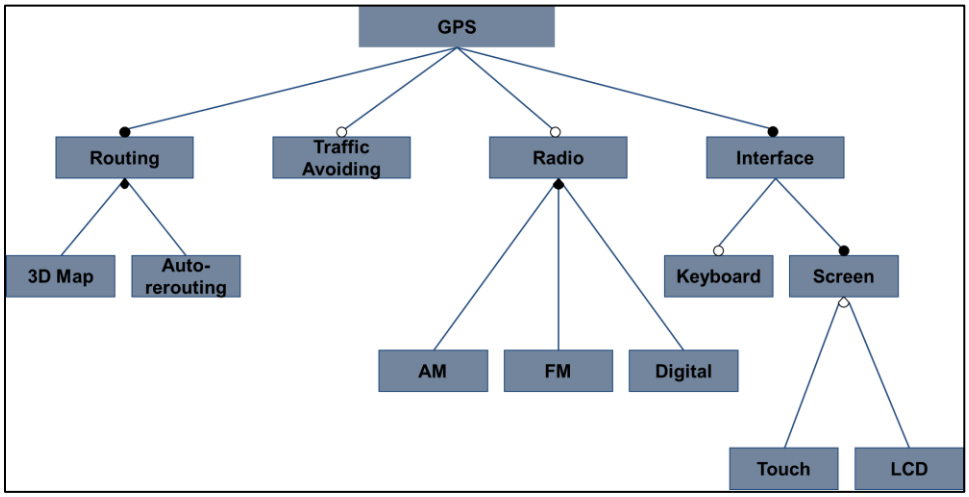


Fig. 1. Feature model of GPS [8].

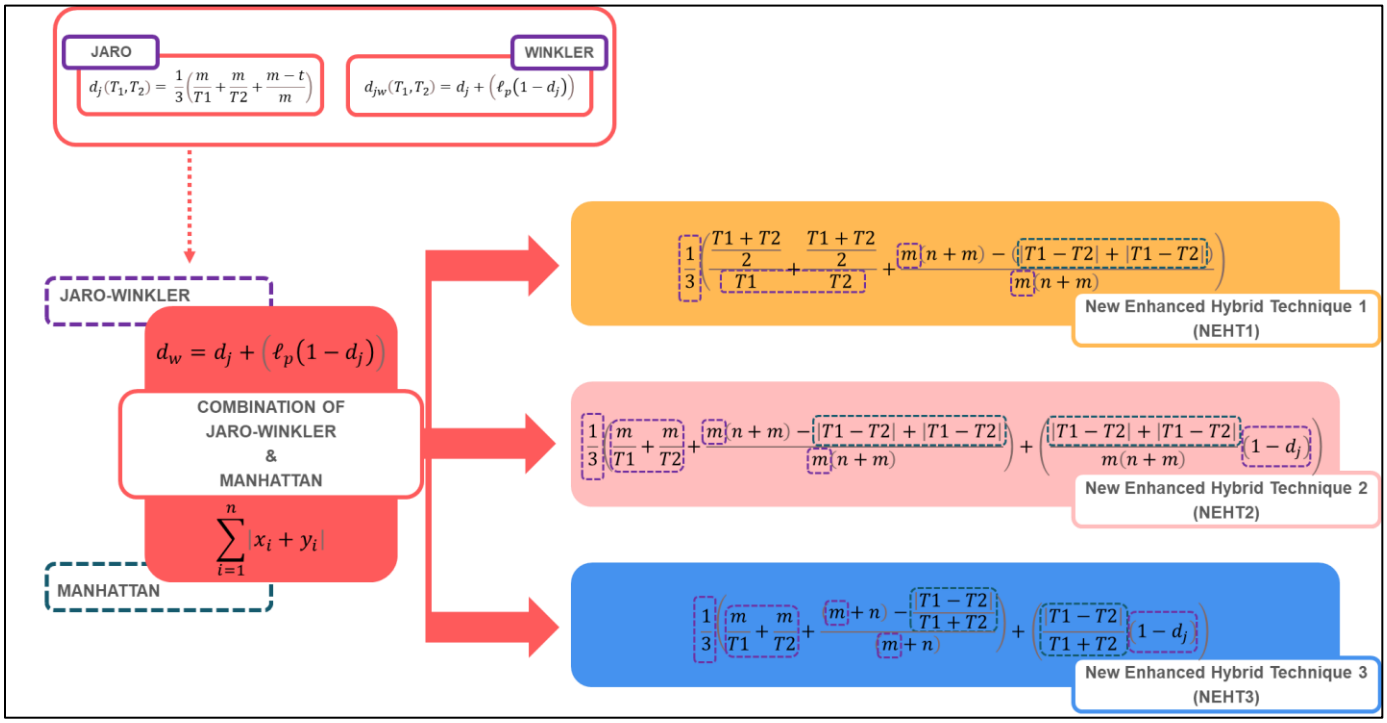


Fig. 2. New enhanced hybrid techniques.

IV. Result and Discussion

Table II shows the similarity and dissimilarity percentages between different pairs of test cases *(T1, T2, T3, T4, T5, T6)*, with NEHT1, NEHT2, and NEHT3 representing different methods or conditions. The values range from 0% *(complete dissimilarity)* to 100% *(complete similarity)*. Since this is an initial result, results use a single FM to represent a dataset. For NEHT1, T1 vs T4, T2 vs T3, T2 vs T6, and T3 vs T6 recorded complete similarity (100.00%), proving the formula is very effective in similarity calculation. Values for NEHT2 and NEHT3 were similar in T1 vs. T4, T2 vs. T3, T2 vs. T6, and T3 vs. T6, which means both proposed techniques provide a consistent way to determine similarity level. The majority of the results show that NEHT1 is effective at determining the degree of similarity.

TABLE II. Calculation for Degrees of Similarity and Dissimilarity

Test Case	NEHT1	NEHT2	NEHT3	NEHT1	NEHT2	NEHT3
	Similarity (%)			*Dissimilarity (%)*		
T1 vs T2	86.79	82.25	70.95	13.21	17.75	29.05
T1 vs T3	97.90	83.40	83.79	2.10	16.60	16.21
T1 vs T4	100.00	71.42	71.42	0.00	28.58	28.58
T1 vs T5	99.56	95.92	95.95	0.44	4.08	4.05
T1 vs T6	97.90	83.40	83.79	2.10	16.60	16.21
T2 vs T3	100.00	87.87	87.87	0.00	12.13	12.13
T2 vs T4	99.65	89.26	89.99	0.35	10.74	10.01
T2 vs T5	93.70	76.68	73.27	6.30	23.32	26.73
T2 vs T6	100.00	87.87	87.87	0.00	12.13	12.13
T3 vs T4	94.57	79.68	77.49	5.43	20.32	22.51
T3 vs T5	96.94	79.87	79.59	3.06	20.13	20.41
T3 vs T6	100.00	87.87	87.87	0.00	12.13	12.13
T4 vs T5	98.81	78.95	79.21	1.19	21.05	20.79
T4 vs T6	94.57	79.68	77.49	5.43	20.32	22.51
T5 vs T6	96.94	79.87	79.59	3.06	20.13	20.41

Fig. 3 and Fig. 4 show similarity and dissimilarity rates of fault detection for the proposed methods *(NEHT1, NEHT2, NEHT3)*. The similarity percentages vary for different methods and test cases. Similar variations may be seen in the dissimilarity percentage, which illustrates how the different approaches of assessing differences differ. There is no uniform trend in how the methods rank similarity or dissimilarity across all test cases. Some test cases consistently show high similarity across all methods, while others show varying degrees of dissimilarity. The author claims that this enhancement will increase the SBP technique's effectiveness [1]. Sulaiman et al. [25] stated that similarity and dissimilarity strategies were introduced to tackle scalability problem in the current priority technique. This method provides a straightforward, scalable, and efficient method for prioritizing and reducing the number of test cases. TCP for SPL can be significantly improved by leveraging high similarity in calculation of string distance. High similarity values are advantageous in locating similar test cases across various SPL. As a result, fewer testing efforts are duplicated, and the existing test cases can be reused. Furthermore, low dissimilarity values can improve coverage, ensure effective bug correction, and improve the fault localization.

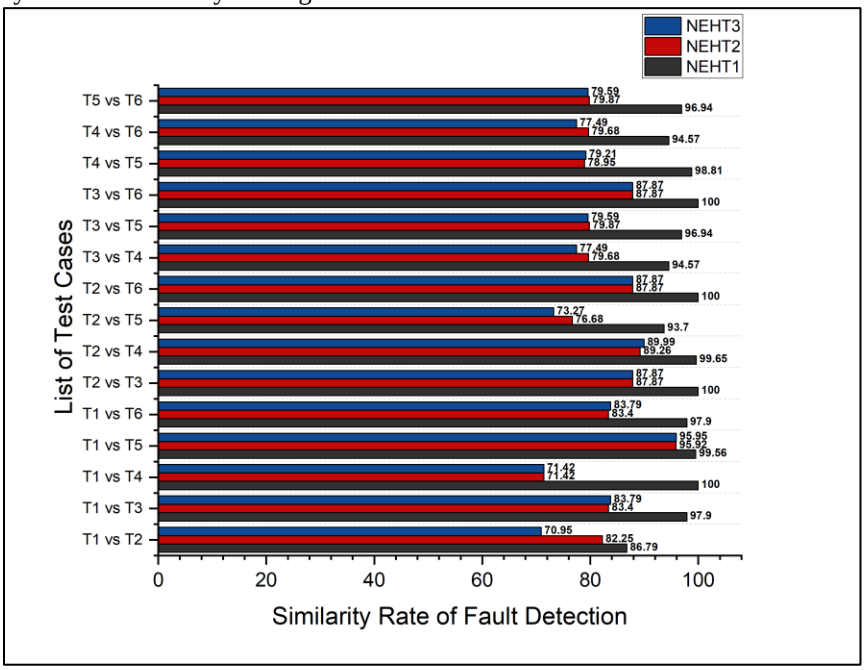


Fig. 3. Similarity rate of fault detection result.

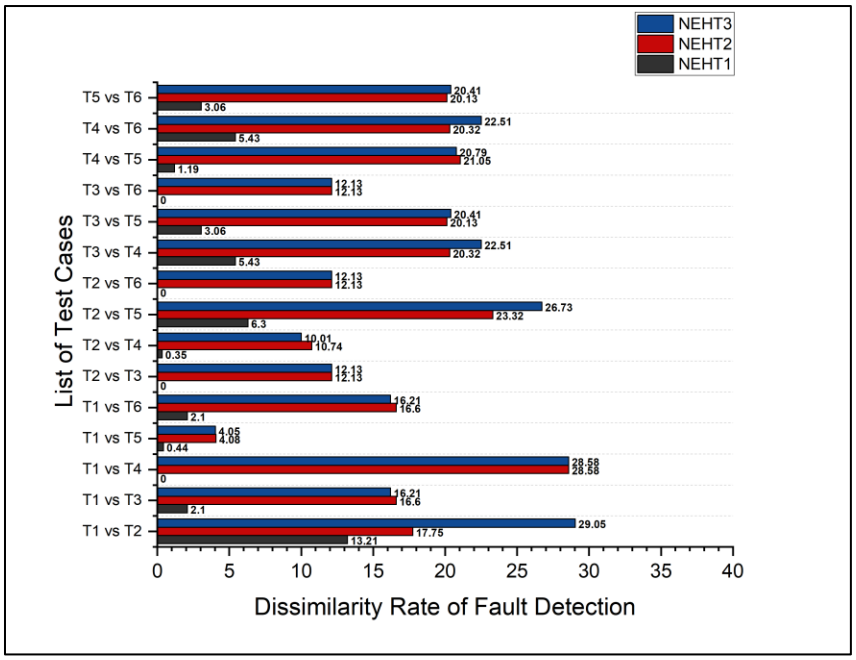


Fig. 4. Dissimilarity rate of fault detection result.

V. CONCLUSION

One of the main advantages of SPLT is its ability to save time and resources. Testers can concentrate on the common characteristics shared by all products in the software family rather than testing each product individually. This makes it possible to quickly find and fix errors, at the same time shortens the development process and lowers expenses. For DBP, dissimilarity test case has been proven to be one of the techniques that can speed up failure detection process. This research employed six different test cases to be tested using three proposed hybrid techniques based on the combination of Jaro-Winkler and Manhattan string distances for early fault identification rate. The findings indicated that NEHT1 has a higher rate of fault identification compared to the other two proposed techniques. In order to increase the success rate of NEHT1's fault identification, we plan to make improvements to it in the future. In addition, we intend to use a variety of case study types for this research project. The limitations of current test case prioritization methods, particularly their struggles in accurately capturing semantic similarity, render them unsuitable for the challenges at hand. Traditional approaches fall short in providing a comprehensive solution for the nuanced characteristics of test cases. The proposed method is chosen to overcome these limitations by introducing a hybrid technique that specifically addresses the semantic aspects of test cases. This strategic choice is aimed at mitigating the deficiencies of existing methods and advancing the field of TCP towards a more precise and effective paradigm.

ACKNOWLEDGMENT

This research was supported by Ministry of Higher Education (MOHE) through Fundamental Research Grant Scheme (FRGS/1/2022/ICT01/UTHM/03/2).

REFERENCES

[1] S. A. Halim, D. N. A. Jawawi, and M. Sahak, "Similarity Distance Measure and Prioritization Algorithm for Test Case Prioritization in Software Product Line Testing," Journal of Information and Communication Technology, vol. 18, no. 1, pp. 57–75, 2019, doi: 10.32890/jict2019.18.1.8281.

[2] T. Ferreira, S. R. Vergilio, and M. Kessentini, "Variability testing of software product line: A preference-based dimensionality reduction approach," Inf Softw Technol, vol. 152, no. September 2021, p. 107031, 2022, doi: 10.1016/j.infsof.2022.107031.

[3] S. Langstrom, "An Investigative Study of Testing Strategy and Test Case Creation in a Hardware-Software Co-design Environment Using Software Product Line Theory," Open Access in DiVA, 2021.

[4] R. A. Sulaiman, D. N. A. Jawawi, and S. A. Halim, "Cost-effective test case generation with the hyper-heuristic for software product line testing," Advances in Engineering Software, vol. 175, Jan. 2023, doi: 10.1016/j.advengsoft.2022.103335.

[5] P. Martou, K. Mens, B. Duhoux, and A. Legay, "Test scenario generation for feature-based context-oriented software systems," Journal of Systems and Software, vol. 197, Mar. 2023, doi: 10.1016/j.jss.2022.111570.

[6] J. Lee, S. Kang, and P. Jung, "Test coverage criteria for software product line testing: Systematic literature review," Inf Softw Technol, vol. 122, no. December 2019, p. 106272, 2020, doi: 10.1016/j.infsof.2020.106272.

[7] T. T. Nguyen, K. T. Ngo, S. Nguyen, and H. D. Vo, "Detecting false-passing products and mitigating their impact on variability fault localization in software product lines," Inf Softw Technol, vol. 153, Jan. 2023, doi: 10.1016/j.infsof.2022.107080.

[8] A. Saini, Rajkumar, A. Kumari, and S. Kumar, "A Proposed Method of Machine Learning based Framework for Software Product Line Testing," 2022 International Conference on 4th Industrial Revolution Based Technology and Practices, ICFIRTP 2022, pp. 10–13, 2022, doi: 10.1109/ICFIRTP56122.2022.10059409.

[9] S. Kumar, Rajkumar, and M. Rani, "Collaborative Filtering-based Test Case Prioritization and Reduction for Software Product-Line Testing," in IEEE Region 10 Annual International Conference, Proceedings/TENCON, Institute of Electrical and Electronics Engineers Inc., Oct. 2019, pp. 498–503. doi: 10.1109/TENCON.2019.8929705.

[10] A. D. Shrivathsan et al., "Novel Fuzzy Clustering Methods for Test Case Prioritization in Software Projects," Symmetry (Basel), vol. 11, no. 11, Nov. 2019, doi: 10.3390/sym11111400.

[11] Z. Q. Zhou, C. Liu, T. Y. Chen, T. H. Tse, and W. Susilo, "Beating Random Test Case Prioritization," IEEE Trans Reliab, vol. 70, no. 2, pp. 654–675, 2021, doi: 10.1109/TR.2020.2979815.

[12] I. Hajri, A. Goknil, F. Pastore, and L. C. Briand, "Automating system test case classification and prioritization for use case-driven testing in product lines," Empir Softw Eng, vol. 25, no. 5, pp. 3711–3769, Sep. 2020, doi: 10.1007/s10664-020-09853-4.

[13] M. L. Mohd-Shafie, W. M. N. W. Kadir, H. Lichter, M. Khatibsyarbini, and M. A. Isa, "Model-based test case generation and prioritization: a systematic literature review," Softw Syst Model, vol. 21, no. 2, pp. 717–753, Apr. 2022, doi: 10.1007/s10270-021-00924-8.

[14] S. Akhmedova, V. Stanovov, and Y. Kamiya, "A Hybrid Clustering Approach Based on Fuzzy Logic and Evolutionary Computation for Anomaly Detection," Algorithms, vol. 15, no. 10, Oct. 2022, doi: 10.3390/a15100342.

[15] M. Khatibsyarbini, "A Study of Test Case Prioritization Technique Based on String Distance Metrics," Universiti Teknologi Malaysia, Johor Bahru, 2019.

[16] U. Markiegi, A. Arrieta, L. Etxeberria, and G. Sagardui, "Dynamic test prioritization of product lines: An application on configurable simulation models," Software Quality Journal, vol. 29, no. 4, pp. 943–988, Dec. 2021, doi: 10.1007/s11219-021-09571-0.

[17] T. K. Akila and M. Arunachalam, "Test case prioritization using modified genetic algorithm and ant colony optimization for regression testing," International Journal of Advanced Technology and Engineering Exploration, vol. 9, no. 88, pp. 384–400, Mar. 2022, doi: 10.19101/IJATEE.2021.874727.

[18] M. Khatibsyarbini, M. A. Isa, D. N. A. Jawawi, H. N. A. Hamed, and M. D. Mohamed Suffian, "Test Case Prioritization Using Firefly Algorithm for Software Testing," IEEE Access, vol. 7, pp. 132360–132373, 2019, doi: 10.1109/ACCESS.2019.2940620.

[19] S. Khoshmanesh and R. Lutz, "Does Link Prediction Help Detect Feature Interactions in Software Product Lines (SPLs)?," in Proceedings - 7th International Workshop on Artificial Intelligence and Requirements Engineering, AIRE 2020, 2020, pp. 87–90. doi: 10.1109/AIRE51212.2020.00020.

[20] C. Birchler, S. Khatiri, P. Derakhshanfar, S. Panichella, and A. Panichella, "Single and Multi-objective Test Cases Prioritization for Self-driving Cars in Virtual Environments," Proc ACM Hum Comput Interact, vol. 5, no. CSCW1, Apr. 2021, doi: 10.1145/1122445.1122456.

[21] S. Ali et al., "Towards Pattern-Based Change Verification Framework for Cloud-Enabled Healthcare Component-Based," IEEE Access, vol. 8, pp. 148007–148020, 2020, doi: 10.1109/ACCESS.2020.3014671.

[22] H. Hemmati, "Advances in Techniques for Test Prioritization," in Advances in Computers, Academic Press Inc., 2019, pp. 185–221. doi: 10.1016/bs.adcom.2017.12.004.

[23] S. Lity, M. Nieke, T. Thum, and I. Schaefer, "Retest test selection for product-line regression testing of variants and versions of variants," Journal of Systems and Software, vol. 147, pp. 46–63, Jan. 2019, doi: 10.1016/j.jss.2018.09.090.

[24] E. Ufuktepe and T. Tuglular, "Application of the law of minimum and dissimilarity analysis to Regression Test Case Prioritization," IEEE Access, vol. 11, pp. 57137–57157, 2023, doi: 10.1109/ACCESS.2023.3283212.

[25] R. A. Sulaiman, D. N. A. Jawawi, and S. A. Halim, "A Dissimilarity with Dice-Jaro-Winkler Test Case Prioritization Approach for Model-

Based Testing in Software Product Line," KSII Transactions on Internet and Information Systems, vol. 15, no. 3, pp. 932–951, Mar. 2021, doi: 10.3837/tiis.2021.03.007.

[26] T. Pospisil, J. Sobotka, and J. Novak, "Enhanced Adaptive Random Test Case Prioritization for Model-Based Test Suites," Acta Polytechnica Hungarica, vol. 17, no. 7, pp. 125–144, 2020, doi: 10.12700/APH.17.7.2020.7.7.

[27] M. Al-Hajjaji, T. Thum, M. Lochau, J. Meinicke, and G. Saake, "Effective product-line testing using similarity-based product prioritization," Softw Syst Model, vol. 18, no. 1, pp. 499–521, 2019, doi: 10.1007/s10270-016-0569-2.

[28] M. Mendonca, M. Branco, and D. Cowan, "S.P.L.O.T. - Software product lines online tools," Proceedings of the Conference on Object-Oriented Programming Systems, Languages, and Applications, OOPSLA, no. May 2014, pp. 761–762, 2009, doi: 10.1145/1639950.1640002.

[29] C. Henard, M. Papadakis, G. Perrouin, J. Klein, and Y. Le Traon, "PLEDGE: A Product Line Editor and Test Generation Tool," ACM International Conference Proceeding Series, pp. 126–129, 2013, doi: 10.1145/2499777.2499778.

[30] J. M. Keil, "Efficient Bounded Jaro-Winkler Similarity Based Search," Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI), vol. P-289, pp. 205–214, 2019, doi: 10.18420/btw2019-13.

A Novel CNN-based Model for Medical Image Registration

Hui GAO, Mingliang LIANG*
Zhengzhou Railway Vocational and Technical College, Henan, 451460, China

***Abstract*—The registration of the deformable image is applied widely to image diagnosis, the monitoring of the disease, and the navigation of the surgery with the aim of learning the correspondence of the anatomist among an image of motion and an image of static. The procedure of the registration of an image mainly includes three steps: the creation of a model of the deformation, a function design for the mensuration of the similarity, and the step of learning for the optimization of the parameter. In the current article, 2-stream architecture is designed, which has the ability to sequentially estimate the fields of the registration of the multi-level by a couple of the pyramids of the feature. In this paper, a 3D network of the encoder-decoder with the 2-stream is designed, which calculates 2 pyramids of the feature of the convolutional as separately by 2 volumes of the input. Also, the registration of the pyramid of the sequential is proposed, which in it, a trail of the modules of the pyramid registration (PR) for the prediction of the fields of the registration of the multi-level is designed as straight by the pyramids of the feature of the decoding. In addition, the modules of PR can be augmented with the computation of the 3D correlations of the local among the pyramids of the feature, which this work leads to the further improvement of the presented approach. Thus, it is capable of collecting the detailed anatomical structure of the brain. The proposed method is tested in three criterion datasets about the registration of MRI of the brain. The evaluation outcomes display that the presented approach outperforms the advanced approaches with a big value.**

***Keywords*—Image registration; convolutional neural network; Pyramid Registration (PR); encoder-decoder**

I. INTRODUCTION

Today, various imaging methods such as MRI, CT, PET, SPECT, and ultrasound imaging are used for the identification of the anatomical structure and the physiological performance of the human body so that each one of them provides specific information to doctors [1], [2]. The medical images usually do not match each other, and to use their information together, it is necessary to apply the methods of the registration of the image [3], [4], [5], [6], [7]. The registration of the image is the procedure of the orientation of the spatial of 2 images from the same scene in a way that their features can be easily related to each other. These images may be created by different sensors or one sensor at different times [3].

Due to the used images, the registration of medical images can be divided into two types: (1) the registration of the same type; (2) the registration of the different types [3]. In the registration of the same type, the used images are prepared by the same type of sensor. For example, to check the drug performance on cancer tumors, the MRI image is recorded from the patient at certain time intervals. With the registration of these images, it is possible to check the progress of the disease or the progress of its improvement [8]. However, in the registration of the different types, the goal is the registration of the recorded images by two different types of sensors. For example, it can refer to the registration of image of MR anatomical by functional (physiological) PET image and with the registration of these two types of images, can achieve the anatomical features and the functional features in an image [9].

The registration of medical images is divided into two categories in terms of the application: 1) The image registration of various persons. 2) The registration of 2 images from the identical person. The registration of images from the different persons is used for their physiological anatomical comparison and also in the preparation of the medical atlases. Also, the image registration of one person is more used to integrate the information of the different images or to evaluate the treatment by observing the changes in the recorded images at different times [3]. In the image registration process, the motion image $I_M(x)$ ($I_M: \Omega \to R$ which in it, Ω displays the image scope of I) is transformed for the matching on the fixed image $I_F(x)$ ($I_F: \Omega \to R$) which in it, $\mathrm{x} = (\mathrm{x}.\mathrm{y}.\mathrm{z}) \in \Omega$ is the coordinates of spatial of the pixels. Namely, in the image's registration, the purpose is the finding of the transformation $y = T(x)$ ($such\ that\ \mathrm{T}: \Omega \to \Omega$) in the way that the points of the corresponding on 2 images $I_F(x)$ and $I_M(y)$ are coincided with each other. For this purpose, different transformations can be used, and these transformations differ from each other in terms of the computational complexity and the flexibility [4].

The transformations used in the image registration are distributed into 2 general groups: the non-hard transformations and the hard transformations. The hard transformations (such as the common geometric transformations including the Affine, the rigid body, the translation, the rotation, the rescaling, etc.) can be parametrically defined, and they are the same for all image pixels. These types of transformations are usually used in the preparation of the medical atlases and in the general registration of the medical images [5], [6], [7], [9], [10], [11], [12], [13], [14]. On the other hand, the non-hard transformations (in comparison to the hard transformations) have more flexibility for image registration. In this type of transformation, a vector field is transferred for the description of the changes between the static image and the motion image so that with the use of it, the corresponding point with each pixel of the static image can be identified in the motion image. Usually, to achieve a proper registration between 2 images, it is essential to use the hard transformation and the non-hard transformation together. In this way, first, by finding the

appropriate hard transformation, the images are generally matched with each other, and then, by optimizing the appropriate non-hard transformation, the remaining local differences between them are compensated [15].

In this paper, a non-hard transformation is used for the image registration. In the current article, a 2 -stream architecture is designed, which has the ability to sequentially estimate the fields of the registration of the multi-level by a couple of the pyramids of the feature. In this paper, a 3D network of the encoder-decoder with the 2-stream is designed, which calculates 2 pyramids of the feature of the convolutional as separately by 2 volumes of the input. Also, the registration of the pyramid of the sequential is proposed, which in it, a trail of the modules of the pyramid registration (PR) for the prediction of the fields of the registration of the multi-level is designed as straight by the pyramids of the feature of the decoding. In addition, the modules of PR can be augmented with the computation of the 3D correlations of the local among the pyramids of the feature, which this work leads to the further improvement of the presented approach. Thus, it is capable of collecting the detailed anatomical structure of the brain.

This article is introduced with the aim of improving the registration of successive pyramids. This is done with advanced PR modules that increase performance. The overall contributions can be summarized as follows: (1) a two-stream 3D encoder-decoder network is designed to compute two convolution feature pyramids separately from two input volumes, and generate more robust deep features for deformation estimation. (2) Sequential pyramid registration is proposed in which a sequence of registration fields is estimated by a set of designed pyramid registration modules. The estimated registration fields perform successive sweeps on the decoding layers, which gradually refine the feature pyramids from coarse to fine. It equips the model with a strong ability to handle large deformations. (3) The PR module can be enhanced by computing local 3D correlations (between two feature pyramids) followed by multiple residual convolutions, which gather richer local details of the anatomical structure for better estimation of deformation fields. It leads to the improvement of previous methods. In addition, 3D correlations with more complex layers in the advanced PR module can enlarge the receiver field, which further increases the ability to handle large deformations. The continuation of the current article is as follows. Section II presents an overview of related works. Section III, the proposed two-stream architecture is presented. Section IV, the used datasets, the designed experiments, and the obtained outcomes are provided. Eventually, Section V presents the general conclusions and suggestions for future research.

II. Related Works

So far, the researchers have used the non-hard transformation and the hard transformation for the registration of the images of the medical on various works. In the current part, in summary, the articles in this field are provided. For example, Horm and Schunk [16], for the first time, have used the optical flux model for the non-hard registration of the images. This model is proposed by adapting the motion of the relative of objects and also the motion of the relative of the viewer. In this method, the authors suppose the brightness intensity of pixels corresponding to the static images and the motion images do not differ as significantly. In this way, by writing the Taylor series on the brightness intensity function of the motion image, the light flux relation can be obtained for the estimation and for the evolution of the transfer vector field (the non-hard transformation) [16]:

$$\nabla I.u + I_t = 0 \quad (1)$$

where, $\nabla = [\frac{\partial}{\partial x}.\frac{\partial}{\partial y}.\frac{\partial}{\partial z}]^T$ is the gradient operator. Also, $I(x.t)$ ($x = [x.y.z]^T$ displays the coordinates, and t is the time) represents a series of consecutive frames and $u(x) = [u_1(x).u_2(x).u_3(x)]^T$ ($u: \Omega \to R^3$) is the displacement vector that should be optimally adjusted. It should be noted that here $u_1(x)$, $u_2(x)$ and $u_3(x)$ represent the factors from the displacement vector on the axis of x, the axis of y, and the axis of z (and in coordinates of x). In the application of the optical flux method for the image registration, it is assumed that I_F and I_M are two consecutive temporal frames from I (namely, $I(x.0) = I_F(x)$ or $I(x.1) = I_M(x)$). In this case, the temporal derivative $I_t = \partial I/\partial t$ (at $t = 0$) after the discretization will be equivalent to the disagreement between the image of static and the image of motion (namely, $I_t(x.0) \approx I_M(x) = I_F(x)$). It is obvious that by solving the above equation under the latter conditions, the vector field u will be optimized to maximize the likeness among the corresponding cases in I_F and I_M.

The optical flux model has been used widely in the non-hard registration of the images of the medical, which, for example, can refer to the research of Palos et al. [15] After the general registration of the images based on the hard Affine transformation, they used the speed function of the grayscales difference in the optical flux model for the adaption of the details. In study [17], the authors have provided a novel registration method between the MRI images and the images of the ultrasound. To decrease the speckle noise in the procedure of registration, they have used the semi-automatic segmentation method for the magnetic resonance images and for the ultrasound images by using the active contour model. Then, they built a strong optical flux model between the segmented ultrasound images and the magnetic resonance images, and they estimated the flow field vector by using the Gaussian pyramid. Cooper and Ritter [18] have used the optical flux model for the evaluation of registration in 2D images of medical and 3D images of medical and also for the calculation of disagreements in images of static and images of motion. Also, Cao et al.[19] have used the symmetric optical flux model for the registration of 4-dimensional images of CT of the chest to deal with the image registration problems (which are caused by the variation of the local illumination intensity and the large displacements).

In the other article, namely in [20], the authors have applied the model of the elastic for the registration of the images of the brain of the human:

$$\mu\Delta u + (\lambda + \mu)\nabla(\nabla \cdot u) + f = 0 \quad (2)$$

μ and λ are the constants of the elasticity, and f is an external force. The above model is extended based on the relation of the linear elasticity of Navier for the field of the

displacement u. It describes the deformation of a material of the elastic of the deformable beneath the impression of the outside force f.

The model of the elastic has also been applied widely for the registration of medical images. For the example, it can refer to the work of Marami et al. [21]. They have estimated the prostate deformation, which is made from the crab of the prostate, by using the method based on the finite element and the model of the elastic. For this goal, the T2-weighted image of MRI from the prostate tumor before the treatment was compared with the T2-weighted image of MRI during the treatment. Mahapatra and Sun [22] have used the elastic model along with the field of the random of Markov and the integration of the information of the segmentation to improve medical image registration. In study [23], the authors also used the elastic model to align the perfusion sequence of the images of MRI of the cardiac. According to the assumption that states in the aligned time curve should have the property of sparseness, they have presented a framework for the alignment of the time-dependent physical phenomena. The introduced method had satisfactory results in the elastic deformation, and the efficiency of the approach had an important betterment in comparison to previous methods. Khallaghi et al. [24] have presented a dynamic three-dimensional registration algorithm based on the elastic model, and they have shown its performance by using the similarity method based on intensity and the volume overlap on three different clinical datasets (including the artery, the liver, and the kidney). Their feature-based algorithm had the proper time and the proper accuracy.

In another work, Christensen [25] provided the model of the fluid of the viscous for the registration images by using the relation of Nessavier-Stokes, which is according to the following relation:

$$\mu\Delta v + (\lambda + \mu)\nabla(\nabla \cdot v) + f = 0 \quad (3)$$

Where the velocity field $v(x) = [v_1(x). v_2(x). v_3(x)]^T$ $(\mathrm{V}: \Omega \rightarrow R^3)$ is described as a derivative of the temporal field of displacement u according to the below relation:

$$v = \frac{du}{dt} = \frac{\partial u}{\partial t} + v_1 \frac{\partial \mathrm{u}}{\partial \mathrm{x}} + v_2 \frac{\partial \mathrm{u}}{\partial \mathrm{y}} + v_3 \frac{\partial \mathrm{u}}{\partial \mathrm{z}} \quad (4)$$

In this method, in order to achieve the large transformations (while maintaining the image continuity), the field of the velocity is applied rather than a field of the displacement. In general, in elastic models and in viscous fluid models, the force of the outside changes the form of the motion image in the direction of the registration with the static image. This force is elected due to the usage of the image and the kind of the image, and it can be the similarity criterion gradient, the disagreement of the grayscale levels, or the interval among the curves of the corresponding 2 images. Another example, it can refer to the work of Agostino et al [26]. They have used the common information criterion gradient as a force of the outside on the model of the fluid of the viscous for registration of images of the non-homogeneous brain.

In 2009 [27], the authors provided a standard database to evaluate the registration accuracy in the non-hard transformation-based methods. They have identified a set of index points in the four-dimensional CT images from the chest of five patients, on the condition of the most aspiration and on the condition of the most exhalation, with high accuracy. Obviously, if the registration of two images from the above set is more accurate, next, the points of the index of the corresponding 2 images will be closer to each other. Next, they introduced the MLS method for the registration of the images. In this method, the index points are first matched based on the maximum correlation criterion, and then the optimal Affine transformation is obtained by using the least squares error method. The above database has been used in several pieces of research to evaluate the precision of registration of the image. For example, in another work, which is called 4DLTM, the same group estimated the movement direction of each index pixel as a polynomial function with the registration of the images of a period from the inhalation and the exhalation. For this purpose, they have used the assumption of the optical flux model that states that the brightness of the pixels is constant during the displacement [28]. In another research, with the presentation of the ALK method, they have improved the performance of the previous algorithm by using the polynomial estimation of the image brightness intensity and the Tikhonov regularizer [29]. Also, in the CCLG method, instead of the optical flux, the mass conservation equation has been used for the modeling of the movement of the voxels [30]. In the LFC method, the movement of the voxel is modeled by using the compressible flux based on the sum of the non-linear squares [31]. Due to the use of temporal information, the last three algorithms can be considered as the four-dimensional methods.

Also, the Demons method uses the static image as a local force for the movement of the voxels of the motion image in order to match the static image [32]. The methods below are all improved versions of the Demons algorithm for image registration based on their grayscale level difference: PF, EPF, AF, DF, ADF, and IC [33]. By using the frame registration method and also by using the improved version of the squared error sum criterion, Li et al. [34] have presented an algorithm called BM for the transformable registration of the images.

III. The Presented Approach

In the current part, the details of the presented approach are presented. The proposed method in this paper consists of three main components, which are as follows: (a) a network of the encoder-decoder of 2 -stream for the calculation of the pyramids of the feature, (b) the registration of the pyramid of the sequential, and (c) the bettered modules of PR.

A. Basic Topics

The purpose of the registration of 3D images of the medical is the estimation of the field of the deformation Φ that it can warp the volume of the motion $M \subset R^{H \times W \times D}$ to the volume of the static $F \subset R^{H \times W \times D}$ such which the volume of the warped $W = M \circ \Phi \subset R^{H \times W \times D}$ can be precisely equaled with constant F . $M \circ \Phi$ is used for the determination of the implementation of the field of the deformation Φ on the volume of motion by using the operation of the warping. The registration of the image is described as a problem of the optimization:

$$\hat{\Phi} = arg\ min_{\Phi}\ \zeta(F. M. \Phi) \quad (5)$$

$$\zeta(F.M.\Phi) = \zeta_{sim}(F.M \circ \Phi) + \lambda\zeta_{smooth}(\Phi) \quad (6)$$

ζ_{sim} displays the function which measures the likeness among the image of the warped $(M \circ \Phi)$ and the image of the static (F). ζ_{smooth} displays a regular limitation in the field of the deformation (Φ) that enforces spatial uniformity. ζ_{sim} and ζ_{smooth} can be described in different shapes. The last endeavors have been assigned to the development of a strong method for computation of the field of the deformation Φ.

B. The Proposed 2-Stream Model

The proposed method is based on an architecture of the encoder-decoder that is improved with the introduction of a two-stream design. The proposed architecture is displayed in Fig. 1. It can be seen from Fig. 1 that the base of the proposed method includes an encoder-decoder of 2-stream along with the collective factors. An encoder with a similar framework to U-Net [35] is used, which consists of 5 blocks of the convolutional. For the blocks, the bating for 1-th block of the convolution, every block has a layer of the 3D convolution for the down-sampling by the stride equal to 2, which this layer is coupled with an operation of ReLU. Therefore, the encoder decreases the resolution of the spatial from the volumes of the input with a factor equal to 16, as displayed in Fig. 2. In the step of the decoding, the connections of the skip are applied on the maps of the convolution of the corresponding procedure of the encoding and the procedure of the decoding. The maps of the convolutional with the low resolution are up-sampled. Then, they are joined to the maps with higher resolution. Next, a convolutional layer with a size equal to $3 \times 3 \times 3$ and an operation of ReLU is applied, as displayed in Fig. 1. Eventually, 2 pyramids of the feature by the features of the convolutional with the multi-resolution are achieved, which are calculated separately by the volume of the motion and by the volume of static.

The presented 2 -stream model lets us calculate the pyramids of the feature separately by 2 volumes of the input, and next, it allows us to predict the deformable fields from the more robust distinct learned features of the convolutional, that this point is the guidance for the bettered efficiency. The mentioned design differs from the existing networks with 1-stream, like [36]. These existing networks calculate the features of the convolutional by 2 integrated volumes, and they appraise the fields of the deformation with the use of the filters of the convolution with 1-stream. In addition, the presented 2-stream framework can calculate 2 pairwise pyramids of the feature, which in it, the fields of the layered deformation can be sequentially appraised in the various levels. The mentioned point permits the method to create a trail of the fields of the deformation with the designing of a novel approach for the registration of the sequential pyramid.

Advert, in the presented approach, the applied base in [37] is modified with the increment of the convolution blocks from 4 to 5. Also, for 5 layers, the channel number from 32 channels per layer to [8.16.16.32.32] is reduced. In addition, in the proposed method, the modified units in [37] have been removed for the maintenance of a lightweight, effective model. These changes lead to a significant reduction in the parameters of the model $(410K \rightarrow 175K)$, while the model maintains a similar performance.

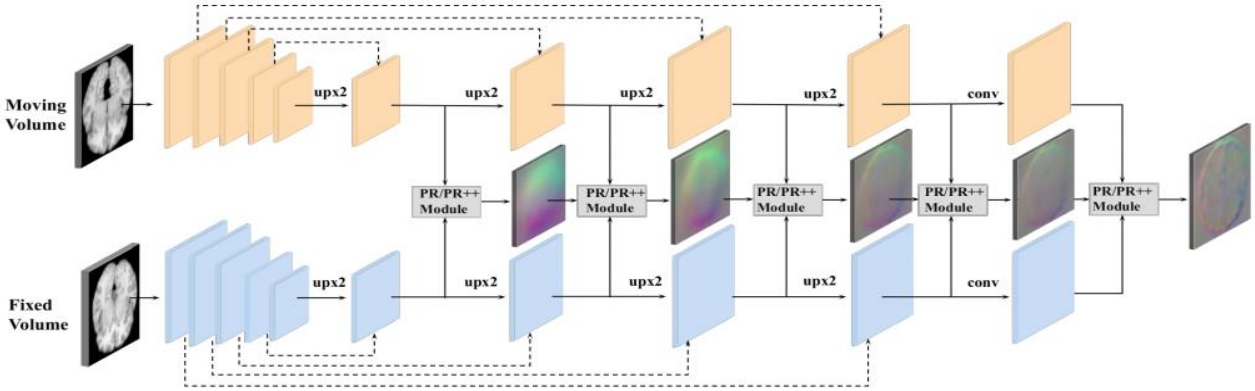


Fig. 1. The proposed method for registration of the medical image.

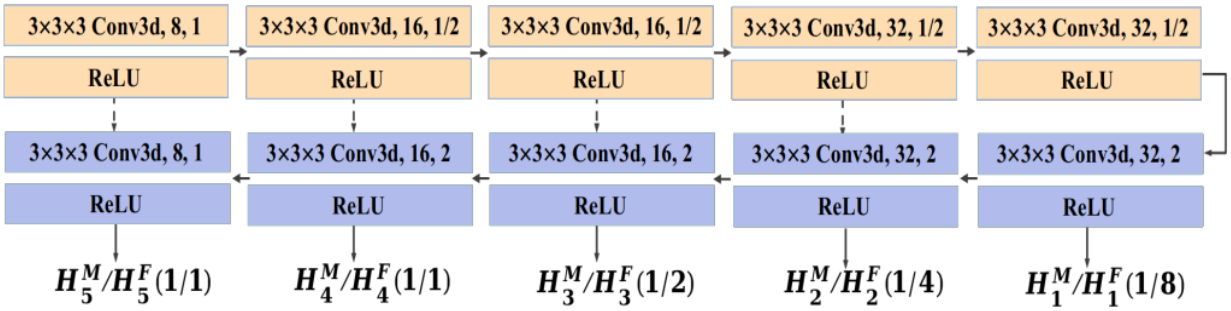


Fig. 2. The basis of our presented approach, which includes one encoder (with the color of yellow) and one decoder (with the color of blue). H_l^M and H_l^F show the computed maps of the feature by volume of motion and volume of static, respectively.

C. The Sequential Pyramid Registration

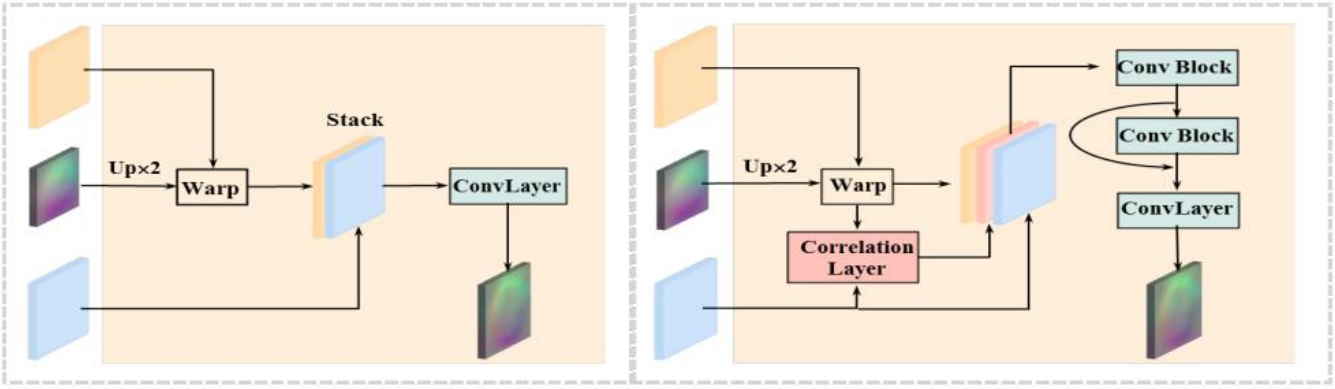
In the presented method in the current paper, a novel registration of the pyramid is provided with the designing of a collection of the modules of PR that are sequentially performed in every layer of the decoding. This work permits the method to estimate the fields of the multiscale deformation by enhancing resolution and also allows the model to create a trail from the fields of the deformation of the pyramid, as displayed in Fig. 1.

Each module of PR apprises a field of the deformation in every layer of the decoding. As the input, the module of PR applies a couple of the features of the convolutional along with a computed field of the deformation by the prior layer (the batting for the 1-th layer of the decoding in which the field of the deformation is not accessible). As the output, the module of PR obtains an appraised field of the deformation in a level of resolution, and this field is applied to the subsequent level of the pyramid. The module of PR consists of a trail from the warping operation, the stacking operation, and the convolution operation (as displayed in Fig. 3(a)). These operations are repeatedly performed on the layers of the decoding.

In the case of sequential operations, in particular, the 1-th field of the deformation (Φ_1) is calculated in 1-th layer of the decoding. First, 2 calculated features of the convolutional in 1-th layer of the decoding are superimposed, and next, one convolution of 3D by the size equal to $3 \times 3 \times 3$ is applied for the estimation of the field of the deformation. The field of the deformation (Φ_1) displays the maps in 3D with the identical form along with the corresponding maps of the feature of the convolution. It can exploit the information of the background with the coarse level, like the structure of the anatomical with top level from the brain, that this information is encoded in the computed features of the convolutional in the subsequent layer of the decoding through the warping of the feature: (1) the top field of the deformation are up-sampled with the use of the interpolation of the bilinear by the coefficient equal to 2, which it is denoted by $u(\Phi_1)$; (2) next, it is used for the warping of the maps of the convolutional volume of the motion on subsequent layers with the use of an operation of the network sampling, as displayed in Fig. 3(a). Next, the warped maps of the convolutional are superimposed anew by the corresponding features of the convolutional, which are generated from the constant volume. In the following, a convolution operation is placed for the estimation of the novel field of the deformation. The mentioned procedure is recurred in every layer of the decoding. It is described as follows:

$$\Phi_l = C_l^{3\times3\times3}(H_l^M \circ u(\Phi_{l-1}).H_l^F) \quad (7)$$

where, $l = 1.2.\cdots.N$ displays the number of the layers of the decoding. $C_l^{3\times3\times3}$ represents a three-dimensional convolution in the layer of the decoding l. The functor $\circ$ represents the operation of the warping, which this operation maps H_l^M to H_l^F by using $u(\Phi_{l-1})$. H_l^M and H_l^F are the convolutional feature pyramids, which are calculated by the volume of motion and by the volume of static in the layer of the decoding l.



(i) PR module. (ii) Improved the PR module.

Fig. 3. The presented module of (i) PR and (ii) the improved PR. The module of the improved PR betters the module of PR with the computation of the features of the correlation via the increment of the convolutions of the residual.

Regarding the 3D layer of the correlation, on the module of the improved PR, a 3D layer of the correlation is defined for calculation of the correlations of the local among 2 volumes of the input on the space of the feature of the convolution. It permits us to collect the related features that are not addressed directly in the main module of PR. However, these features can emphasize the details of the local representation of the deep. In particular, suppose P_i^W, P_j^F represent the voxels of focal of the blocks of 3D (by the size equal to $(2k+1)^3$), which are sampled by the maps of the feature of the warped motion volume and the maps of the feature of the static volume. The correlation among 2 blocks of the sampled three-dimensional can be calculated as follows:

$$C(w_i.f_i) = \frac{1}{(2k+1)^3} \sum_{n_w.n_f \in [-k.k]^3} P_{i+n_w}^W \times P_{j+n_f}^F \quad (8)$$

$n \in [-k.k]^3$ displays n iterations in a three-dimensional neighbouring equal to $[-k.k] \times [-k.k] \times [-k.k]$ from P_i^W or P_j^F. According to a block of 3D of the local in the maps of the feature of the motion volume, the computation of the correlations of the dense on total blocks of 3D sampled by the maps of the feature of the static volume is time-consuming. Thus, according to a block of 3D along by P_i^W, the correlations of the local with the sampling from a collection of P_j^F is calculated in a 3D neighborhood with a size equal to $d \times d \times d$. It can be performed as the convolutions of 3D. A stride

equal to $s_w = 1$ is applied to the sample as densely P_i^W by the maps of the features of the warped. Then, the neighboring of the correlation is adjusted by $d = 3$ in the maps of the feature of the static, which in it, P_j^F with the stride equal to $s_f = 1$ is sampled. Every block of the sampled has an identical size equal to $[-k.k] \times [-k.k] \times [-k.k]$. Also, the straight correlation among 2 blocks of the sampled is calculated with the use of the relation in Eq. (8). It produces the three-dimensional correlation maps (P^C) with the shape $[2 \times FL(d/s_f) + 1]^3 \times (H/s_w) \times (W/s_w) \times (D/s_w)$, where $[2 \times FL(d/s_f) + 1]^3 = 27$ displays the channel's number. FL represents a calculation of the floor. The created maps of the correlation have a similar form of 3D with the maps of the feature of the motion volumes and the static volumes. It ensures which 3 maps can be combined for more processing.

In the case of the convolutional enhancement, the proposed 2 -stream framework calculates 2 pyramids of the feature separated by 2 volumes of the input. Nevertheless, the funds of the work of image registration are the learning of the correspondence of the anatomical of the powerful among 2 volumes on the space of the feature, which inspired us to model a novel approach for the more aggregation of the features of the calculated pyramid. The main subordinate of the presented module of the improved PR is to present a strong method for the learning of the details of the local of the richer by 2 features, that it ensures the further precise estimation of the fields of the deformation in the different scales. To richen the features of the learned, the calculated maps of the correlation are accumulated by 2 features of the pyramid in every decoding layer: the features of the warped by the volume of the motion and the features of the pyramid from the volume of the static (as displayed on Fig. 3(b)). The maps of the correlation have 27 channels in total layers of the decoding whenever the channel number of 2 features of the pyramid changes in the various layers as below: [8.16.16.32.32].

For this purpose, as shown in Fig. 3(b), 2 blocks of the convolution of 3D are applied to process the features of the stacked. Each convolution block contains two convolution layers with a size equal to $3 \times 3 \times 3$, which in the following, an operation of ReLU is done. 1-th block of the convolution significantly decreases the stacked feature channels from [43.59.59.91.91] to [8.16.16.32.32], which this reduction is stable by the used channels number on the module of PR for the computational performance. Additionally, a connection of the residual is used on a 2-th block of the convolutional in the attempt to preserve more background information and also in the attempt to extract the distinctive features of 2 volumes. Eventually, a layer of convolution is applied for the estimation of the field of the deformation. The novel module of the improved PR is used in the proposed framework of the 2-stream registration, which this work leads to an improved version of the proposed method.

D. The Field of the Final Deformation

The presented method creates five consecutive fields of the deformation with increasing resolution, which these fields are shown as $[\Phi_1.\Phi_2.\Phi_3.\Phi_4.\Phi_5]$. To calculate the field of the final deformation, a field of the deformation of the estimated is up-sampled with a coefficient equal to 2. Next, it is warped with the field of the below deformation that this field is estimated. This up-sampling operation and this warping operation are performed as frequently and as ordinally to create the final field of the deformation (see Fig. 4) that encodes the rich information of the background of the multi-level by the deformations of the multiscale. It permits the model to propagate the robust information of the background among the layers of the decoding of the hierarchy, in which the fields of the deformation of the predicted are progressively filtered in a manner of the coarse-to-fine. Thus, it collects information on the background of the top level and the features of the low level. The information of the background of the top-level equips the proposed method to the capability for the function by the deformations of the large-scale, whenever the features of the low-scale allow it to model the detailed information of the structure of the anatomical. The modules PR and modules of the improved PR are integrated into the proposed 2-stream framework, which results in a model of the trainable end-to-end. A correlation of the cross of the local of the negative is applied as the function of the loss, and it is combined by a regularization of the smooth. For example, it can mention a diffusion regularizer, which calculates the proximate gradients of the spatial with the use of the difference among the voxels of the neighboring.

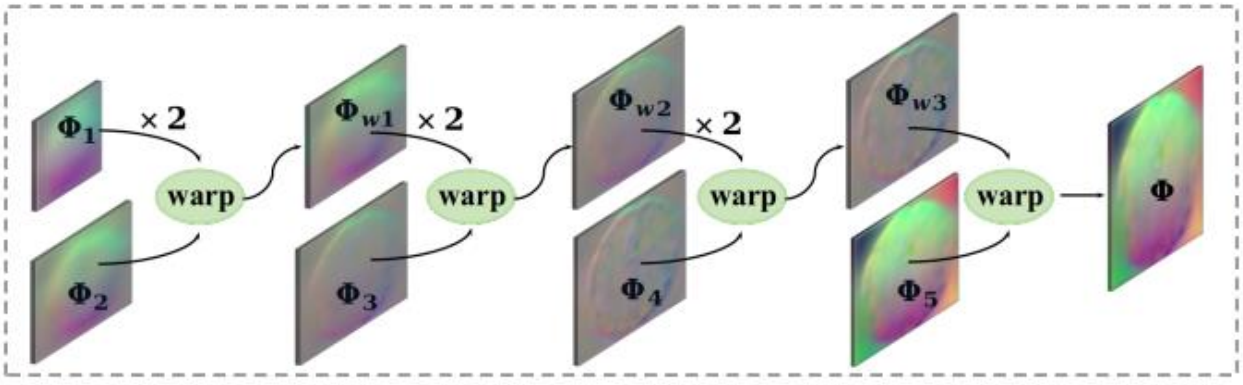


Fig. 4. The final field of the deformation is calculated from the sequential warping of the late field and the prior field.

IV. The Tests and the Analysis of the Results

In the current part, the experiments and the obtained outcomes by the proposed method of the registration of the brain MRI images are shown. In these experiments, first, a series from the atlas-based registration experiments is presented, where a field of the registration is calculated among one atlas or one volume of the reference. An atlas displays one reference or one volume of the average. Generally, it is made by frequently co-aligning the dataset from the volumes of MRI of the brain and averaging them together [40]. In this paper, a calculated atlas is used by using an external dataset [39]–[41]. Fig. 5 displays an instance from the image coupled with the use of the identical static atlas for total samples. Then, the result of the size of the dataset of the training in the image registration is briefly reviewed. Also, the results are provided in a dataset that includes the segmentations of the manual. Finally, the proposed method is trained by using a random pair of the subjects of the training as the input and the registration of the test among a pair of the subjects of the unseen. The language of the programming of Python has been used to implement these experiments. The presented method is implemented on a computer with 8G RAM and Core(TM) i7 CPU 3.0 GHz Intel(R). The network of the convolutional is performed on GPU, and the card used for the graphics on the current approach is GEFORCE 840M from NVIDIA.

A. The Used Datasets and the Evaluation Criteria

In this paper, the multi-study large-scale datasets of the scan of MRI of the brain of the T1-weighted are used from 3 publicly accessible datasets: ABIDE [42], OASIS [43], and FreeSurfer Buckner40 [39]. The details of the acquisition, the scopes of age of the subject, and the conditions of the health are various for every dataset. Total scans are down-sampled into a network with a size equal to $256 \times 256 \times 256$ and with the isotropic voxels with a length equal to $1\ mm$. The standard stages of the preprocessing (the normalization of the spatial and the extraction of the brain for every scan) were performed with the use of FreeSurfer [39], and the obtained images were cropped into $160 \times 192 \times 224$. Total MRI images were segmented as anatomically by FreeSurfer. Also, the control of the quality was applied with the use of the inspection of the visual for the detection of the errors of the gross in the results of the segmentation and in the alignment of the affine. The datasets are divided by a ratio equal to 85, 7.5, and 7.5 for validation, testing, and training, respectively. Additionally, it should be noted that the Buckner40 dataset was only used for the test step by using manual segmentation.

The obtaining of the registration of the real dense images for this data is not described as well because the multiple fields of the registration can provide analogous images of the warped. Here, first, the proposed approach is evaluated with the use of the volumetric overlay of the segmentations of the anatomical. If the field of the registration φ displays an exact match, then the areas f, $m^{\circ}\varphi$, which correspond with the identical structure of the anatomical, must be overlapped as well (see Fig. 5). The overlap of the volume among structures is quantified with the use of the score of Dice.

In addition, symmetric normalization (SyN) [42] is used and the registration algorithm with superior performance [44] in a comparative study. To implement the SyN, the software package of the generally accessible Advanced Normalization Tools (i.e., ANTs) [45] is used, which comes with a cross-correlation similarity criterion. During the experiments, a step size of SyN equal to 0.25 is used and the Gaussian parameters $(9\ and\ 0.2)$ in 3 scales by a maximum epoch equal to 201. The package of NiftyReg is used as the 2-th base. the 2D layer of the transformer of the spatial of the linear interpolation to $n - D$ is developed. Now, $n = 3$ is applied. Also, the ADAM optimizer [46] by the rate of learning equal to 0.0001 is used. The implementation includes a default with the iterations equal to 150000.

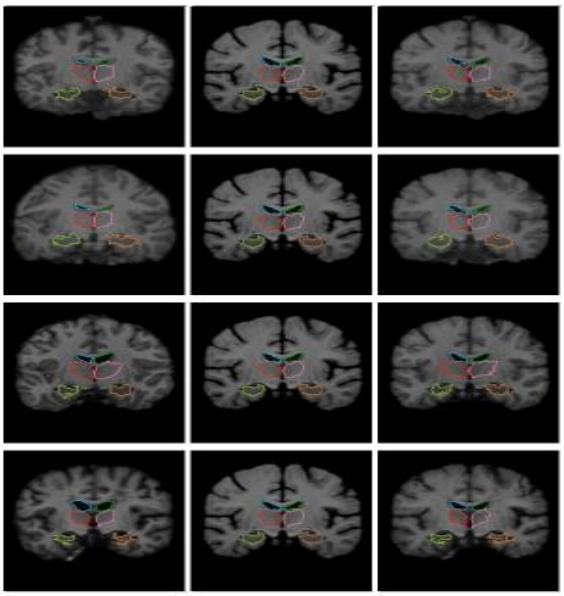


Fig. 5. The example of the extracted MRI crown crops from the couples of the input (columns $1-2$) and the obtained $M^{\circ}\varphi$ from the presented approach (third column). Our presented approach can handle the different changes in the structure form, for example, the expansion / shrinking of ventricles in row 2 and row 3 or the stretching of the hippocampus in row 4.

B. The Experiments and the Obtained Results

As mentioned, first the proposed method is trained for the atlas-based registration. First the presented network is trained. Next, the results of the Dice score on the dataset of the test are reported. Table I shows the mean of the calculated scores of Dice for total subjects in structures of the baseline method (the Affine alignment), the ANTs method, the approach of NiftyReg, and the presented approach. Due to the Dice scores, the presented approach has comparable efficiency with the ANTs method and the approach of NiftyReg, and it is significantly superior to the alignment of Affine. The examples of the outcomes of the visual images of the warped from the proposed method are displayed in Fig. 5 and Fig. 6. The proposed method is capable of performing the changes of the considerable form for the different structures. Additionally, Fig. 7 shows the scores of Dice for every method in the box form. The presented approach obtains comparable scores of Dice with the approach of ANTs and the approach of NiftyReg for all structures. It performs better in several structures, like the ventricles of the lateral, and it also performs worse in other structures, such as the hippocampus.

Table I contains the number of voxels, and in these voxels, the determinant of Jacobian is the non-positive. It is discovered which whole approaches lead to the deformations by the little islands from these voxels, yet they are different in the majority of the vast of the voxels. Fig. 6 and Fig. 8 show several examples of the deformation fields of the presented approach. The presented approach does not have any explicit restrictions for the diffeomorphic deformations. The ANTs method and the NiftyReg method include implementations that can forcefully countenance the diffeomorphic transformations. However, during the parameter searches, these cases negatively affect the time of the run or the outcomes. In the current article, the base implementations are performed by the configurations that have obtained the foremost scores of Dice. Thus, it is found that this work creates good deformation regularization.

Next, the result of the size of the dataset of the training in the precision and the relationship among the depreciated optimization and the optimization of the sample-specific is evaluated. The proposed method is trained on the subsets that are different from the dataset of the training, and then the scores of Dice are reported. The test dataset is run with the fine-tuning of the obtained displacements by the proposed method for 100 iterations per the test pair. Fig. 9 shows the obtained outcomes. The small size of the dataset of the training, equal to 10 scans, makes the scores of Dice for the training and for the test slightly lower in comparison with the bigger training dataset size. Nevertheless, there is an important disagreement on the scores of Dice when the network is trained by 100 scans or by the total dataset.

TABLE I. THE MEAN SCORES OF DICE FOR THE ALIGNMENT OF AFFINE, THE ANTS, THE NIFTYREG, AND THE PROPOSED APPROACH. THE VOXELS NUMBER AND THE PERCENT OF THE VOXELS BY A NON-POSITIVE DETERMINANT OF JACOBIAN FOR EVERY APPROACH ARE DISPLAYED

Method	Dice	$\lvert J_\varphi \rvert \leq 0$	$\%\ of\ \lvert J_\varphi \rvert \leq 0$
Affine	0.584 (0.157)	0	0
ANTs	0.749 (0.136)	9662 (6258)	0.140 (0.091)
NiftyReg	0.755 (0.143)	41251 (14336)	0.600 (0.208)
Proposed Method	0.754 (0.144)	191352 (5985)	0.374 (0.118)

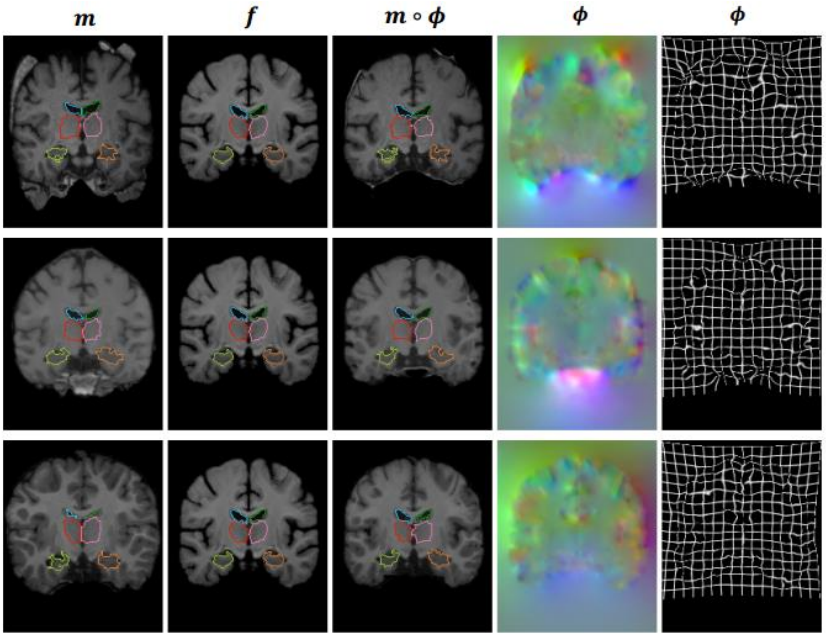


Fig. 6. An example of the fields of the deformation φ (columns $4-5$), which are exploited with the registration of a motion image (column 1) to a static image (column 2). The warped volume m°φ is displayed in column 3.

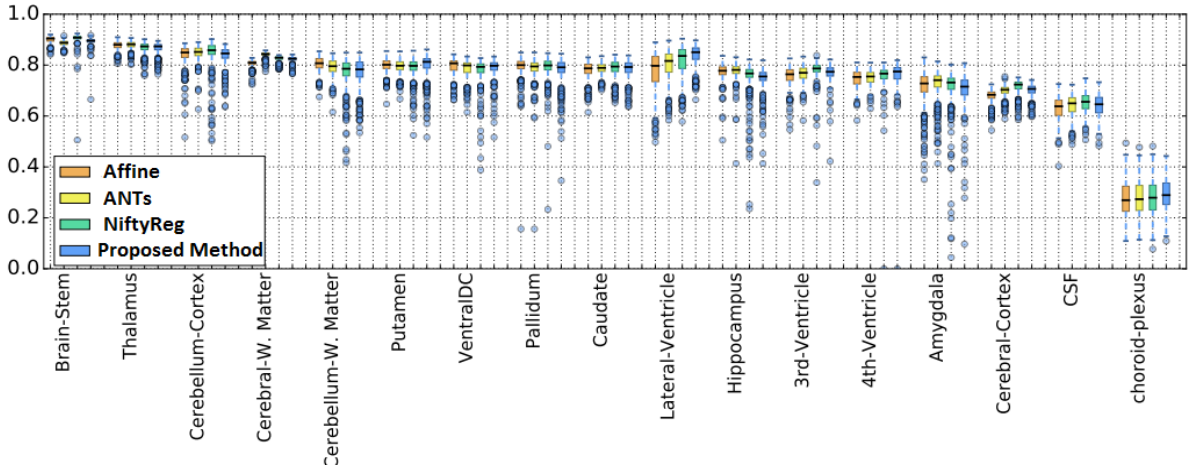


Fig. 7. The plots of the box of the scores of Dice for the various structures of the anatomical for the alignment of Affine, the ANTs, the NiftyReg, and the proposed method.

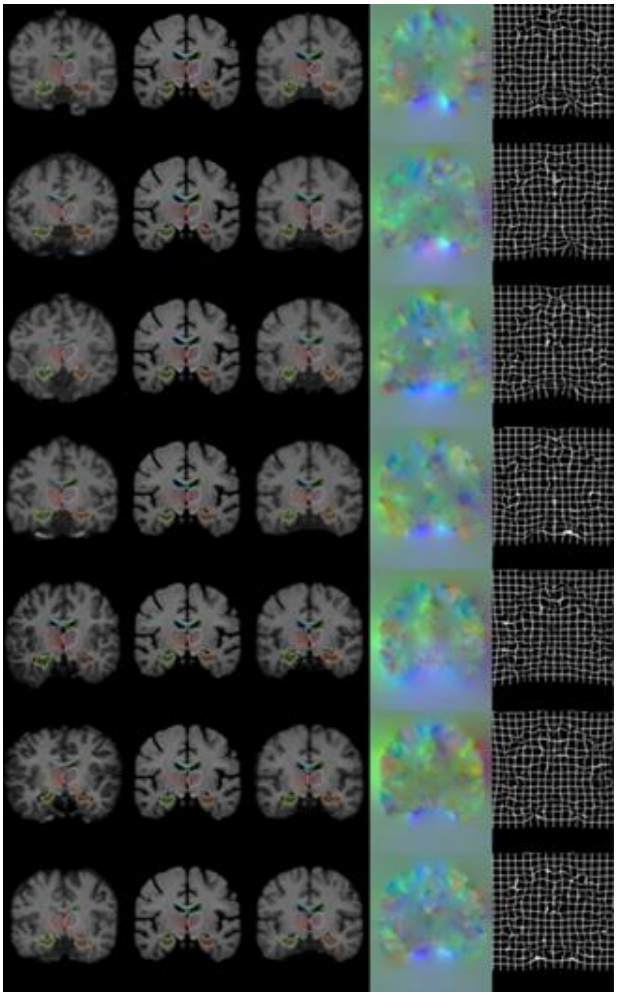


Fig. 8. An example of the fields of the deformation φ (columns $4-5$), which are exploited with the registration of a motion image (column 1) to a static image (column 2). The warped volume m°φ is displayed in column 3.

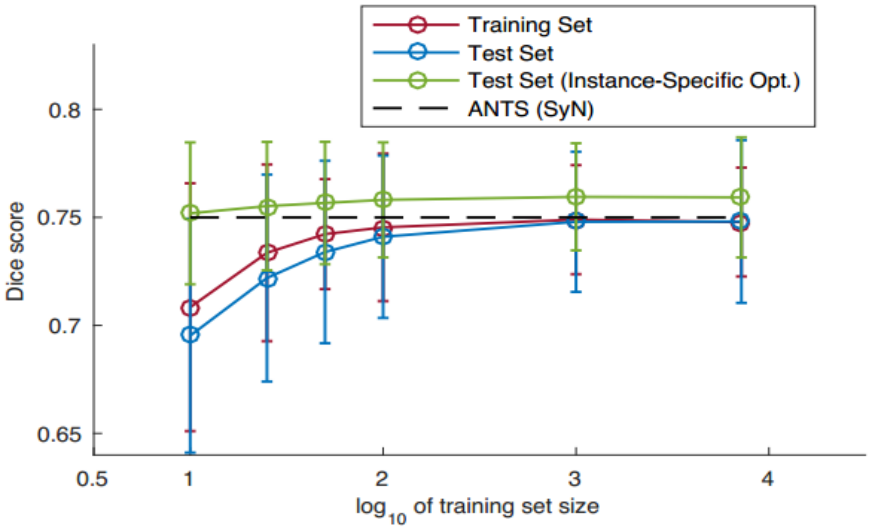


Fig. 9. Result of the size of the dataset of the training in the score of Dice.

TABLE II. THE RESULTS OF THE MANUAL ANNOTATION EXPERIMENT FOR THE AFFINE ALIGNMENT, THE ANTS, THE NIFTYREG, THE PRESENTED APPROACH, AND THE PRESENTED APPROACH BY OPTIMIZATION OF THE SAMPLE-SPECIFIC

Method	Dice
Affine	0.608 (0.175)
ANTs	0.776 (0.130)
NiftyReg	0.776 (0.132)
Proposed Method	0.773 (0.134)
Proposed Method with Instance-Specific Optimization	0.784 (0.132)

Since manual segmentation is not accessible for the uttermost datasets, the accessibility of the FreeSurfer segmentation enables a vast scope of the above experiments. In the current experiment, the registration test on the Buckner40 dataset (the us-seen) is used, which has 39 scans. The dataset of Buckner40 includes the specialist hand-lines from the identical used structures of the anatomical on the prior tests that here from them are used for the evaluation. The outcomes of the score are displayed in Table II. These outcomes display that our presented approach has similar behavior with the approach of ANTs and the approach of NiftyReg, which is stable with the 1-th test. The presented approach further improves the outcomes with the optimization of the sample-specific. In this dataset, the obtained results by the proposed method obtain slightly lower scores. However, they are bettered with optimization of the sample-specific to be comparable by the approach of ANTs and the approach of NiftyReg.

In the next experiment, the presented approach is trained for the registration of the subject-to-subject. In respect, there is a further variation on every registration. Thus, the feature number for every layer of the network is doubled. The efficiency of the presented approach is calculated by optimization of the sample-specific. Table III displays the mean scores of Dice of the test step in 250 test pairs, which are randomly selected for registration. The obtained scores of Dice from the presented approach (with the doubled number of features) are comparable with the approach of ANTs and the approach of NiftyReg. Also, the scores of Dice of the presented approach with the sample-specific optimization are comparable with both baseline methods.

C. Discussion

The proposed method in this paper, with the unsupervised loss, has the similar performance to the state-of-the-art methods ANT and NiftyReg, in terms of Dice score, while it reduces the computation time from hours to minutes on a CPU and less than a second on a GPU. The proposed method is flexible and it handles partial observed or coarsely specified auxiliary information during training, which this can lead to improved Dice score while maintaining improved runtime. Also, the proposed method performs the degenerate optimization and it learns general performance parameters that are optimal for the entire training dataset. The sample-specific optimization improves the performance of the proposed method by one dice point. This is a small increase and it indicates that the depreciated optimization can lead to near-optimal registration.

The performance gain varies based on the quality and number of anatomical sections available. Given a labeled anatomical structure during training, subjects' registration accuracy for that label increases without negatively affecting other anatomy. If half or all labels are observed, or even a coarser segmentation is provided in training, the registration accuracy for all labels will improve during testing. While with one type of auxiliary data, the experiment was done in this study, but the proposed method can use other auxiliary data such as different methods or anatomical key points.

TABLE III. THE RELATED RESULTS TO THE SUBJECT-TO-SUBJECT REGISTRATION BY USING THE AFFINE ALIGNMENT, THE ANTS, THE NIFTYREG, THE PRESENTED APPROACH, AND THE PRESENTED APPROACH BY OPTIMIZATION OF THE SAMPLE-SPECIFIC

Method	Dice
Affine	0.579 (0.173)
ANTs	0.761 (0.117)
NiftyReg	0.772 (0.117)
Proposed Method	0.763 (0.052)
Proposed Method with Instance-Specific Optimization	0.773 (0.119)

V. CONCLUSIONS AND SUGGESTIONS

In this paper, a two-stream pyramid registration network with the improved PR module is presented for unsupervised 3D medical image registration. The presented approach, due to the design, has a 2-stream architecture that permits it to calculate 2 pyramids of the feature of the convolutional as separately by 2 volumes of the input. Next, the registration of the pyramid of the sequential by a collection of the modules of PR is provided for the estimation of a trail from the fields of the registration that these fields can filter the learned features of the pyramid incrementally in a manner of the coarse-to-fine through the warping of the sequential. The module of PR is augmented with the residual complexity enhancement and with the computation of the local correlation features. The proposed approach has a competitive performance by the approach of ANTs and the approach of NiftyReg according to the score of Dice. The obtained results from the tests in 3 datasets show that the proposed approach is flexible, and it controls the partially observed information of the auxiliary during the training; this point can result in improved scores of Dice. The proposed method performs the depreciated optimization, and it learns the general performance parameters that are optimal for the full dataset of the training. Also, the outcomes have displayed which optimization of the sample-specific betters the efficiency of the proposed approach.

The proposed method can use the other data of the auxiliary, like the various methods or the anatomical key points, which can be addressed in the subsequent works. Additionally, the proposed approach is a generic mechanism of learning, and it is not restricted to a specific kind of image or anatomy. Its effectiveness can be checked on other applications of the registration of images of the medical, like the scans of MRI of the cardiac or the images of CT of the lung. Also, by providing a proper function of the loss, like the information of the mutual, the proposed approach can do the registration of the multimodal. Thus, it is suggested to be considered in the subsequent works.

ACKNOWLEDGMENT

This work was supported by the project of Research and Development of Key Technologies for SLAM Intelligent Robots Based on AI (No.212102210281) and Henan Province Higher Education Teaching Reform Research and Practice Project: "Competition Leading, Diversified Collaboration - Research and Practice on Building a System for Improving the Entrepreneurship and Employment Ability of Vocational College Graduates (No. 2021SJGLX1042)";

REFERENCES

[1] M. Saadatmand-Tarzjan and H. Ghassemian, "On analytical study of self-affine maps," Journal of Iranian Association of Electrical and Electronics Engineers, vol. 12, no. 3, pp. 77–92, 2016.

[2] M. Saadatmand-Tarzjan and H. Ghassemian, "On analytical study of self-affine maps," Journal of Iranian Association of Electrical and Electronics Engineers, vol. 12, no. 3, pp. 77–92, 2016.

[3] P. Parsa and R. Safabakhsh, "A New Method for Image Segmentation based on Multi-Objective Differential Evolution Fuzzy Clustering," Journal of Iranian Association of Electrical and Electronics Engineers, vol. 13, no. 2, pp. 103–114, 2016.

[4] D. L. G. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," Phys Med Biol, vol. 46, no. 3, p. R1, 2001.

[5] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "Elastix: a toolbox for intensity-based medical image registration," IEEE Trans Med Imaging, vol. 29, no. 1, pp. 196–205, 2009.

[6] J. Flusser and T. Suk, "A moment-based approach to registration of images with affine geometric distortion," IEEE transactions on Geoscience and remote sensing, vol. 32, no. 2, pp. 382–387, 1994.

[7] D. Li and Y. Zhang, "A novel approach for the registration of weak affine images," Pattern Recognit Lett, vol. 33, no. 12, pp. 1647–1655, 2012.

[8] P. Riyamongkol, W. Zhao, Y. Liu, L. Belayev, R. Busto, and M. D. Ginsberg, "Automated registration of laser Doppler perfusion images by an adaptive correlation approach: application to focal cerebral ischemia in the rat," J Neurosci Methods, vol. 122, no. 1, pp. 79–90, 2002.

[9] J. L. Boes et al., "Image registration for quantitative parametric response mapping of cancer treatment response," Transl Oncol, vol. 7, no. 1, pp. 101–110, 2014.

[10] P. Riyamongkol and W. Zhao, "The Hopfield neural network model for solving affine transformation parameters in the correlation method," in 2006 IEEE Region 5 Conference, IEEE, 2006, pp. 249–253.

[11] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," Int J Comput Vis, vol. 24, no. 2, pp. 137–154, 1997.

[12] Y. Chenoune, Y. Bouaoune, E. Delechelle, E. Petit, J. Garot, and A. Rahmouni, "MR/CT multimodal registration of short-axis slices in CT volumes," in 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2007, pp. 4496–4499.

[13] C. Niu, "Medical image registration based on mutual information using kriging probability density estimation," in 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, IEEE, 2006, pp. 3097–3099.

[14] J. Lin, Z. Gao, B. Xu, Y. Cao, and Z. Yingjian, "The effect of grey levels on mutual information based medical image registration," in The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2004, pp. 1747–1750.

[15] S. Samant, P. K. Nanda, and S. Sahoo, "Multimodal image registration based on weighted feature and mutual information," in 2011 IEEE Recent Advances in Intelligent Computational Systems, IEEE, 2011, pp. 420–424.

[16] G. Palos, N. Betrouni, M. Coulanges, M. Vermandel, V. Devlaminck, and J. Rousseau, "Multimodal matching by maximisation of mutual information and optical flow technique," in The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2004, pp. 1679–1682.

[17] B. K. P. Horn and B. G. Schunck, "Determining optical flow," Artif Intell, vol. 17, no. 1–3, pp. 185–203, 1981.

[18] M. Hou, C. Chen, D. Tang, S. Luo, F. Yang, and N. Gu, "Magnetic microbubble-mediated ultrasound-MRI registration based on robust optical flow model," Biomed Eng Online, vol. 14, pp. 1–15, 2015.

[19] J. Cooper, "Optical flow for validating medical image registration," in Proceedings of the 9th IASTED International Conference on Signal and Image Processing, ACTA Press/IASTED, 2003, pp. 502–506.

[20] Z. Cao, E. Dong, Q. Zheng, W. Sun, and Z. Li, "Accurate inverse-consistent symmetric optical flow for 4D CT lung registration," Biomed Signal Process Control, vol. 24, pp. 25–33, 2016.

[21] R. Bajcsy and S. Kovačič, "Multiresolution elastic matching," Comput Vis Graph Image Process, vol. 46, no. 1, pp. 1–21, 1989.

[22] B. Marami et al., "Elastic registration of prostate MR images based on estimation of deformation states," Med Image Anal, vol. 21, no. 1, pp. 87–103, 2015.

[23] D. Mahapatra and Y. Sun, "Integrating segmentation information for improved MRF-based elastic image registration," IEEE Transactions on Image Processing, vol. 21, no. 1, pp. 170–183, 2011.

[24] L. Cordero-Grande, S. Merino-Caviedes, S. Aja-Fernandez, and C. Alberola-Lopez, "Groupwise elastic registration by a new sparsity-promoting metric: Application to the alignment of cardiac magnetic resonance perfusion images," IEEE Trans Pattern Anal Mach Intell, vol. 35, no. 11, pp. 2638–2650, 2013.

[25] S. Khallaghi, C. G. M. Leung, K. Hastrudi‐Zaad, P. Foroughi, C. Nguan, and P. Abolmaesumi, "Experimental validation of an intrasubject elastic registration algorithm for dynamic‐3D ultrasound images," Med Phys, vol. 39, no. 9, pp. 5488–5497, 2012.

[26] G. E. Christensen, R. D. Rabbitt, and M. I. Miller, "Deformable templates using large deformation kinematics," IEEE transactions on image processing, vol. 5, no. 10, pp. 1435–1447, 1996.

[27] E. D'agostino, F. Maes, D. Vandermeulen, and P. Suetens, "A viscous fluid model for multimodal non-rigid image registration using mutual information," Med Image Anal, vol. 7, no. 4, pp. 565–575, 2003.

[28] R. Castillo et al., "A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets," Phys Med Biol, vol. 54, no. 7, p. 1849, 2009.

[29] E. Castillo, R. Castillo, J. Martinez, M. Shenoy, and T. Guerrero, "Four-dimensional deformable image registration using trajectory modeling," Phys Med Biol, vol. 55, no. 1, p. 305, 2009.

[30] C. B. Hoog Antink, T. Singh, P. Singla, and M. Podgorsak, "Evaluation of advanced Lukas–Kanade optical flow on thoracic 4D-CT," J Clin Monit Comput, vol. 27, pp. 433–441, 2013.

[31] E. Castillo, R. Castillo, B. White, J. Rojo, and T. Guerrero, "Least median of squares filtering of locally optimal point matches for compressible flow image registration," Phys Med Biol, vol. 57, no. 15, p. 4827, 2012.

[32] E. Castillo, R. Castillo, B. White, J. Rojo, and T. Guerrero, "Least median of squares filtering of locally optimal point matches for compressible flow image registration," Phys Med Biol, vol. 57, no. 15, p. 4827, 2012.

[33] J.-P. Thirion, "Image matching as a diffusion process: an analogy with Maxwell's demons," Med Image Anal, vol. 2, no. 3, pp. 243–260, 1998.

[34] X. Gu et al., "Implementation and evaluation of various demons deformable image registration algorithms on a GPU," Phys Med Biol, vol. 55, no. 1, p. 207, 2009.

[35] M. Li, Z. Xiang, L. Xiao, E. Castillo, R. Castillo, and T. Guerrero, "GPU-accelerated block matching algorithm for deformable registration of lung CT images," in 2015 IEEE International Conference on Progress in Informatics and Computing (PIC), IEEE, 2015, pp. 292–295.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[37] D. Kuang and T. Schmah, "Faim–a convnet method for unsupervised 3d medical image registration," in Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10, Springer, 2019, pp. 646–654.

[38] Y. Hu, E. Gibson, D. C. Barratt, M. Emberton, J. A. Noble, and T. Vercauteren, "Conditional segmentation in lieu of image registration," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, Springer, 2019, pp. 401–409.

[39] M. Kang, X. Hu, W. Huang, M. R. Scott, and M. Reyes, "Dual-stream pyramid registration network," Med Image Anal, vol. 78, p. 102379, 2022.

[40] B. Fischl, "FreeSurfer," Neuroimage, vol. 62, no. 2, pp. 774–781, 2012.

[41] M. De Craene, A. du Bois d'Aische, B. Macq, and S. K. Warfield, "Multi-subject registration for unbiased statistical atlas construction," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004: 7th International Conference, Saint-Malo, France, September 26-29, 2004. Proceedings, Part I 7, Springer, 2004, pp. 655–662.

[42] R. Sridharan et al., "Quantification and analysis of large multimodal clinical image studies: Application to stroke," in Multimodal Brain Image Analysis: Third International Workshop, MBIA 2013, Held in Conjunction with MICCAI 2013, Nagoya, Japan, September 22, 2013, Proceedings 3, Springer, 2013, pp. 18–30.

[43] A. Di Martino et al., "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," Mol Psychiatry, vol. 19, no. 6, pp. 659–667, 2014.

[44] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," J Cogn Neurosci, vol. 19, no. 9, pp. 1498–1507, 2007.

[45] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," Med Image Anal, vol. 12, no. 1, pp. 26–41, 2008.

[46] A. Klein et al., "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," Neuroimage, vol. 46, no. 3, pp. 786–802, 2009.

[47] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," Neuroimage, vol. 54, no. 3, pp. 2033–2044, 2011.

Recognition of Depression from Video Frames by using Convolutional Neural Networks

Jianwen WANG, Xiao SHA*
Department of Computer Science, Hebei University of Water Resources and Electric Engineering, Hebei 061001, China

***Abstract*—The disturbances of the mood are relevant to the emotions. Specifically, the behaviour of persons with disturbances of mood, like the depression of the unipolar, displays a powerful correlation of the temporal by the emotional girths of the arousal and the valence. Moreover, the psychiatrists and the psychologists take into account the audible signs of the facial and the audible signs of the voice when they assess the condition of the patient. Depression makes audible behaviours like weak expressions, the validation of the contact of the eye and the use of little flat-voiced sentences. Artificial intelligence has combined various automated frameworks for the detection of depression severity by using hand-crafted features. The method of deep learning has been successfully applied to detect depression. In the current article, a federate architecture, which is the network of the neural of the deep convolutional basis on the attention of global, is proposed to diagnose the depression. This method uses CNN with the attention mechanism and also uses the integration of the weighted spatial pyramid pooling for the learning of the deep global representation. In this method, two branches are introduced: the CNN based on local attention focuses on the patches of the local, while the CNN based on global attention attains the universal patterns from the whole face area. For taking the data of the supplementary among two parts, a CNN basis on the local-global attention is proposed. The designed experiments have been done in two datasets, which are AVEC2014 and AVEC2013. The results show that our presented approach can extract the depression patterns from the video frames. Also, the outcomes display that our presented approach is superior to the best methods based on the video for the detection of depression.**

***Keywords*—Deep learning; depression recognition; Convolutional Neural Network (CNN); attention mechanism**

I. INTRODUCTION

By 2020, depression had the 4-th rank among the most earnest issues of the health of the mentally [1]. Generally, it does temperate damage to the life of the individual. Also, it has a special effect on the society and the family. In several instances, depression may cause self-annihilation. Therefore, it is essential to discover an impressive solution for the diagnosis of depression and the treatment of depression of the clinical.

In recent years, a multitude of approaches have been proposed based on the different perspectives to help psychologists or doctors with the detection and treatment of clinical depression; these methods have mainly used emotional computations, machine learning communities, computer vision, etc. for estimating the depression's severity on the basis of the audiovisual cues, the common methods usually include three sequential methods: a) The extraction of the feature, b) The aggregation of the feature, and c) The regression (the classification). The extraction of the feature acts as an important task in the detection of depression from the videos. It is very important to extract a distinctive feature descriptor for the diagnosis and the estimation of depression [1]. Due to the extraction of the feature, the approaches can be hastily distributed to the features of the hand-crafted and the features of the deep learning.

The features of the hand-crafted use the knowledge domain for the designing of the features which are relevant as closely to the depressive signs [2], [3]. Although the hand-crafted features representation is considered for the obtention of the superior performance for the depression severity assessment, the below subjects have been related by the researchers. First, the exploitation of the features of the hand-crafted is time-consuming because these features require particular knowledge. The patterns of the binary of the local consist of 3 planes of the orthogonal [4], and they are heavy in terms of computational. Second, the hand-crafted features have been criticized due to the lack of related significant information on the patterns of depression [5].

Newly, the learned deep features with the use of CNNs have been applied widely to represent the deep features, and they have done great in depression diagnosis [6]. DepressNet [6] is a new framework for the learning of annotated depression representations. The selected methods of CNN (e.g., GoogleNet [7], AlexNet [8], VGG-Net [9], etc.) are pre-trained in the big datasets of the picture of the facial [10] and then it is trained in the dataset of AVEC2013 [11] and the dataset of AVEC2014 [12] with the use of fine-tuning. Its performance surpasses the many approaches to the diagnosis of depression based on the video. [13] uses the composition from the RNN [14] and the 3D-CNN [15] for the learning of the representation of the consecutive features of the spatial-temporal in 2 various scales of the areas of the face. [6], [13] adopt the model of the deep of the pre-trained for the fine-tuning of 2 datasets of the depression for the estimation of the depression.

In general, the diagnosis of depression is a problem of regression or a problem of classification according to machine learning. The purpose of the dataset of AVEC2013 and the dataset of AVEC2014 is prediction of the depression scores. It is proposed that almost all of the non-literal behaviours in the interaction of humans are the anent of the area of the face [16]. Regarding the estimation of depression based on the video, the salient area of the face is applicable for anticipating depression severity, as has been proposed in [6].

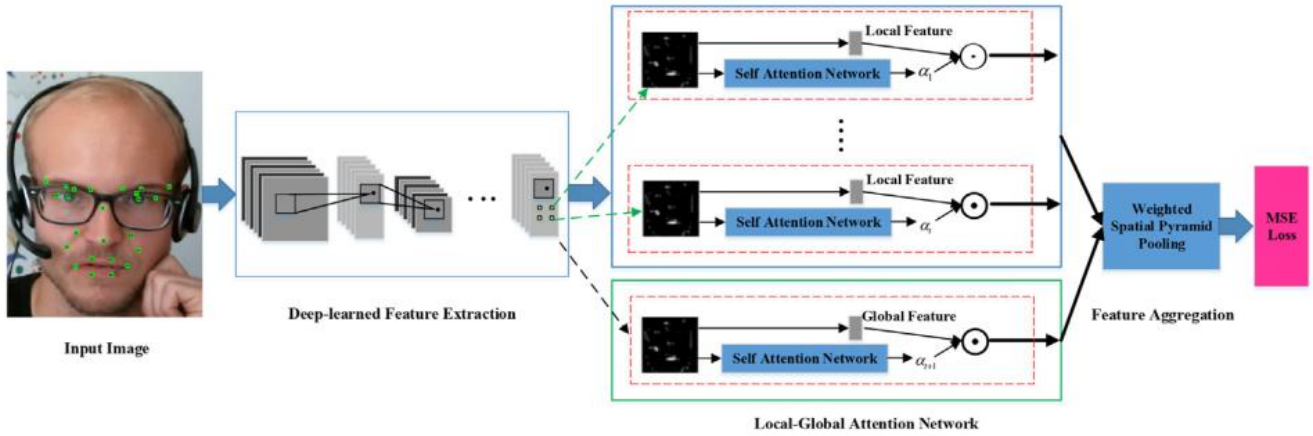


Fig. 1. Framework of our presented method to detect depression.

In the current article, the exploration of the techs on the basis of the facial look for depression diagnosis is concentrated. In order to deal with the mentioned problems, a new approach is proposed for depression detection by using the face video frames, which is called the DCNN, based on the attention of local-global. As displayed in Fig. 1, our proposed method includes 3 parts: i) The module of the extraction of the deep features learned from Depressed-CNN, ii) The convolutional neural network based on the local-global attention, iii) The module of the weighted spatial pyramid pooling (i.e., WSPP). The purpose of the universal attributes is to explain the set of depression-specific patterns, when the local attributes focus on the record of the specific patterns in the patch regions, which can extract the distinctive features in the patches of the prominent maps of the feature. The information extraction of the features of the local and the features of the global is crucial for the better performance of the depression diagnosis.

In order to clearly state the differences between the proposed method and the previous methods, we summarize the key contributions of the present paper as follows: 1) An end-to-end framework with the deep global-local attention is proposed, which effectively uses the face dynamics as a non-verbal metric to estimate the severity of the depression scale, which it has been neglected in the previous methods. 2) To encode the robust feature representations, a sophisticated CNN-based feature extraction network is designed. This network preserves the valuable and the distinctive features useful for the analysis of depression. 3) A CNN with a self-attention network effectively describes the discriminative patterns of the faces. By adopting the attention mechanism, the global-local attention-based CNN can automatically preserve the valuable feature and filter out the redundant face information. The continuation of the current article is as follows: Section II provides an overview of previous works. In Section III, our proposed approach is provided. In Section IV, the used datasets and the experiment results are provided. Section V discusses the general conclusions and future perspectives.

II. An Overview of Related Works

Many works have been done for the analysis of depression on the dataset of AVEC2013 and the dataset of AVEC2014. In the following, we introduce some presented approaches for the analysis of depression in the video frames. Liu et al. [17] have designed a region-based global network with partial attention and relational attention that this network learns the relationship between the partial features and the features of the global. In [18], the authors have introduced a framework with the use of CNN and the mechanism of attention to automatically detect depression by facial changes, whose performance overtakes most methods of facial depression detection. By focusing on the attention mechanisms and by paying attention to the facial details, this method has achieved promising results. In [19], authors have presented the deep network of the regression for the learning of the representation of the depression features visually and interpretatively, and its results show that the area near the eye plays a significant task in the diagnosis of depression. Al Jazaery and Guo [20], with the use of the 3D-CNN and the RNN, have learned as automatically the features of the spatial-temporal areas of the face in 2 various scales that can model the information of the local of the spatial-temporal and the information of the global of the spatial-temporal by the steady expressions of the facial to forecast the depression level.

On [2], the local features of the phase of the quantization by 3 planes of the orthogonal are extracted by using the coding of the sparse, and then, they are displayed with the discriminant map and by the decision surface fusion method for the generation of the features of the top-level. The regression of the vector of the support is adopted for evaluation of the severity of the depression. On [21], a new feature of the temporal dynamic, which is called the strong patterns of the binary of the average of the local by 3 planes of the orthogonal, is applied to provide the expressions of the dynamic of the temporal of the face. For the creation of a representation of the vector of the feature, the vector of Fisher of the Dirichlet process learns a richer intermediate representation from the MRLBP-TOP features in the subsequences. Next, for every sample of the video, a representation of the discriminative is generated by using the statistical aggregation approaches. In [22], the authors train an architecture of CNN for a combination of the appearance of the facial and the dynamics

of the facial to evaluate the depression diagnosis scale. The authors have related the superior outcomes over the other approaches based on the visual.

In [23], the authors design a system of artificial intelligence to estimate the depression scale. This system can combine the pattern of the supplementary among the features of the hand-crafted and the features of the deep learning. In the cues of the visual, the features of deep learning are exploited, and these features contain some related discriminative information to depression. In the cues of the audio, the features exploit the descriptors of the spectral with low level and the coefficients of the brain of the frequency of the Mel to take the expression of the vocal by the clips of the audio. The temporal movement in the space of the various features is defined with the histogram of the history of the dynamic from the feature. [6] propos Depress-Net, which is the deep network of the regression for prediction of the severity of the depression by the alone images. The map of the activation of the depression is applied in order to show the areas of the salient from the image of the face to determine the depression scale. In the meantime, the authors have designed a Depress-Net of the multi-area for modelling the various patterns of the various areas to better the total outcomes. Vast tests have been done on the dataset of AVEC2013 and the dataset of AVEC2014. Its efficiency has shown that the presented method outperforms the best visual-based methods of depression detection.

In [24], the authors extract the features of the global-local of the 3D-CNN to enhance the method's efficiency. The presented model is equipped with the 3D pooling of the average of the global for the representation of the patterns of the temporal-spatial for the diagnosis of depression. The empirical outcomes display that the integration of the features of the local and the features of the global of the 3D-CNN achieves promising efficiency. On [25], the authors have proposed to exploit automatically the basic human behaviours as the descriptors with low-dimensional from every frame. Two representations of the feature of the spectral, namely the spectral heat-maps and the vectors of the spectral, have been presented for the capturing of the associated multiscale patterns with the depression. The authors have presented these two spectral representations for the prediction of depression by using CNN and artificial neural networks. In [26], the 2-stream framework of the spatial-temporal for the depression diagnosis is presented. Eke, the researchers present the time-averaged integration method for the generation of the time-slice features. The tests on the dataset of AVEC2013 and the dataset of AVEC2014 have shown that their presented framework achieves comparable performance for depression detection.

III. Our Presented Approach

In the current part, the details of our presented approach are described. The framework of the diagnostic of depression, which is the deep and end-to-end, is shown in Fig. 1. First, the area of the face is cropped by the clips of the video with the use of the OpenFace toolbox [27]. Then, the usual CNN for the extraction of the feature is implemented, and this CNN obtains the maps of the feature of the face. In order to filter the features of the additional, the various networks of self-attention are introduced in the maps of the feature of the local and the maps of the feature of the global. WSPP is adopted to create the scale-variable features representation on the multiscale feature maps. Finally, two layers of the fully connected and a layer of the loss of the MSE are used to predict the severity of the depression. On each of the below sub-parts, the details of every part from our presented approach are provided.

A. The Extraction of Feature for the Obtention of the Face Feature Maps

The CNNs have been applied widely, and it has been proven that these networks are impressive in extracting the features in the field of emotional computations, like the recognition of the expression of the depression diagnosis and so on. To overcome the small datasets, our presented method for the analysis of depression is inspired by the described method in [28] with the un-deep frameworks. Since the deep frameworks, we are able to model a distinct presentation to predict the depression severity range. To extract the deep learning features, the presented shape of the framework of Depressed-CNN is shown in Table I. Meantime, the Depressed-CNN architecture is shown in Fig. 2.

TABLE I. Our Proposed Configuration for the Depressed-CNN Architecture

Name	Input	Operation	Kernel	Output
Conv1	224 × 224 × 3	Convolution	3 × 3, ReLu	224 × 224 × 64
Pool1	224 × 224 × 64	Pooling	2 × 2	112 × 112 × 64
Conv2	112 × 112 × 64	Convolution	3 × 3, ReLu	112 × 112 × 128
Pool2	112 × 112 × 128	Pooling	2 × 2	56 × 56 × 128
Conv3	56 × 56 × 126	Convolution	3 × 3, ReLu	56 × 56 × 256
Pool3	56 × 56 × 256	Pooling	2 × 2	28 × 28 × 256
Conv4_1	28 × 28 × 256	Convolution	3 × 3, ReLu	28 × 28 × 512
Conv4_2	28 × 28 × 512	Convolution	3 × 3, ReLu	28 × 28 × 512
CMP	28 × 28 × 512	Pooling	4 × 4	28 × 28 × 128
Con	CMP+Conv4_2	Concatenate	/	28 × 28 × 640

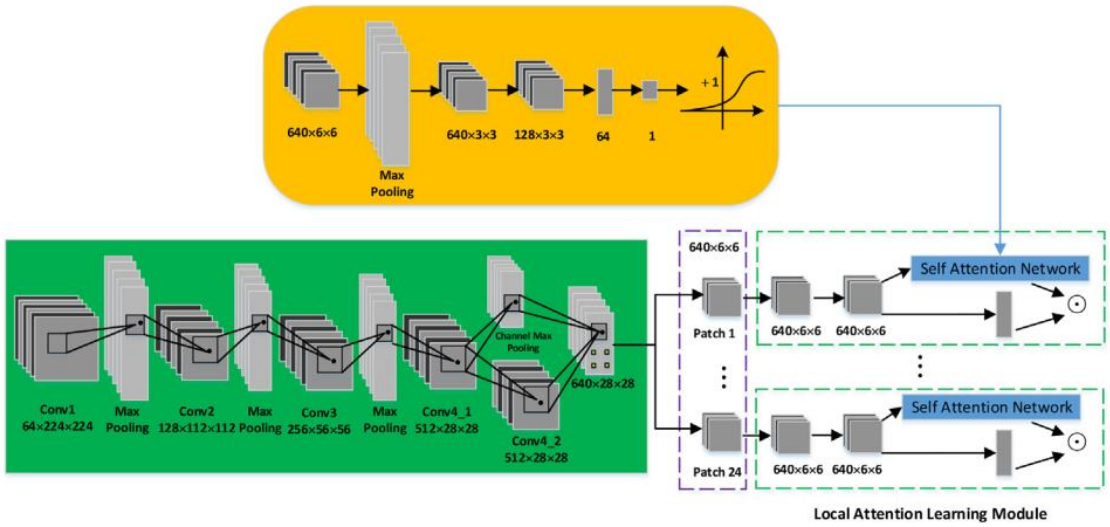


Fig. 2. Our detailed view of CNN based on the local attention.

The features of deep learning are exploited by Depressed-CNN as below. The image size input of the face is equal to 224×224 by three channels of colour. Inspired by [29], for total convolution layers, a filter by a little kernel of 3×3 is used to encode the left/right concept and the up/down concept for the extraction of the information of the spatial-temporal. After 3 operations of the convolutional and after three max-pooling operations, $Conv4\_1$ will be $[512 \times 28 \times 28]$. In the proposed method, the max-pooling is done on a window of 2×2 by $stride = 2$. To prevent the loss of information and also to preserve the distinct feature, the max-pooling of the channel (CMP) is used for the pooling of the map of the feature on the direction of the channel by the size of the kernel equal to 4, $stride = 4, pad = 0$. The output is a 3D map of the feature with a size equal to $[128 \times 28 \times 28]$. This structure is similar to this form because the popular max-pooling calculates the value of the maximum on the direction of the spatial while the max-pooling of the channel calculates the value of the maximum on the direction of the channel. In addition, for the creation of a strong representation of the feature, the map of the feature of $Conv4\_2$ is created on $Conv4\_1$. Next, CMP and the map of the feature of $Conv4\_2$ are concatenated together to achieve the feature of the final by size equal to $[640 \times 28 \times 28]$.

B. Convolutional Neural Network based on Local Attention

The various components of the image of the face have a great contribution to the diagnosis of depression. Inspired by [27], the area of the face is cropped to the various patches for the capturing of the representation of the distinct features for depression analysis. The CNN based on the local attention consists of two main steps: the generation of the patch and the capture of the salient feature. In the below, these 2 stages are described in the detail.

-The Generation of the Patch: For the analysis of the depression by using the area of the face, the patches of the local may have discriminating features to estimate the depression severity. As the popular scheme of the extraction of the feature of deep learning, the CNN restriction is the learning of the geometric transformations. Since some facial muscles contain specific information for the state recognition that is relevant as closely to the depression [30], the area of the face is cropped to the distinct patches. Also, in [31], [32], the researchers believe that the information of the multi-view from the areas of the salient is significant for image recovery. For the mentioned purpose, it is suggested that the proposed method adopts the various patches to take the distinct representations to detect depression. In the current article, the toolbox of OpenFace is used for the detection of the face region and also for the aligning of each frame from the video sequences with the size equal to 224×224 in three colour channels. Next, 68 landmarks of the face are recognized (Fig. 3). In order to find the effective patches of the face which are relevant as closely to the depression, 16 points are selected from 68 points, and these points cover the distinct face patches. Then, 8 points, which are covered on the eyes and the cheeks of the face, are recalculated. In total, 24 face patches were extracted. Our presented approach is displayed in Fig. 3.

The stages of the process of processing are as follows: (a) The OpenFace toolbox is used to discover the face zone and fix the face on every image from the trail of the video. 26 points are selected. These points cover the main areas of the face, namely, the eyes, the nose and the mouth. The elected points have the following indexes: 18, 19, 20, 37, 38, 39, 41, 42, 22, 23, 25, 26, 27, 44, 45, 46, 48, 47, 28, 30, 49, 51, 59, 53, 55, 57. (b) 4 pairs of the face landmarks (38.20), (41.42), (45.25), (48.47) are taken. Finally, 16 points from 26 points are recalculated. (c) The middle point of each pair of the points is calculated. (d) For the detection of the area around the mouth, two pairs of points (59.18) and (57.27) are selected, and then the middle points are calculated. The indices of the middle point are 22, 21. Next, 2 points, which have a similar space from the corners of the mouth, are calculated. For the coordinates of the goal edges of the direction of left, we describe it as $(U.V) = (U_{left} - 16.V_{left} - 16)$. The coordinates of the point of the target in the right direction it is described as $(U.V) = (U_{right} - 16.V_{right} - 16)$. The interval between the corner of the mouth and the edge of the target is calculated, and then the indices are determined as 24,23. (e)

Next, it selects 24 landmarks of the face, which cover the key areas of the face. (f) Due to the location of 24 points, 24 patches are cut. The patches are created by the image of the face. Nevertheless, in the proposed method, the generation of the patch is done in the maps of the feature for obtention of the areas of the salient.

The Registration of the Features of the Salient: As displayed in the box of green in Fig. 2, the process of the generation of the patch is performed with the use of the maps of the feature of CNN instead of the images of the face. This point is to maximize the utilization of the operations of the convolution and the strengthening of the fields of the receptive of the neurons, which reduces the model size. The output of the method of the extraction of the feature is the maps of the feature with a size equal to $[640 \times 28 \times 28]$. For the patch generation, 24 local zones equal to $[640 \times 6 \times 6]$ are obtained. To achieve the attributes of the salient, we apply the module of the learning of the attention of the local by CNN based on the local attention to learn the features of the patches of the face automatically. The module of learning the attention of the local is displayed on the box of yellow on the midmost from Fig. 2 by 2 dashed rectangles of green. On every module of the learning of the local attention of the patch-specific, after the generation of the patch, the created maps of the feature are entered in 2 layers of the convolution. Next, the second maps of the feature are entered into 2 branches. The 1-th branch considers the maps of the feature as the features of the local level of the vector. For the 2-th branch, a network of self-attention is applied for the focus on the distinct representation areas by the patches of the spatial. Next, the feature of the patch is recomputed by the vector of the weight.

Formally, let P_{a_j} is j-th patch of the map of the feature by the size equal to $[640 \times 6 \times 6]$. $\hat{P}_{a_j^1} = f(P_{a_j})$ is the 1-th map of the feature in the dashed rectangle of green on the top of Fig. 5, which map has a size equal to $[640 \times 6 \times 6]$. After a filter with a size equal to 1×1, the 2-th map of the feature is $\hat{P}_{a_j^2} = f(P_{a_j})$, which this map has a size equal to $[640 \times 6 \times 6]$. f represents the operation of the convolution in the architecture of CNN. Next, the 2-th map of the feature is entered into 2 branches. The 1-th branch converts the 2-th map of the feature to a vector of the local feature. Suppose φ_j is the map of the feature of the local that takes the 2-th map of the feature as the input. φ_j can be defined as the follows:

$$\varphi_j = \varphi(\hat{P}_{a_j^2}) \tag{1}$$

φ is the operation of the vector transformation. The 2-th branch is the network of self-attention. This network consists of an operation of the max-pooling, an operation of the convolution, 2 layers of the fully connected and an operation of the sigmoid. The function of the sigmoid is applied to limit the output scope α_j in 0-1. In it, 0 represents a related patch, and 1 displays a critical patch for the detection of depression. The weight α_j can be described as the follows:

$$\alpha_j = \omega_j(\hat{P}_{a_j^2}) \tag{2}$$

Where α_j is the scalar, and ω_j displays the operation of the network of the self-attention. After the operations of 2 branches, α_j is applied in the feature of the local φ_j to create a distinct feature:

$$\rho_j = \alpha_j \cdot \varphi_j \tag{3}$$

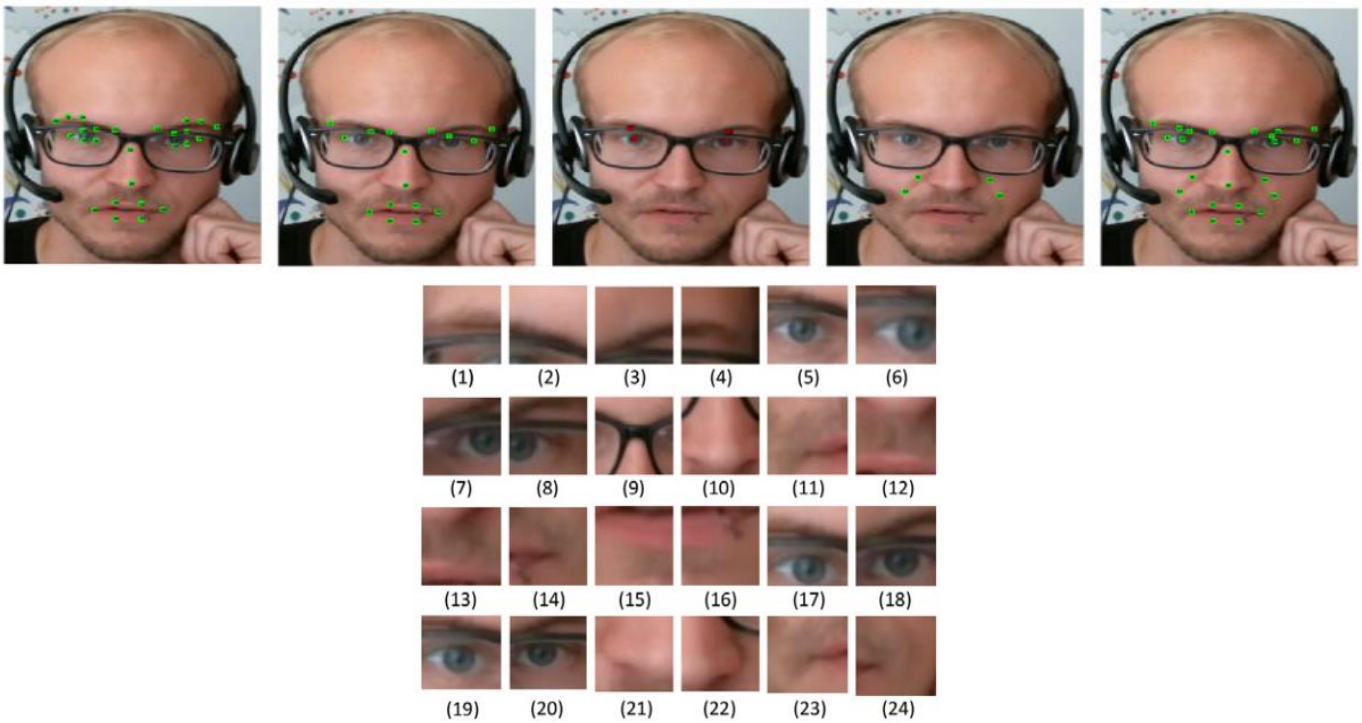


Fig. 3. An example of the face patch generation by using the mentioned patch generation process.

The output feature contains a distinct representation of the depression. Specifically, each module of local attention learning is weighted with weights, which are learned automatically with the network of self-attention. For CNN based on local attention, it is a framework of the depression of the local end-to-end with the below parts: the extraction of the deep learned features from CNN, the generation of the patch and the mechanism of the attention. The feature basis on the patch, which is learned by the proposed deep network, can discover the visual sinking pattern in the face region, and with it, the automatic estimation of the depression is feasible.

C. Convolutional Neural Network based on Global-Local Attention

In this section, we first provide a description from CNN based on global attention, and then, we state the final proposed method, namely CNN based on global-local attention. As described above, CNN, based on local attention, can automatically learn the distinctive features by using the mechanism of attention to analyze depression. Nevertheless, patches of the CNN based on the local attention may miss some additional information, which this information includes the face images and the general information of the semantics for the pattern of the depression. Therefore, for improvement of performance and also for the learning of the deep information of the semantics, the CNN based on local-global attention is proposed. For the part of the extraction of the feature, a similar operation of CNN based on local attention is used. In the presented implementation, it is proposed that the module of the learning of the attention of the global be used to represent the information of the semantics of the global in the depression diagnosis (the dashed rectangle of red in Fig. 4.

The module of the learning of global attention includes an operation of the max-pooling, an operation of the convolution and 2 operations of the branch. Ere entering to 2 branches, the map of the feature $Conv6$ can be defined as g by a size equal to $[512 \times 14 \times 14]$. The 1-th branch converts the maps of the feature g to the vectors of the feature of the global. Suppose Ψ is a feature of the global, and it is defined as the follows:

$$\Psi_{j+1} = \varphi(g) \tag{4}$$

φ is the operation of the vector transformation. The 2-th branch is the network of self-attention that includes an operation of the max-pooling, an operation of the convolution, 2 layers of the fully connected and an operation of the sigmoid. The weight α_{j+1} can be described as the follows:

$$\alpha_{j+1} = \omega_{j+1}(g) \tag{5}$$

Where α_{j+1} is the scalar, and ω_{j+1} represents the operation of the network of the self-attention. α_{j+1} is weighted in the feature of the global φ_{j+1} to obtain the feature containing the useful information ρ_{j+1}:

$$\rho_{j+1} = \alpha_{j+1} \cdot \varphi_{j+1} \tag{6}$$

In addition, to encode the complementary representations in CNN based on the local attention and in CNN based on the global attention, a final end-to-end architecture, which is called the CNN based on the local-global attention (the rectangle of red in Fig. 5) is proposed to connect this CNNs together.

D. Weighted Spatial Pyramid Pooling

The key purpose of the current article is the evaluation of depression severity. To increase the depression severity indices, the deep representation should capture the distinctive facial features at the various measures. Therefore, an important number of the samples, in total cases of the natural, are required in the training phase. The sources of the spatial of the natural changes of the face include the positions, the facial area size and the angles. In the proposed method, the features of the deeply learned local and the features of the deeply learned global by the mechanism of attention are used to obtain the information of the discriminative directly to diagnose depression. Nevertheless, the movement of the head may make changes in face size within a trail of the image. Thus, the changes may result in the blurring of the face, and this blurring affects the face image clarity.

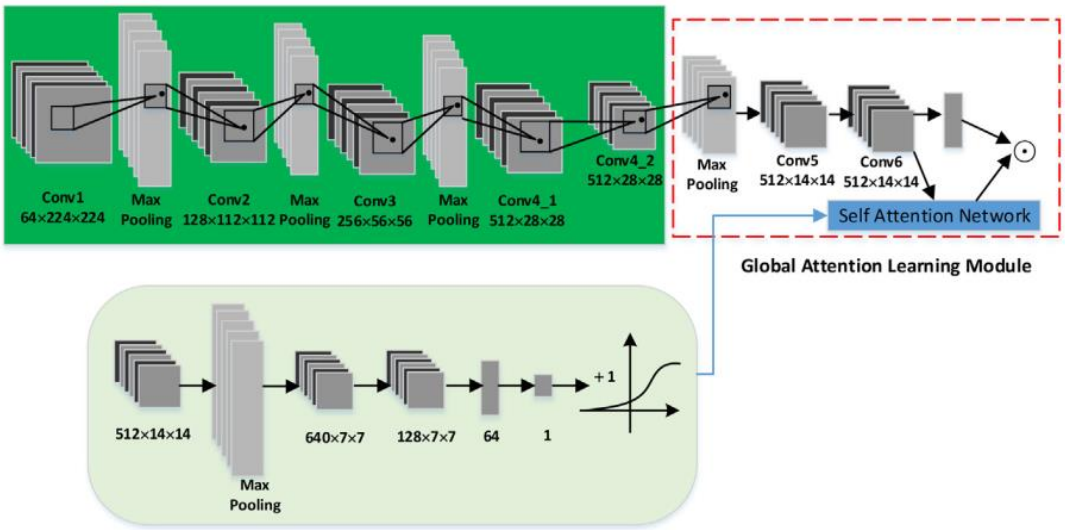


Fig. 4. Our detailed view of CNN based on the global attention.

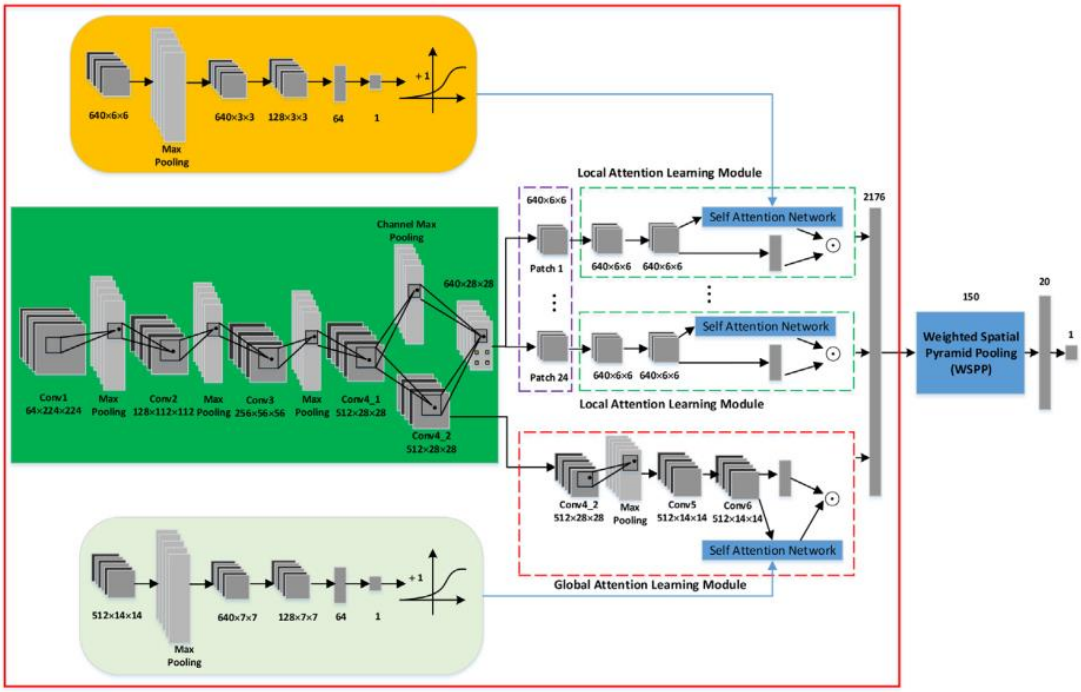


Fig. 5. Our view of the method of CNN based on local-global attention with WSPP.

To achieve a representation of the static scale, a layer of WSPP is applied for the representation of the multiple scales in the output upside of CNN based on local-global attention. The WSPP idea is to segment the map of the feature to the various parts from the scales of the finer-to-coarser, which is finished by the aggregation of the features of the local. The layer of WSPP can better the non-scaling. Also, it reduces the problem of over-fitting. In this paper, we use the definition of the initial weight on the spatial pyramid kernel, which this definition is presented in [33]. The features at the resolutions of the finer are related to the weight of the heavier ones, and the attributes of the coarser solutions are supported by the load of the lower ones. Fig. 6 describes the step of the representation of the feature of WSPP. The output shape of CNN based on the local-global attention is equal to $2048 \times 1 \times 1$.

On each spatial pyramid, the max-pooling is used for the combination of responses of every filter. The WSPP output is equal to the sum of the overall number from the pyramids. The output of CNN based on the local-global attention has the shape of $[batch\_size. 2048]$. A vector of the feature of the final by the size equal to $150D$ is obtained. The size of the window of WSPP is equal to $25 \times 151 \times 1.102 \times 1.204 \times 1$, and their corresponding stride is equal to $25 \times 151 \times 1.102 \times 1.204 \times 1$. The vectors with the fixed dimensions are fed into the layer of the fully connected.

For a network of the deep, the function of the loss plays an important task in the regression of the target. The analysis of the depression can be considered as a problem of the regression. Thus, in the proposed method, the loss of Euclidean is applied as the function of the loss, and this function is appropriate for our proposed method. The function of the loss of Euclidean L computes the squares sum of the disagreement among the values of the actual and the estimated values. It can be expressed as follows:

$$\mathrm{L} = \frac{1}{2M} \sum_{i=1}^{M} ||\hat{p}_i - p_i||^2 \qquad (7)$$

M is the sample number, and $\hat{p}_i$ displays the architecture output. Also, p_i displays the label. In this way, the final proposed architecture for the depression scale estimation is obtained, and all parts of it are shown in Fig. 5.

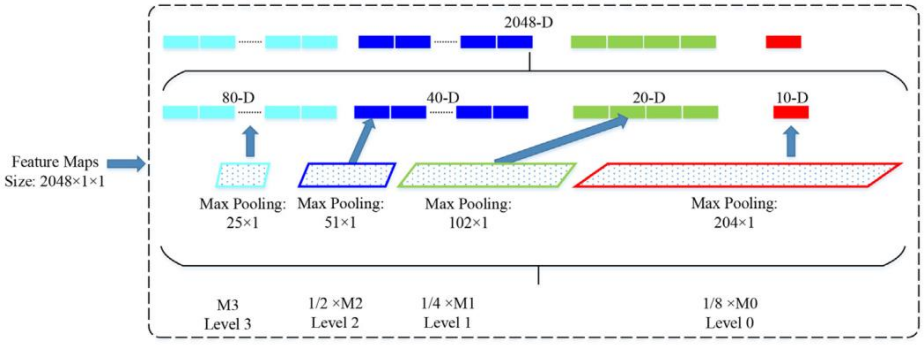


Fig. 6. Our view from WSPP.

IV. The Experiments and the Results Evaluation

In the current part, the details of our used datasets, the performed tests and also, and the obtained outcomes are presented. The Python programming language has been used for the implementation of these experiments. The presented method is implemented on a computer with 8G RAM and Core (TM) i7 CPU 3.0 GHz Intel(R). The network of the convolutional is designed on GPU, and the card used for the graphics on our approach is GEFORCE 840M from NVIDIA.

A. The Used Datasets

To prove the performance of the proposed approach for depression detection, the tests on two datasets are conducted: AVEC 2014 and AVEC 2013. The distribution of the scores of BDI-II on the dataset of AVEC 2013 and the dataset of AVEC 2014 is displayed in Fig. 7. In the dataset of AVEC 2013, 150 clips of the video exist. These clips are taken from 82 participants on the interaction of the computer-human by a webcam and a microphone for the record of the data. The scope of the age for the total people on the dataset is equal to 18 years to 63 years by a mean age equal to 31.5 years and also with a standard deviation equal to 12.3 years. The recorded clips are adjusted to 30 frames every second by a resolution equal to 640×480. The dataset of AVEC 2013 is distributed to 3 partitions: the development, the test and the training. In each partition, this dataset has 50 videos. Every video has a corresponding label with its level of depression intensity, and this level is evaluated based on the questionnaire of BDI-II.

The dataset of AVEC 2014 is the subset of the dataset of AVEC 2013. In it, there are 2 works: Northwind and FreeForm. These two works have 150 clips. In the work of FreeForm, the persons answered multiple questions, like the description of a sorrowful memory in childhood or the expression of popular food. In the work of Northwind, the persons had to study a selective from a fairy tale aurally. Similar to the AVEC 2013 dataset, the AVEC2014 dataset has 3 partitions: the development, the test and the training. The tests are done by using the partition of the training and the partition of the development from two tasks as data of training, and then, the partition of the test is applied for measurement of the model performance.

B. The Experiment Settings and the Evaluation Criteria

To obtain fast convergence and the optimization of the model, we apply the AdamW [34] by the adaptive learning rate strategy. The size of the batch is adjusted to 64, the rate of the dropout is adjusted to 0.2, and the learning coefficient is adjusted to 0.1. Regarding the rate of learning, the tests are done to check the efficiency of our presented approach at different rates.

The efficiency of base approaches in the AVEC 2013 dataset and the AVEC 2014 dataset is evaluated based on two evaluation criteria: the MAE and the RSME. MAE and RMSE are considered as criteria during the experiment for a fair comparison. These criteria are defined as follows:

$$\text{MAE} = \frac{1}{M}\sum_{j=1}^{M}|\hat{p}_j - p_j| \tag{8}$$

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(\hat{p}_j - p_j)^2} \tag{9}$$

Where M displays the overall number of samples of the video and p_j and $\hat{p}_j$ represent the actual score and the estimated score of BDI-II from j-th video.

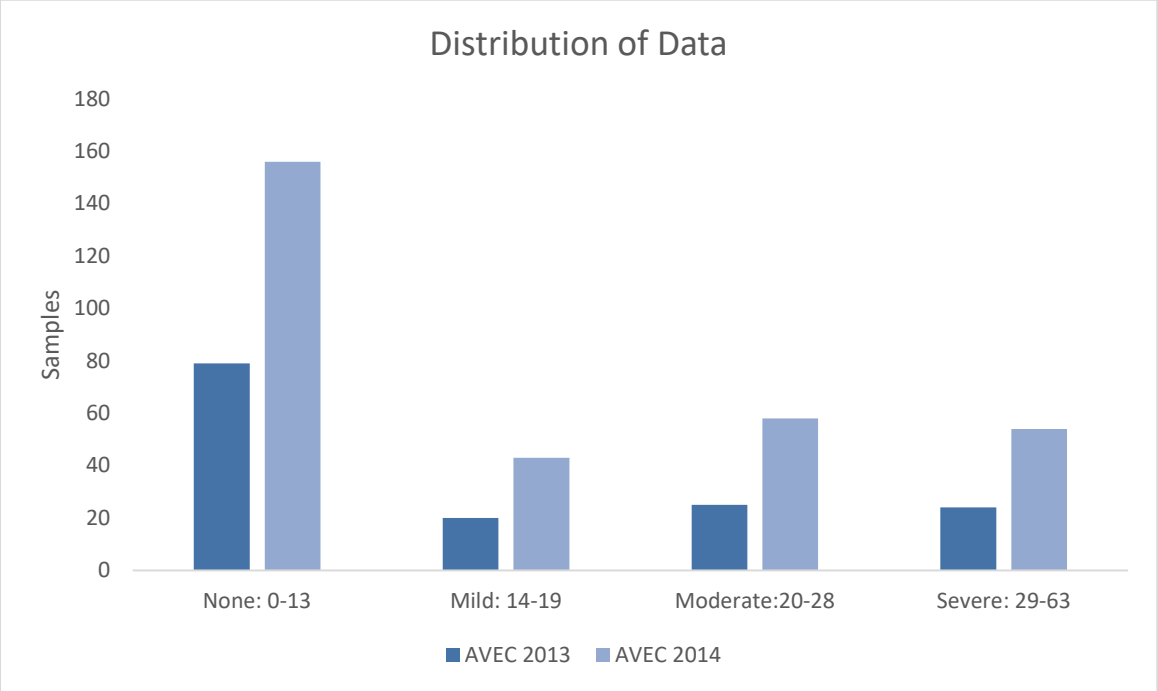


Fig. 7. Distribution of the scores of BDI-II on the dataset of AVEC 2013 and the dataset of AVEC 2014.

C. Evaluation of the Obtained Results

In the current part, first, we conduct an erosion check to investigate the performance of the components of the individual on our proposed approach. Next, our proposed framework is compared by the multiple other methods in this field to demonstrate its promising performance. The initialization of the model by LA-CNN (CMP-) and LA-CNN (CMP+), respectively, indicates the use and the non-use of the CMP technology in CNN based on local attention. The initialization of the model by GA-CNN shows that just the mechanism of the attention of the global is applied to the diagnosis of depression. The residual of the initialization of the model in Table II and Table III combine 2 or 3 separate components to take the information of the supplementary among them. Table II shows the detection results in the partition of the test from the dataset of AVEC2013. It is seen that $G1$ achieves the foremost efficiency. For the dataset of AVEC2014, we conducted various experiments for the verification of the efficiency of our presented approach. In Table III, it can be seen that similar perceptions with the AVEC2013 dataset are obtained. Additionally, from the 2 tables, it can be seen that the various approaches of CNN based on local attention have better performance than the different models of CNN based on global attention and $C1$. The performance in two depression datasets displays which capability of our presented approach is suitable for the evaluation of the depression severity scale from the video sequences. These observations show that with the combination of the components of an individual, the total efficiency is improved more over the use of a component of an individual. This point implies that we need to integrate the models of local and the models of global to diagnose depression. For a fair comparison with the other methods, just $G1$ and $G2$ are adopted to evaluate the depression diagnosis models, which are the foremost outcomes of our presented approach.

Regarding the learning rate, we changed this rate in the scope of 0.000001-0.1 and other parameters are static. As displayed in Tables IV and V, the increment of the rate of learning decreases the efficiency. Since the larger rate of learning leads to poor performance, the minimum rate of learning is appropriate for our presented approach to learn the significant patterns which are relevant to depression closely.

TABLE II. THE EFFICIENCY OF THE VARIOUS COMBINATIONS OF OUR PRESENTED METHOD IN THE AVEC2013 DATASET

Model Setting	RMSE	MAE
LA-CNN (CMP-)	8.74	7.19
LA-CNN (CMP+)	8.65	7.12
A1: LA-CNN (CMP-)+WSPP	8.56	6.81
B1: LA-CNN (CMP-)+WSPP	8.40	6.52
GA-CNN	9.05	7.43
C1: GA-CNN+WSPP	8.98	7.36
D1: LA-CNN(CMP-)+GA-CNN	8.71	7.15
E1: LA-CNN(CMP+)+GA-CNN	8.63	7.02
F1: LA-CNN(CMP-)+GA-CNN+WSPP	8.52	6.80
G1: LA-CNN(CMP+)+GA-CNN+WSPP	8.30	6.48

TABLE III. THE EFFICIENCY OF THE VARIOUS MIXTURES OF OUR PRESENTED METHOD IN THE AVEC2014 DATASET

Model Setting	RMSE	MAE
LA-CNN (CMP-)	8.72	7.16
LA-CNN (CMP+)	8.70	7.11
A2: LA-CNN (CMP-)+WSPP	8.51	6.82
B2: LA-CNN (CMP-)+WSPP	8.34	6.51
GA-CNN	9.02	7.41
C2: GA-CNN+WSPP	8.91	7.32
D2: LA-CNN(CMP-)+GA-CNN	8.62	6.91
E2: LA-CNN(CMP+)+GA-CNN	8.57	6.82
F2: LA-CNN(CMP-)+GA-CNN+WSPP	8.48	6.70
G2: LA-CNN(CMP+)+GA-CNN+WSPP	8.19	6.42

TABLE IV. THE RESULT OF THE RATE OF LEARNING IN THE EFFICIENCY OF OUR PRESENTED METHOD BY USING THE AVEC 2013 DATASET IN THE $G1$ MODE

Learning Rate	RMSE	MAE
0.000001	8.28	6.48
0.00001	8.50	6.79
0.0001	8.69	7.11
0.001	9.01	7.38
0.01	9.18	7.42
0.1	9.27	7.54

TABLE V. THE RESULT OF THE RATE OF LEARNING IN THE EFFICIENCY OF OUR PRESENTED METHOD BY USING THE AVEC 2014 DATASET IN THE $G2$ MODE

Learning Rate	RMSE	MAE
0.000001	8.18	6.40
0.00001	8.46	6.78
0.0001	8.71	7.13
0.001	8.94	7.37
0.01	9.12	7.39
0.1	9.26	7.55

In the following, to prove the efficiency of our presented approach, the comparison is made between the existing methods and our presented approach. It should be kept in mind that, as mentioned, the outcomes of our presented approach to compare the existing methods are provided on $G1$ mode and $G2$ mode. The quantitative performance comparison results for the dataset of AVEC 2013 and the dataset of AVEC 2014 are presented in Tables VI and VII. In particular, the presented models in [35]–[39] are based on the representations of the hand-crafted. Our approach performs better than the other approach in terms of the features of the hand-crafted emphasis on the experiences of the researchers, and it is hard to describe fully the depressive symptoms. In the approaches with the use of the DCNN, the presented method in [40] trains the deep methods in the big dataset, and next, it fine-tunes the dataset of AVEC 2013 and the dataset of AVEC 2014.

As shown in Tables VI and VII, our approach obtains the foremost efficiency between the approaches of the end-to-end in the used datasets. The presented method in [40] proposes a model of CNN based on the visual for depression detection by dividing roughly the area of the face into 3 parts and, next, by combining the total image of the face to better the model's detection performance. Our superior efficiency is according to the combination of the mechanism of the attention of the local and the mechanism of the attention of the global to extract the depression features. The results of the presented method in [40] display that their approach relies on the attention to just an area and also relinquishes other details of the face, which helps in depression detection. Reciprocally, the presented method in [41] obtains the acceptable efficiency sans a pre-trained network. The researchers segment the face area based on the points of the facial feature, and then, they clog the map of the feature to extract the information of the feature of the local. Our presented approach performs superior over these approaches with a significant margin.

D. Discussion

The obtained results from the experiments show that the proposed hand-crafted features as well as the feature aggregation method do not obtain higher RMSEs in compared to the method proposed in the present paper. From these results, it can be concluded that the proposed method in AVEC2013 and AVEC2014 can automatically learn the local and global feature information of the face area and outperform the best advanced methods for depression diagnosis. This further shows the effectiveness of the proposed method for diagnosis and analysis of depression.

In compared to the obtained results by similar methods, our method has improved the accuracy of depression diagnosis. There is a potential reason that the two-stage framework (eg, pre-training, fine-tuning) can effectively use their advantage to detect depression. By comparing the obtained results from similar methods, the RMSEs do not exceed them, our method is trained from scratch to be an end-to-end design for depression detection. For AVEC2014, as shown, our method achieves the comparable results to most video-based depression detection methods on the test set. In addition, these results of different components are better than the results obtained in AVEC2013.

Finally, in order to have an intuitive view of the prediction of the depression score from the images of the face, the images of the visualized face are provided in Fig. 8. The 1-th column from Fig. 8 displays the basic images. 2-th column to 5-th columns show the different areas from the facial image. The heat map on the images of the face is the fireplace area that the approach has learned. Before the merging of the attention maps, the proposed method can refer to multiple locations simultaneously. The proposed approach, in particular, relies on the areas of the movement of the associated facial muscles with depression, like the eyes, the mouth and the eyebrows. However, it ignores the irrelevant areas.

TABLE VI. COMPARISON OF OUR PRESENTED APPROACH TO SIMILAR APPROACHES BASED ON THE DATASET OF AVEC 2013

Methods	MAE	RSME
LPQ in [11]	10.88	13.61
PHOG in [42]]	-	10.45
LPQ-TOP in [2]	8.22	10.27
MRLBP-TOP, DPFV in [21]	7.55	9.20
LSOGCP in [39]	6.91	9.17
2D-CNN in [40]	6.50	8.41
3D-CNN in [41]	6.83	8.46
Proposed Method (G1)	6.48	8.30

TABLE VII. COMPARISON OF OUR PRESENTED APPROACH TO SIMILAR APPROACHES BASED ON THE DATASET OF AVEC 2014

Methods	MAE	RSME
[36]LGBP-TOP in	8.86	10.86
LBP-TOP in [43]	7.08	8.91
MRLBP-TOP, DPFV in [21]	7.21	9.01
LSOGCP in [39]	7.19	9.10
ResNet-50 in [44]	7.13	8.23
2D-CNN in [40]	6.51	8.39
3D-CNN in [41]	6.78	8.42
Proposed Method (G2)	6.42	8.19

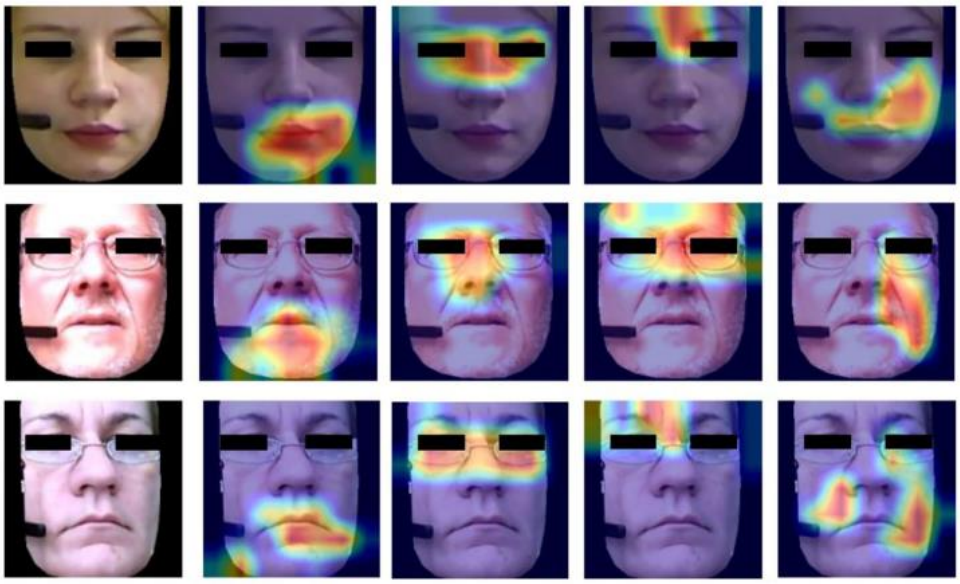


Fig. 8. The examples of the visualization of the face images with the different areas from the face.

V. CONCLUSIONS AND SUGGESTIONS

In this research, CNN, by the mechanism of attention, is used for the designing of an end-to-end integrated approach to the diagnosis of depression based on the video. We ratiocinate which a functional ability to take the feature pattern from the "encoded" depression on the areas of the face is important. In particular, a new framework is proposed. This framework consists of two branches: CNN based on local attention and CNN based on global attention. A CNN based on the local attention focuses only on the local patches. A CNN based on global attention learns the patterns of the global from the total area of the face. To take the information of the supplementary among 2 branches, a CNN basis on the local-global attention is presented. Finally, to achieve the informative patterns of the depression, a WSPP is applied to learn the final feature representations. The extensive tests in 2 datasets, namely AVEC2014 and AVEC2013, have displayed that the ability of our presented approach is higher than the diagnosis models of depression that are almost video-based.

Hereafter, the dataset from the further depressed patients will be gathered for the learning of the stronger features representation by the various appearance images. Additionally, the examination of learning of the multimodal depression representation (the audio, the video, the text, etc.) seems to be an interesting topic. In addition, we will investigate the more explainable patterns of the representation and the stronger approaches of the regression by the discriminant DCNN. Also, the presented model based on deep learning can aid doctors in the evaluation of depressed people.

ACKNOWLEDGMENT

This work was supported by the Hebei University of Water Resources and Electric Engineering Basic Research Project Funding: "Research on Social Recommendation Algorithms Based on Disentangled Graph Neural Networks" (No. SYKY2311).

REFERENCES

[1] S. Song, S. Jaiswal, L. Shen, M. Valstar, Spectral representation of behaviour primitives for depression analysis, IEEE Transactions on Affective Computing (2020), 1-1.

[2] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," IEEE Transactions on Information Forensics and Security, vol. 10, no. 7, pp. 1432–1441, 2015.

[3] L. He, D. Jiang, and H. Sahli, "Multimodal depression recognition with dynamic visual and audio cues," in 2015 International conference on affective computing and intelligent interaction (ACII), IEEE, 2015, pp. 260–266.

[4] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in 2015 International conference on affective computing and intelligent interaction (ACII), IEEE, 2015, pp. 255–259.

[5] S. Song, L. Shen, and M. Valstar, "Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral

features," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 158–165.

[6] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," IEEE Trans Affect Comput, vol. 11, no. 3, pp. 542–552, 2018.

[7] C. Szegedy et al., "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[8] C. Yan, B. Gong, Y. Wei, Y. Gao, Deep multi-view enhancement hashing for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020), 1–1.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[10] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," arXiv preprint arXiv:1411.7923, 2014.

[11] C. Yan, B. Shao, H. Zhao, R. Ning, Y. Zhang, F. Xu, 3d room layout estimation from a single rgb image, IEEE Transactions on Multimedia (2020), 1–1.

[12] M. Valstar et al., "Avec 2014: 3d dimensional affect and depression recognition challenge," in Proceedings of the 4th international workshop on audio/visual emotion challenge, 2014, pp. 3–10.

[13] M. Al Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," IEEE Trans Affect Comput, vol. 12, no. 1, pp. 262–268, 2018.

[14] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in Proceedings of the 18th ACM international conference on multimodal interaction, 2016, pp. 445–450.

[15] Guo, W., Yang, H., Liu, Z., Xu, Y., and Hu, B. (2021). Deep neural networks for depression recognition based on 2d and 3d facial expressions under emotional stimulus tasks. Front. Neurosci. 15:609760. doi: 10.3389/fnins.2021.609760.

[16] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," Image Vis Comput, vol. 32, no. 10, pp. 641–647, 2014.

[17] Z. Liu, X. Yuan, Y. Li, Z. Shangguan, L. Zhou, and B. Hu, "PRA-Net: Part-and-Relation Attention Network for depression recognition from facial expression," Comput Biol Med, vol. 157, p. 106589, 2023.

[18] M. Niu, L. He, Y. Li, and B. Liu, "Depressioner: Facial dynamic representation for automatic depression level prediction," Expert Syst Appl, vol. 204, p. 117512, 2022.

[19] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," IEEE Trans Affect Comput, vol. 11, no. 3, pp. 542–552, 2018.

[20] M. Al Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," IEEE Trans Affect Comput, vol. 12, no. 1, pp. 262–268, 2018.

[21] L. He, D. Jiang, and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding," IEEE Trans Multimedia, vol. 21, no. 6, pp. 1476–1486, 2018.

[22] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," IEEE Trans Affect Comput, vol. 9, no. 4, pp. 578–584, 2017.

[23] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," IEEE Trans Cogn Dev Syst, vol. 10, no. 3, pp. 668–680, 2017.

[24] W. C. de Melo, E. Granger, and A. Hadid, "Combining global and local convolutional 3d networks for detecting depression from facial expressions," in 2019 14th ieee international conference on automatic face & gesture recognition (fg 2019), IEEE, 2019, pp. 1–8.

[25] S. Song, S. Jaiswal, L. Shen, and M. Valstar, "Spectral representation of behaviour primitives for depression analysis," IEEE Trans Affect Comput, vol. 13, no. 2, pp. 829–844, 2020.

[26] M. A. Uddin, J. B. Joolee, and Y.-K. Lee, "Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm," IEEE Trans Affect Comput, vol. 13, no. 2, pp. 864–870, 2020.

[27] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in 2016 IEEE winter conference on applications of computer vision (WACV), IEEE, 2016, pp. 1–10.

[28] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2983–2991.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[30] He, L., Tiwari, P., Lv, C., Wu, W., and Guo, L. (2022b). Reducing noisy annotations for depression estimation from facial images. Neural Netw. 153, 120–129. doi: 10.1016/j.neunet.2022.05.025.

[31] C. Yan, B. Gong, Y. Wei, and Y. Gao, "Deep multi-view enhancement hashing for image retrieval," IEEE Trans Pattern Anal Mach Intell, vol. 43, no. 4, pp. 1445–1451, 2020.

[32] C. Yan, B. Shao, H. Zhao, R. Ning, Y. Zhang, and F. Xu, "3D room layout estimation from a single RGB image," IEEE Trans Multimedia, vol. 22, no. 11, pp. 3014–3024, 2020.

[33] J.B.J. Md Azher Uddin, Y.-K. Lee, Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm, IEEE Transactions on Affective Computing (2020), 1–1.

[34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," arXiv preprint arXiv:1711.05101, 2017.

[35] He, L., Guo, C., Tiwari, P., Su, R., Pandey, H. M., and Dang, W. (2022a). Depnet: an automated industrial intelligent system using deep learning for video-based depression analysis. Int. J. Intell. Syst. 37, 3815–3835. doi: 10.1002/int.22704.

[36] M. Valstar et al., "Avec 2014: 3d dimensional affect and depression recognition challenge," in Proceedings of the 4th international workshop on audio/visual emotion challenge, 2014, pp. 3–10.

[37] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," IEEE Transactions on Information Forensics and Security, vol. 10, no. 7, pp. 1432–1441, 2015.

[38] L. He, D. Jiang, and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding," IEEE Trans Multimedia, vol. 21, no. 6, pp. 1476–1486, 2018.

[39] M. Niu, J. Tao, and B. Liu, "Local second-order gradient cross pattern for automatic depression detection," in 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), IEEE, 2019, pp. 128–132.

[40] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," IEEE Trans Affect Comput, vol. 11, no. 3, pp. 542–552, 2018.

[41] L. He, C. Guo, P. Tiwari, H. M. Pandey, and W. Dang, "Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence," International Journal of Intelligent Systems, vol. 37, no. 12, pp. 10140–10156, 2022.

[42] Valstar, M., Schuller, B., Smith, K., Almaev, T. R., Eyben, F., Krajewski, J., et al. (2014). "AVEC 2014: 3D dimensional affect and depression recognition challenge," in Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (Orlando, FL), 3–10. doi: 10.1145/2661806.2661807.

[43] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in 2015 International conference on affective computing and intelligent interaction (ACII), IEEE, 2015, pp. 255–259.

[44] W. C. De Melo, E. Granger, and A. Hadid, "Depression detection based on deep distribution learning," in 2019 IEEE international conference on image processing (ICIP), IEEE, 2019, pp. 4544–4548.

MG-CS: Micro-Genetic and Cuckoo Search Algorithms for Load-Balancing and Power Minimization in Cloud Computing

Jun ZHOU*, Youyou Li
College of Artificial Intelligence, Jiaozuo University, Jiaozuo, Henan, 454000, China

***Abstract*—Cloud computing has emerged as a transformative technology, offering remote access to various computing resources. However, efficiently managing these resources while curbing escalating energy consumption remains a critical challenge. In response, this paper presents the Micro-Genetic Algorithm with Cuckoo Search (MG-CS), a novel approach for enhancing cloud computing efficiency. MG-CS optimizes load balancing and power reduction and significantly contributes to reducing operational costs, ensuring compliance with service level agreements, and enhancing overall service quality. Our experiments showcase MG-CS's versatility in achieving a well-balanced distribution of workloads, resource optimization, and substantial energy savings. This multifaceted approach redefines cloud resource management, offering an environmentally sustainable and cost-effective solution. By introducing MG-CS, this research addresses the pressing challenges in cloud computing, aligning it with environmental responsibility and economic efficiency.**

***Keywords*—Resource utilization; cloud computing; energy consumption; optimization**

I. INTRODUCTION

Cloud computing enables cloud users to access a wide range of configurable computing resources, such as networks, servers, storage, services, and applications, conveniently and on-demand [1]. It has become a transformative technology widely discussed and currently prevalent in numerous commercial sectors. The cloud environment is categorized into private, public, and hybrid/federated clouds [2]. A private cloud represents a dedicated computing environment exclusively utilized by a single organization. It offers benefits like isolation, customization, and heightened security. The hosting can either be on-premises or managed by a third-party provider [3].

On the other hand, a public cloud operates as a shared cloud computing environment accessible to the general public. It provides advantages such as convenience, cost-effectiveness, and scalability, with resources delivered by third-party service providers via the Internet [4]. A hybrid/federated cloud integrates elements of both private and public clouds. This approach enables organizations to distribute workloads across multiple cloud deployment models, offering flexibility, seamless integration, and redundancy. Multi-provider clouds are becoming increasingly popular in cloud infrastructure, where multiple providers are used to distribute workloads across the environment. Organizations can enhance flexibility, redundancy, and resource allocation by leveraging multiple providers. Moreover, there are specialized cloud environments designed to cater to specific services [5]. IoT cloud services are a prime example that caters to IoT devices' data analysis and management. These services are equipped with capabilities to process and derive insights from the massive volumes of IoT-generated data efficiently. Mobile cloud services employ cloud computing to provide applications and services to mobile devices. This approach allows mobile users to access cloud applications and data, providing flexibility, scalability, and improved performance [6].

Cloud computing encompasses three primary cloud service models, each catering to specific needs: Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and Infrastructure-as-a-Service (IaaS) [7]. Software applications are provided to users over the Internet by SaaS, a cloud computing model that uses a subscription-based approach. SaaS enables users to access and utilize these applications via a web browser, eliminating the need for local installation and maintenance. The responsibility of hosting, maintaining, and updating the software lies with the SaaS provider. PaaS, however, provides developers with a platform and environment to build, deploy, and manage applications without the complexities of managing the underlying Infrastructure. PaaS includes essential tools, runtime environments, databases, and other services required for seamless application development and deployment. Finally, IaaS grants users' access to virtualized computing resources via the Internet. Users can rent virtual machines, storage, and networking components pay-as-you-go. IaaS empowers organizations with the flexibility to create and manage their virtual data centers without the burden of owning physical hardware.

Cloud computing is built upon service-oriented architecture, which enables it to offer various services such as Database-as-a-Service (DbaaS), Identity-as-a-Service (IDaaS), and the broader concept of Anything-as-a-Service (XaaS). This architecture has revolutionized resource management in industry and academia, providing an efficient and dynamic approach [8]. The cloud system's dynamic nature is a crucial characteristic, accommodating numerous users, devices, networks, organizations, and resources that frequently connect and disconnect from the system. This adaptability is essential for meeting the diverse needs of cloud users. Several factors come into play when deciding on the appropriate cloud service model to implement. These factors include flexibility, scalability, interoperability, and service control. Evaluating

these aspects is crucial in determining the best-fit cloud service model to address specific requirements and optimize performance.

In the context of cloud computing, users have the flexibility to request resources from both the cloud service provider and the cloud resource broker. When functioning as a cloud service provider, the cloud resource broker is responsible for selecting the most suitable resource, considering the user's stipulated time constraints and budget considerations. This dynamic approach ensures the seamless delivery of on-demand services to users. Nevertheless, the proliferation of users and applications within the cloud ecosystem can lead to an escalation in workload and web application traffic, particularly for those deployed on virtual machines (cloud resources). To manage this expanding landscape effectively, the cloud resource broker necessitates a proficient algorithm capable of distributing tasks equitably among the active virtual machines. Such an algorithm becomes instrumental in minimizing the proportion of tasks that are rejected due to resource constraints. The overarching goal of load distribution within the cloud milieu is to optimize several critical aspects, including scalability, response time, and resource utilization.

Effective load-balancing not only leads to the attainment of minimum makespan times for tasks but also contributes to overall system performance enhancement. Furthermore, load-balancing acts as a preventive measure against system bottlenecks stemming from disparities in load distribution. This realm presents substantial research challenges within the realm of cloud computing, focusing on the equitable distribution of workload among virtual machines. Load-balancing in the cloud encompasses two pivotal stages: task scheduling and virtual machine monitoring. Task scheduling, a well-recognized optimization problem (NP-Complete), becomes intricate due to the heterogeneous resource configuration within the cloud and the swift fluctuations in on-demand requests. The intricate nature of this landscape renders the prediction and computation of all conceivable task-resource mappings within the cloud environment arduous.

Consequently, the development of an efficient task-scheduling algorithm assumes paramount importance. Such an algorithm is instrumental in the judicious distribution of tasks, thereby mitigating scenarios where certain virtual machines endure overload or under-load conditions. These algorithms play a pivotal role in fostering balanced resource utilization and fostering optimal performance within cloud computing systems. As a result, they constitute an indispensable component in the pursuit of achieving equilibrium and excellence within the dynamic cloud computing landscape.

Various techniques, including meta-heuristic algorithms, machine learning, and deep learning, have been integrated into cloud load balancing strategies to address the increasing demand for cloud services and ensure optimal resource utilization. Meta-heuristic algorithms, such as Ant Colony Optimization (ACO) [9], Particle Swarm Optimization (PSO) [10], sine cosine algorithm [11], and imperialist competitive algorithm [12], provide efficient methods for task scheduling and resource allocation, contributing to equitable workload distribution and enhanced system performance. Machine learning techniques enable cloud systems to learn from historical data, adapt to changing workloads, and make real-time load-balancing decisions [13-15]. Deep learning, with its neural networks, enhances predictive accuracy and aids in proactive load management [16, 17]. Cloud load balancing ensures that the various components of these transportation systems, such as ticketing, scheduling, and real-time tracking, operate efficiently and respond to dynamic demands [18, 19]. By utilizing the power of cloud computing and the aforementioned advanced techniques, public transportation services can offer improved reliability, scalability, and cost-effectiveness, ultimately benefiting commuters and the environment.

The demand for cloud services has led to the rapid expansion of extensive data centers, resulting in a significant increase in electricity consumption. This heightened energy consumption has raised concerns about its environmental impact and economic sustainability. Researchers have explored innovative approaches to optimize cloud resource management to address these challenges while simultaneously upholding high-quality service levels. In this context, the integration of metaheuristic algorithms has shown great promise in tackling complex optimization problems frequently encountered in cloud computing. Our paper introduces a novel approach, the Micro-Genetic and Cuckoo Search (MG-CS) algorithm, which is tailored for power reduction and load-balancing in cloud computing. The primary contribution of this research is the development of an efficient and multifaceted approach that concurrently addresses various critical objectives:

- Load balancing excellence: MG-CS aims to achieve a well-balanced distribution of workloads across cloud resources, ensuring optimal resource utilization and averting potential performance bottlenecks.
- Dedicated power minimization: Our approach reduces energy consumption within cloud data centers, promotes environmental sustainability, and optimizes operational costs.
- Strategic cost reduction: We target minimizing resource wastage and optimizing cloud service delivery to make cloud infrastructure more cost-effective.
- Time optimization initiatives: MG-CS endeavors to improve response times and task completion rates, enhancing the overall user experience and operational efficiency of cloud services.
- SLA compliance assurance: Our approach ensures that cloud services meet predefined service level agreements (SLAs), aligning with performance and availability requirements defined by consumers.
- QoS elevation strategies: The research aims to elevate the quality of cloud services, covering aspects of reliability, scalability, and data security, thereby providing an enhanced user experience and meeting customer expectations.

The paper is organized as follows: Section II provides an overview of related work, Section III details our proposed load balancing framework using the MG-CS algorithm, Section IV

presents the experimental results, and Section V concludes our research, summarizing the contributions and potential future directions.

II. Related Work

Yakhchi, et al. [20] presented a method rooted in the CS algorithm to identify over-utilized hosts within a cloud environment. Subsequently, they employed the Minimum Migration Time (MMT) policy to systematically transfer VMs from over-utilized hosts to alternative hosts, ensuring that the migration process did not inadvertently lead to new instances of over-utilization. Following this, the researchers categorized all hosts except the over-utilized ones as underutilized, aiming to efficiently relocate VMs from these underutilized hosts to different hosts and transition the former to a sleep mode. This strategic maneuver effectively optimized both resource utilization and energy consumption. The research employed simulation using the CloudSim simulator, yielding compelling results. Specifically, their approach yielded the lowest energy consumption compared to several well-established algorithms, reaffirming the efficacy of their proposed method.

Sharma, et al. [21] have employed the bat algorithm as an approach to cloud load balancing. The bat Algorithm draws inspiration from the echolocation behavior of bats and has been proposed for this purpose. Bats, in their pursuit of prey, exhibit erratic flight patterns by altering various parameters such as velocity, pulse emission rate, position, frequency, and loudness. These alterations are made based on the proximity between the bat and its prey. The adjustment of velocities and positions of bats is incorporated in a manner similar to the PSO algorithmic. The bat algorithm is structured to achieve optimal results by running the algorithm through multiple iterations. In the context of this study, the bat algorithm is utilized to determine the most suitable server from a pool of available servers for the execution of incoming tasks. When a new task is introduced into the task pool, the load balancer initiates the bat algorithm to identify the best-suited server that matches the requirements of the incoming task. The bat algorithm takes into account factors such as task type and required resources when selecting the optimal VM for task execution. Upon selecting the appropriate server, the load balancer allocates the task to that server. If the load on the chosen server surpasses that of all other servers, the task is then distributed across multiple servers.

Devaraj, et al. [22] introduced an innovative load-balancing algorithm named FIMPSO, which represents a hybrid amalgamation of the Firefly (FF) algorithm and the Improved Multi-Objective Particle Swarm Optimization (IMPSO) technique. The FIMPSO algorithm synergizes the strengths of the FF algorithm to effectively narrow down the search space while harnessing the capabilities of the IMPSO technique to attain enhanced responsiveness. The IMPSO algorithm takes a unique approach to select the global best (gbest) particle. It does so by considering the proximity of a point to a line, enabling the identification of candidates for the gbest particle. This method significantly refines the search process, ultimately facilitating the pursuit of an optimal solution. The proposed FIMPSO algorithm is validated through its notable accomplishment in load balancing. This achievement translates to improved resource utilization and diminished task response times. The outcomes of simulations underscore the superiority of the FIMPSO model in comparison to alternative methods. Specifically, the FIMPSO algorithm exhibited exceptional performance metrics such as average response time (13.58ms), CPU utilization (98%), memory utilization (93%), reliability (67%), and throughput (72%). Additionally, the FIMPSO algorithm achieved an impressive makespan of 148, outperforming all other methodologies considered for comparison.

Jena, et al. [23] introduced an inventive approach to dynamically balance the load across VMs utilizing a hybrid strategy named QMPSO, which amalgamates a modified Particle Swarm Optimization (MPSO) technique with an enhanced Q-learning algorithm. Within the QMPSO algorithm, this fusion mechanism fine-tunes the velocity of MPSO by incorporating insights from both the global best (gbest) and personal best (pbest) solutions. These solutions are derived from the optimal actions identified through the improved Q-learning algorithm. The primary objectives driving this hybridization are to elevate the performance of virtual machines through load balancing, amplify the throughput of VMs, and uphold equilibrium between task priorities by optimizing their waiting times. To validate the robustness of the QMPSO algorithm, a comprehensive comparison was conducted. The algorithm's outcomes, gleaned from both simulation-based assessments and actual platform measurements, were juxtaposed with those generated by existing load-balancing and scheduling algorithms. The empirical evidence unequivocally demonstrated the superiority of the proposed QMPSO algorithm, underscoring its prowess in achieving load-balancing and fine-tuning the performance of virtual machines within a cloud environment.

Sefati, et al. [24] harnessed the Grey Wolf Optimization (GWO) algorithm as a means to attain effective load-balancing while considering the resource reliability capacity. In this endeavor, the GWO algorithm was employed to discern nodes that were either idle or occupied within the cloud environment. Once these nodes were identified, the algorithm proceeded to compute the threshold and fitness function for each node. The researchers conducted a simulation using CloudSim, wherein the proposed approach, leveraging the GWO algorithm, was assessed in comparison to other load-balancing methods. The results of this assessment highlighted significant advantages, including reduced costs and response times. Moreover, the solutions obtained were deemed optimal, serving as a testament to the efficacy of the load-balancing methodology founded on the GWO algorithm.

Latchoumi and Parthiban [25] have introduced a groundbreaking approach, termed the Quasi-Oppositional Dragonfly Algorithm for Load-balancing (QODA-LB), with the primary aim of attaining optimal resource scheduling within a cloud computing framework. The QODA-LB algorithm strategically integrates three pivotal variables – execution time, execution cost, and charge – to formulate an objective function. This objective function serves as the foundation for task allocation to Virtual Machines (VMs), predicated on their inherent potential. A noteworthy aspect of the QODA-LB algorithm is the incorporation of the Quasi-

Oppositional Based Learning principle. This principle confers a distinctive edge by elevating the standard convergence rate of the Dragonfly algorithm (DA). The integration of this principle enhances the efficacy of load-balancing and resource scheduling within the cloud environment. A comprehensive series of experiments was meticulously conducted to assess the QODA-LB algorithm's performance. The ensuing results were scrutinized from diverse angles to validate its heightened efficiency. The outcomes of simulations substantiated the algorithm's exceptional load-balancing efficiency, positioning it as a superior alternative to other foundational approaches for load-balancing and resource scheduling in the realm of cloud computing.

Haris and Zubair [26] introduced a dynamic load-balancing algorithm named Mantaray modified multi-objective Harris hawk optimization (MMHHO) that draws inspiration from hybrid optimization algorithms. This innovative approach leverages the strengths of the Harris Hawk Optimization (HHO) algorithm, enhancing its search space through integration with the Manta Ray Foraging Optimization (MRFO) algorithm. The hybridization process strategically melds various factors, including cost, response time, and resource utilization, to streamline the load-balancing process. The MMHHO algorithm sets its sights on optimizing system performance by bolstering VM throughput, achieving equilibrium in load distribution among VMs, and harmonizing task priorities by adjusting their waiting times. The implementation of the MMHHO-based load-balancing algorithm is realized through the utilization of the CloudSim tool. This platform provides the means to assess the algorithm's effectiveness across various parameters and compare its performance against other established load-balancing algorithms. Upon meticulous analysis and simulation, the results unequivocally underscore the supremacy of the proposed MMHHO load-balancing scheme. In terms of system performance and efficiency, the MMHHO algorithm surpasses its counterparts, thereby validating its potential to elevate load-balancing processes and enhance the overall effectiveness of the system.

III. Proposed Load-Balancing Framework

A. Problem Statement

The importance of autonomic load-balancing in the cloud computing domain stems from its capacity to elevate throughput through the optimized utilization of resources. The load-balancing strategies and power management strategies put forth in this proposal are geared towards the automatic and efficient allocation of computational resources within the cloud infrastructure. This is achieved by evaluating the suitability of all tasks concerning resource availability. The effectiveness of the load-balancing approach is determined using intersection formulas, with the most common ones being represented by Eq. (1) and Eq. (2). These formulas play a crucial role in the assessment of task-resource mapping, enabling the system to achieve improved performance and better resource allocation in the cloud environment.

$$K = (G + 2\sqrt{g})/3G \quad (1)$$

$$X = \begin{bmatrix} 0 & gen \\ 1 & gen \end{bmatrix} X \frac{\pi}{2} \quad (2)$$

Fig. 1 depicts the architecture of the load-balancing framework, encompassing three fundamental stages:

- Optimal resource utilization: This phase focuses on achieving efficient resource utilization by effectively managing cloud resources and handling the workload coming from cloud users. Clustering and VM deployment support are employed to ensure optimal resource provisioning.
- Workload submission and demand-based processing: During this phase, cloud users submit their requests and workloads based on their specific demands. The system takes into account energy consumption while processing and managing the workload.
- Minimizing power consumption: The framework emphasizes minimizing power consumption to reduce the environmental impact and operational costs within the cloud.

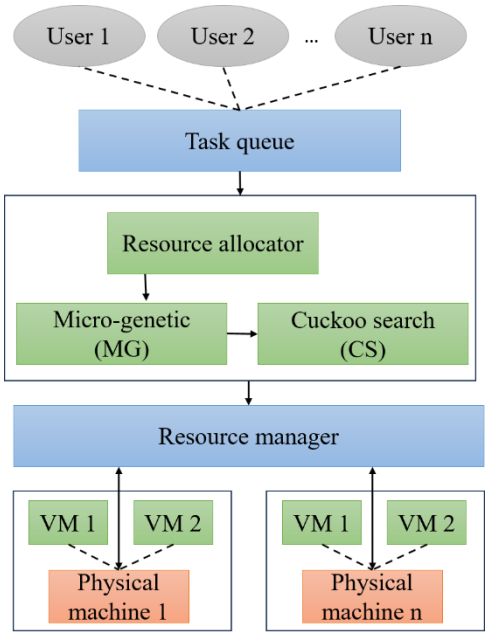


Fig. 1. Proposed load-balancing framework

Key terminologies and components within the load balance framework are as follows:

- Cloud users: Entities, whether individuals or businesses, who make use of cloud storage services to conveniently manage, store, and access their computing resources from any location.
- SLA administration: Service-level agreements (SLAs) provide assurance to customers and enable cloud providers to prioritize the fulfillment of their particular needs and expectations. Active management of SLAs is crucial, as they represent more than just guidelines and function as contracts.

- Workload scheduling and clustering: This method clusters and schedules similar workloads to the corresponding virtual machines based on Quality of Service (QoS) and SLA considerations.
- QoS management: This component is responsible for managing any Quality of Service (QoS) specifications linked to workload access, ensuring that the workload is effectively handled in accordance with its specific QoS requirements.

Algorithm 1 outlines the process of grouping workloads based on their center points by optimizing an objective function to achieve the minimal value. The algorithm aims to create clusters by utilizing the designated center points as group representatives. Each workload in a cluster is highly likely to belong to the cluster represented by its nearest center point. However, workloads can only belong to a single cluster, except for the center point, which may be part of multiple clusters.

Algorithm 1. Workload clustering

No of Clusters commit to complete which will be decided by C.

STEP-1: START

STEP-2: The non-empty subclasses of object C will be divided at random.

STEP-3: Cluster centroids are currently the separating seed points of clusters.

STEP-4: An object will be paired with another object whose seed points are closer.

STEP-5: END

B. Micro-genetic Algorithm

Genetic Algorithms (GAs) belong to the family of Evolutionary Algorithms (EAs) and are widely recognized as one of this family's earliest and most well-known members. In GAs, elitism, which involves preserving the best individuals (solutions) in the population, is promoted through two fundamental mechanisms:

- Environmental selection: Environmental selection aims to remove the worst-fitted individuals from the current population, ensuring they do not contribute to the next generations. This allows the fittest individuals to have a higher chance of survival and progression.
- Parent selection: In the parent selection process, the algorithm promotes generating offspring solutions from the population's best individuals (the elite solutions). Non-elite solutions are excluded from this process, further reinforcing the elitism aspect of the algorithm.

Micro-GAs (MGs) are a specific type of GA with minimal populations. Due to the use of such small populations, MGs exhibit a high level of elitism. In MGs, the environmental selection has lower survivor rates than canonical GAs, as the focus is on preserving only the best solutions. Additionally, the parent selection process only allows elite solutions to generate offspring, further enhancing the elitism effect. The genetic algorithm demonstrates the capability to quickly generate high-quality local optimal solutions while maintaining competitiveness in the long term. This makes it an effective approach for solving problems with computationally intensive fitness functions.

Nevertheless, when dealing with problems that encompass high-dimensional parameter spaces, attaining the convergence of all model parameters within a specified margin of error can present difficulties and consume a substantial amount of time. As the count of model parameters expands, the genetic algorithm necessitates a larger population size, resulting in an increased volume of cost-function analyses. This can be computationally expensive, especially when dealing with high-dimensional problems. In such scenarios, micro-genetic algorithms offer a viable alternative. These algorithms operate with very small populations, which help reduce the computational burden while maintaining a high level of elitism. The smaller population size allows for a more focused search, and the algorithm can swiftly converge to promising solutions without the need for a large number of cost-function evaluations.

C. Cuckoo Search Algorithm

The CS algorithm is inspired by the egg-laying strategy of cuckoo birds, where they lay their eggs in the nests of other bird species. This nature-inspired optimization technique simulates this behavior to explore complex search spaces and discover optimal solutions. Cuckoos employ a Levy flight strategy to select nests, frequently opting for nests where the host bird has recently deposited its own eggs. This behavior enhances the likelihood of their eggs successfully hatching. Notably, certain female cuckoos mimic the colors and patterns of host eggs to decrease the chances of their eggs being rejected, thereby amplifying their reproductive success. The foraging behavior of animals, including insects, follows a quasi-random pattern, effectively resembling a random walk. This behavior has been observed in many animals and has been mathematically modeled as Lévy flights. Lévy flights involve making successive movements with step lengths drawn from a Lévy distribution, which allows for long jumps that facilitate efficient exploration of large search spaces. Based on this concept, researchers have applied Lévy flights to optimization and search problems, resulting in the development of the CS algorithm. Preliminary results have shown promising capabilities of this algorithm in finding optimal solutions for a wide range of optimization problems. By imitating the natural behavior of cuckoos and incorporating Lévy flights, the Cuckoo Search Algorithm offers a powerful and efficient approach for tackling complex optimization challenges. The CS algorithm models the natural behavior of cuckoos and can be described using the following idealized rules:

- Each cuckoo lays a single egg at a time, selecting a nest at random for deposit. Nests with superior egg quality (improved solutions) are more likely to persist across subsequent generations.
- The count of available host nests is constant, represented as 'n,' and the host bird has a probability of detecting an alien egg within the range of [0, 1].

- When an alien egg is detected, the host bird has the choice to either discard it or desert the nest to construct a new one at a distinct location.

To simplify, this final assumption can be approximated using a probability of pa for each of the n nests. With these rules in mind, the fundamental steps of the CS algorithm can be succinctly summarized in pseudocode as follows:

1) Initialize the population of cuckoos (solution candidates).
2) Evaluate the quality (fitness) of each cuckoo.
3) Identify the best cuckoos and their nests for further reproduction
4) Repeat until stopping criteria are met:
5) Generate new cuckoo solutions by performing Levy flights
6) Evaluate the fitness of newly generated cuckoos
7) Replace the old cuckoos with the new cuckoos in the nests based on their fitness
8) Abandon and rebuild nests (cuckoos) with a probability of p_a
9) If a host bird discovers an alien egg with probability p_a:
10) Throw away the alien egg or abandon the nest and build a new one
*11) Identify t*he best solution found and return it as the final result

In the CS algorithm, the movement of each cuckoo from generation *t* to *t+1* is represented by a vector *x* with entries $X_i(t+1)$ and is calculated by Eq. (3).

$$X_i(t+1) = X_i(t) + \alpha \oplus \text{Lévy}(u) \quad (3)$$

Where $X_i(t)$ is the current position of the i^{th} cuckoo at generation *t*, $\alpha > 0$ is the step size, which depends on the scale of the given problem, $\oplus$ represents entry-wise multiplication, and *Lévy(u)* is determined using the Lévy flight, a random step-length process. The expression for *Lévy(u)* is given by:

$$\text{Lévy}(u) = t^{(-\lambda)} \quad (4)$$

Where λ is a parameter, typically within the range $1 < \lambda \leq 3$, *t* is the current generation. The Lévy flight results in a power-law step-length distribution with a heavy tail, making cuckoos more exploratory. In the real world, if a cuckoo's egg closely resembles the host's eggs, it is less likely to be discovered by the host bird. To mimic this behavior, the CS algorithm performs a random walk in a biased way with some random step sizes. This biased random walk, guided by the Lévy flight, allows the algorithm to explore the search space more effectively, discovering better solutions in complex optimization problems.

D. MG-CS Algorithm

Consider a cloud environment with multiple VMs where diverse workloads are dynamically generated based on user demands. The goal is to efficiently distribute these workloads across the available VMs to ensure optimal resource utilization, prevent overloads or under-utilization, and ultimately enhance the overall performance of the cloud system. The algorithm follows the steps described below.

- Initialization: The algorithm begins by randomly entering tasks into a task memory divided into replaceable and non-replaceable tasks. Various parameters such as states, positions, steps, and visual parameters are set up during this phase.
- Task selection: The algorithm selects tasks from the task memory for further processing.
- Crossover and mutation: The selected tasks undergo crossover and mutation operations to generate new potential solutions.
- Patronize: The algorithm evaluates the fitness level of each potential solution.
- New tasks and convergence: The algorithm tracks the best MG values and investigates their behaviors. If the fitness of MG exceeds the predefined threshold (bulletin value), the MG's fitness is updated in the bulletin.
- Filter and external memory: The algorithm uses a filter to refine the solutions and stores valuable information in the external memory, facilitating feedback with both sides of the task memory.
- Final solution: The CS step performs the optimal solution chosen from the population and decodes it to determine the most appropriate resource assignment to tasks based on their availability and throughput.

IV. EXPERIMENTAL RESULTS

The experiment was conducted in CloudSim, a simulation tool devised by cloud laboratories situated in Melbourne. Within this experiment, a total of 50 tasks were examined within a simulation framework encompassing 25 VMs. Each VM was equipped with 2048 MB of RAM. Fig. 2 to 7 present various performance metrics and comparisons of the proposed MG-CS method with existing approaches in a cloud simulation environment. The experiments were conducted with different numbers of workloads and servers. Fig. 2 depicts the availability rate in relation to various workloads. MG-CS exhibited a diverse spectrum of results, with an availability rate of up to 55% with 500 workloads. As the workload increased, the availability rate decreased, reaching 90% with 3000 workloads. Fig. 3 provides an illustration of the reliability rate as it correlates with different workloads. The MG-CS again demonstrated a diverse range of results. It achieved a reliability rate of up to 48% with 1000 workloads, which decreased as the workload increased. Ultimately, it achieved a reliability rate of 72% with 3000 workloads, outperforming the existing system's performance.

In Fig. 4, the resource utilization pattern is displayed alongside varying workloads. The MG-CS approach demonstrated notable efficacy, achieving a peak resource utilization of 80% when subjected to 2500 workloads. Fig. 5 provides a visual representation of the SLA violation rates in relation to varying workloads. Notably, the MG-CS system presented a remarkably low violation rate compared to its counterparts. Specifically, it exhibited a mere 4% violation rate when confronted with 500 workloads and a slightly higher

10.5% violation rate when handling 3000 workloads. Fig. 6 presents the energy consumption during workload processing. The introduced MG-CS approach effectively minimized energy consumption in comparison to comparative ones. To illustrate, when subjected to 2000 workloads, the energy consumption was notably reduced to 400 kW. Fig. 7 provides a visual contrast of execution times across diverse methodologies and workloads. Impressively, the MG-CS system consistently accomplished the processing of 500 to 3,000 workloads within a time span of 5000 to 6,000 seconds. This remarkable efficiency in execution time sets the MG-CS approach apart from existing techniques, highlighting its superior performance.

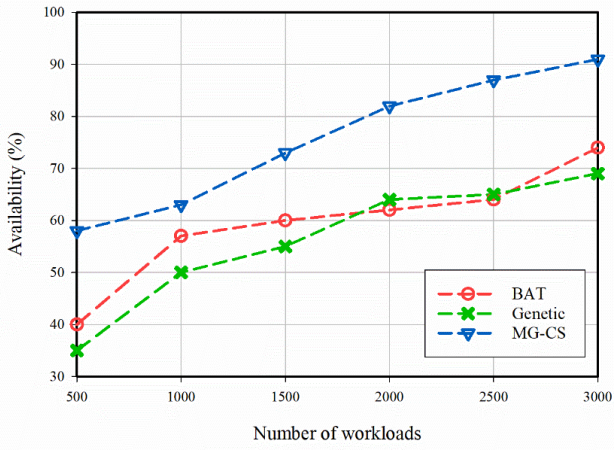


Fig. 2. Availability comparison

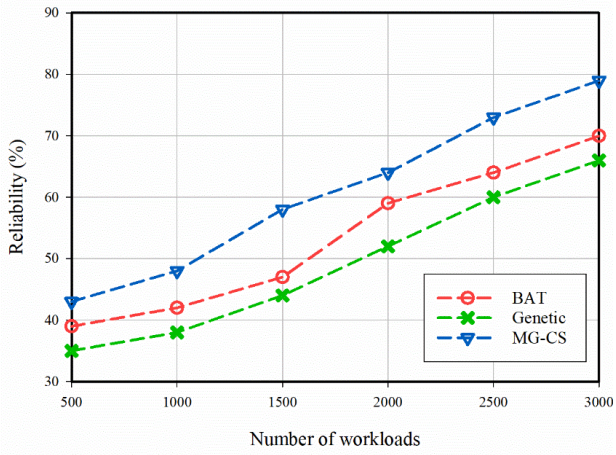


Fig. 3. Reliability comparison

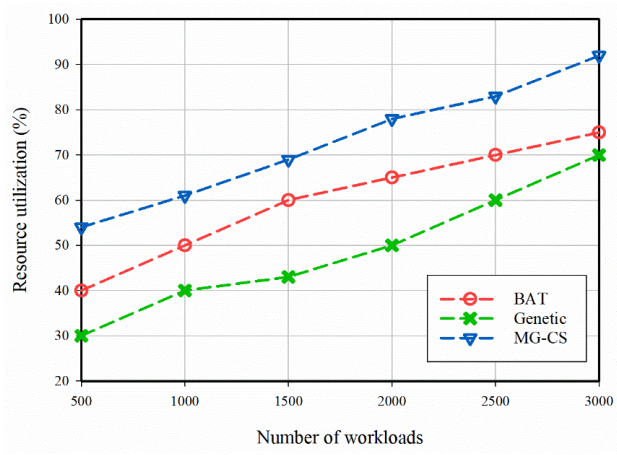


Fig. 4. Resource utilization comparison

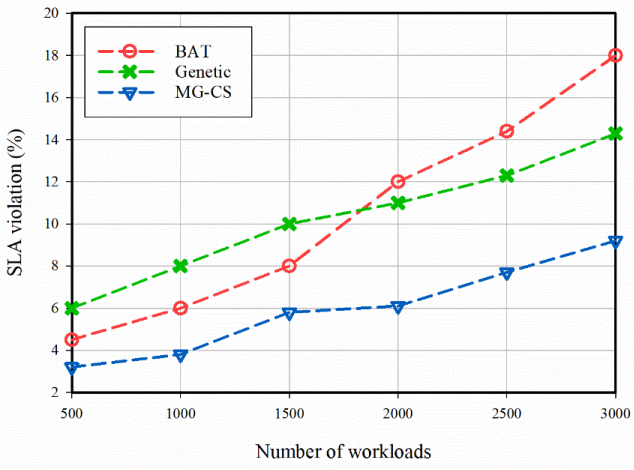


Fig. 5. SLA violation comparison

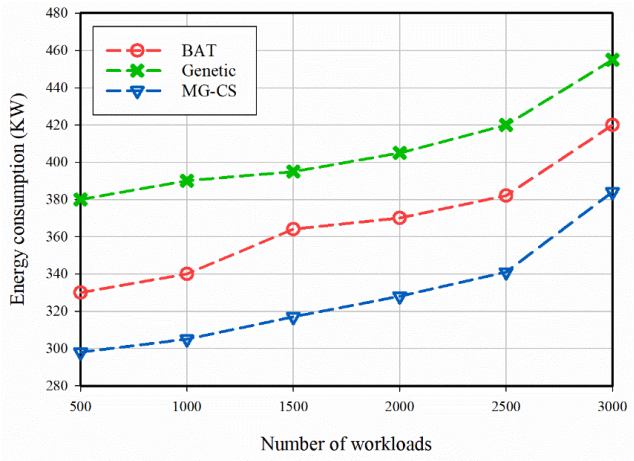


Fig. 6. Energy consumption comparison

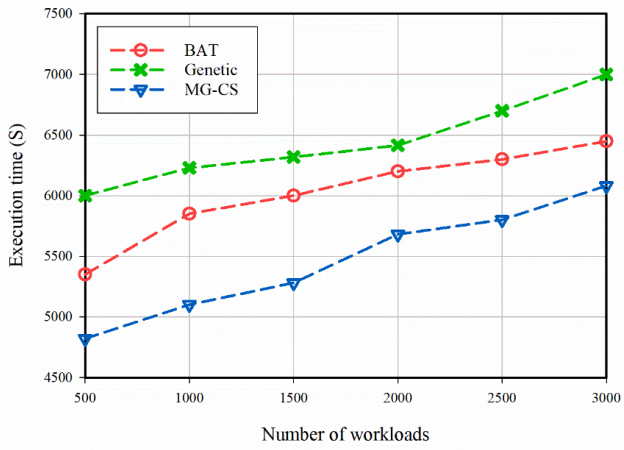


Fig. 7. Execution time comparison

The experimental results, as presented in Fig. 2 to 7, reveal the significant impact of the proposed MG-CS approach on various performance metrics in a cloud simulation environment. Notably, the results illustrate the adaptability and efficacy of MG-CS across a range of workloads and server configurations. In terms of availability, Fig. 2 demonstrates that MG-CS exhibited a diverse spectrum of results, achieving an availability rate of up to 55% with 500 workloads. As the workload increased, availability decreased but remained robust, reaching 90% with 3000 workloads. Similarly, in Fig. 3, the reliability rate showcased a wide range of outcomes. MG-CS achieved a reliability rate of up to 48% with 1000 workloads, surpassing existing systems. Fig. 4 showcases the resource utilization pattern, with MG-CS achieving a remarkable peak utilization of 80% when subjected to 2500 workloads, signifying its efficiency. In terms of SLA violation rates (Fig. 5), MG-CS demonstrated an impressively low violation rate, with just 4% for 500 workloads and a slightly higher 10.5% for 3000 workloads, highlighting its ability to meet service level agreements. Furthermore, Fig. 6 illustrates the energy consumption during workload processing, with MG-CS effectively minimizing energy consumption, reducing it to 400 kW when subjected to 2000 workloads. Lastly, Fig. 7 provides a visual contrast of execution times, underscoring MG-CS's remarkable efficiency in processing 500 to 3,000 workloads within a period of 5,000 to 6,000 seconds. These findings emphasize the significance of the MG-CS approach in enhancing cloud resource management, achieving load balance, and optimizing operational efficiency while meeting service level agreements and reducing energy consumption.

Table I presents the dimensions of diverse synthetic datasets along with the associated task quantities. The "extra-large" dataset encompasses 800-1000 tasks, and each task's magnitude falls within the range of 100,000-200,000MI. Similarly, the "large" dataset comprises 600-700 tasks, with task sizes ranging from 70,000-100,000MI. Correspondingly, the "medium-sized" dataset entails 400-500 tasks, and the tasks vary in size between 50,000-70,000MI. Likewise, the "small-sized" dataset encompasses 100-200 tasks, with task sizes spanning from 30,000-50,000MI. It is noteworthy to mention that task sizes were generated randomly during runtime, and their size is denoted in Millions of Instructions (MI). Moreover, the research utilized a total of 80 servers, each characterized by distinct resource capacities and loads. Each server hosted different types of VM instances, featuring varying CPU and memory capacities, as outlined in Table II. Fig. 8 illustrates the outcomes obtained through the proposed method concerning CPU utilization. The graph clearly demonstrates that, in comparison to FIMPSO, MG-CS consistently achieved the highest CPU utilization across all task categories.

TABLE I. DATASETS DESCRIPTION

Type of tasks	Size of tasks (MI)	Number of tasks
Small	100-200	30000-50000
Medium	400-500	50000-70000
Large	600-700	70000-100000
Extra-large	800-1000	100000-200000

TABLE II. TYPES OF VM INSTANCES

Type of tasks	Memory capacity (GB)	CPU capacity (MIPS)
Small	5	10000
Medium	10	20000
Large	15	25000
Extra-large	20	35000

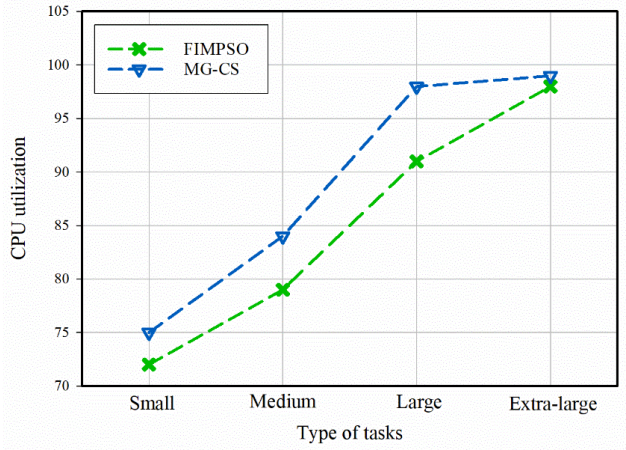


Fig. 8. CPU utilization comparison.

V. CONCLUSION

This paper presented MG-CS, a load-balancing resource allocation approach aimed at improving the utilization of cloud resources. The proposed method's experimental results demonstrate its effectiveness in addressing various QoS factors, including availability, reliability, resource utilization, SLA violation, energy Consumption, and execution time. The proposed method is compared with existing algorithms for QoS parameter efficiency, and the experimental results show that the MG-CS technique outperforms existing GA and BAT algorithms in terms of cost, timing, and energy. The experimental findings prove that MG-CS is the effectiveness of MG-CS in accomplishing both optimal scheduling and load-balancing objectives. Furthermore, the approach ensures the preservation of superior QoS standards while steadfastly adhering to SLA requirements during cloud services. In the

future, the researchers plan to incorporate artificial intelligence self-learning methods to facilitate large-scale data sources. The method can enhance its performance and adaptability by integrating AI capabilities in handling complex and dynamic cloud environments. These advancements are expected to contribute to more efficient and robust cloud resource allocation, making cloud services more reliable and cost-effective for users.

MG-CS offers several notable benefits in the context of cloud load balancing and power minimization. It excels in achieving superior load distribution across cloud resources, ensuring optimal resource utilization, and averting performance bottlenecks. Moreover, the MG-CS algorithm effectively reduces energy consumption within cloud data centers, promoting environmental sustainability and cost efficiency. The method optimizes cloud service delivery, enhancing resource utilization and minimizing operational expenses. Furthermore, the approach enhances response times, task completion rates, and overall QoS, improving the user experience. However, like any approach, there are limitations to consider. The computational complexity of MG-CS may pose challenges in large-scale cloud environments. Additionally, the algorithm's performance could be influenced by the specific workload characteristics, and it may require fine-tuning for optimal results. Moreover, while MG-CS demonstrates robust performance in our experiments, its generalizability to diverse cloud infrastructures and real-world scenarios may need further investigation. These limitations underscore the need for ongoing research to fine-tune and adapt MG-CS for various cloud computing contexts and scenarios.

REFERENCES

[1] B. Pourghebleh, A. A. Anvigh, A. R. Ramtin, and B. Mohammadi, "The importance of nature-inspired meta-heuristic algorithms for solving virtual machine consolidation problem in cloud environments," Cluster Computing, pp. 1-24, 2021.

[2] V. Hayyolalam, B. Pourghebleh, M. R. Chehrehzad, and A. A. Pourhaji Kazem, "Single-objective service composition methods in cloud manufacturing systems: Recent techniques, classification, and future trends," Concurrency and Computation: Practice and Experience, vol. 34, no. 5, p. e6698, 2022.

[3] V. Hayyolalam, B. Pourghebleh, A. A. P. Kazem, and A. Ghaffari, "Exploring the state-of-the-art service composition approaches in cloud manufacturing systems to enhance upcoming techniques," The International Journal of Advanced Manufacturing Technology, vol. 105, no. 1-4, pp. 471-498, 2019.

[4] K. N. Qureshi, G. Jeon, and F. Piccialli, "Anomaly detection and trust authority in artificial intelligence and cloud computing," Computer Networks, vol. 184, p. 107647, 2021.

[5] Q. Yu, L. Chen, and B. Li, "Ant colony optimization applied to web service compositions in cloud computing," Computers & Electrical Engineering, vol. 41, pp. 18-27, 2015.

[6] B. Pourghebleh and N. J. Navimipour, "Data aggregation mechanisms in the Internet of things: A systematic review of the literature and recommendations for future research," Journal of Network and Computer Applications, vol. 97, pp. 23-34, 2017.

[7] F. Ebadifard and S. M. Babamir, "Autonomic task scheduling algorithm for dynamic workloads through a load balancing technique for the cloud-computing environment," Cluster Computing, vol. 24, no. 2, pp. 1075-1101, 2021.

[8] I. Z. Yakubu and M. Murali, "An efficient meta-heuristic resource allocation with load balancing in IoT-Fog-cloud computing environment," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 3, pp. 2981-2992, 2023.

[9] M. Dorigo, M. Birattari, and T. Stutzle, "Ant colony optimization," IEEE computational intelligence magazine, vol. 1, no. 4, pp. 28-39, 2006.

[10] T. M. Shami, A. A. El-Saleh, M. Alswaitti, Q. Al-Tashi, M. A. Summakieh, and S. Mirjalili, "Particle swarm optimization: A comprehensive survey," IEEE Access, 2022.

[11] L. Abualigah and A. Diabat, "Advances in sine cosine algorithm: a comprehensive survey," Artificial Intelligence Review, vol. 54, no. 4, pp. 2567-2608, 2021.

[12] S. Mahmoudinazlou, A. Alizadeh, J. Noble, and S. Eslamdoust, "An improved hybrid ICA-SA metaheuristic for order acceptance and scheduling with time windows and sequence-dependent setup times," Neural Computing and Applications, pp. 1-19, 2023.

[13] B. M. Jafari, M. Zhao, and A. Jafari, "Rumi: An Intelligent Agent Enhancing Learning Management Systems Using Machine Learning Techniques," Journal of Software Engineering and Applications, vol. 15, no. 9, pp. 325-343, 2022.

[14] S. R. Abdul Samad et al., "Analysis of the Performance Impact of Fine-Tuned Machine Learning Model for Phishing URL Detection," Electronics, vol. 12, no. 7, p. 1642, 2023.

[15] S. P. Rajput et al., "Using machine learning architecture to optimize and model the treatment process for saline water level analysis," Journal of Water Reuse and Desalination, 2022.

[16] S. Aghakhani, A. Larijani, F. Sadeghi, D. Martín, and A. A. Shahrakht, "A Novel Hybrid Artificial Bee Colony-Based Deep Convolutional Neural Network to Improve the Detection Performance of Backscatter Communication Systems," Electronics, vol. 12, no. 10, p. 2263, 2023.

[17] W. Anupong et al., "Deep learning algorithms were used to generate photovoltaic renewable energy in saline water analysis via an oxidation process," Water Reuse, vol. 13, no. 1, pp. 68-81, 2023.

[18] S. Vairachilai, A. Bostani, A. Mehbodniya, J. L. Webber, O. Hemakesavulu, and P. Vijayakumar, "Body Sensor 5 G Networks Utilising Deep Learning Architectures for Emotion Detection Based On EEG Signal Processing," Optik, p. 170469, 2022.

[19] M. Khodayari, J. Razmi, and R. Babazadeh, "An integrated fuzzy analytical network process for prioritisation of new technology-based firms in Iran," International Journal of Industrial and Systems Engineering, vol. 32, no. 4, pp. 424-442, 2019.

[20] M. Yakhchi, S. M. Ghafari, S. Yakhchi, M. Fazeli, and A. Patooghi, "Proposing a load balancing method based on Cuckoo Optimization Algorithm for energy management in cloud computing infrastructures," in 2015 6th International Conference on Modeling, Simulation, and Applied Optimization (ICMSAO), 2015: IEEE, pp. 1-5.

[21] S. Sharma, A. K. Luhach, and S. Sinha, "An optimal load balancing technique for cloud computing environment using bat algorithm," Indian J Sci Technol, vol. 9, no. 28, pp. 1-4, 2016.

[22] A. F. S. Devaraj, M. Elhoseny, S. Dhanasekaran, E. L. Lydia, and K. Shankar, "Hybridization of firefly and improved multi-objective particle swarm optimization algorithm for energy efficient load balancing in cloud computing environments," Journal of Parallel and Distributed Computing, vol. 142, pp. 36-45, 2020.

[23] U. Jena, P. Das, and M. Kabat, "Hybridization of meta-heuristic algorithm for load balancing in cloud computing environment," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 6, pp. 2332-2342, 2022.

[24] S. Sefati, M. Mousavinasab, and R. Zareh Farkhady, "Load balancing in cloud computing environment using the Grey wolf optimization algorithm based on the reliability: performance evaluation," The Journal of Supercomputing, vol. 78, no. 1, pp. 18-42, 2022.

[25] T. P. Latchoumi and L. Parthiban, "Quasi oppositional dragonfly algorithm for load balancing in cloud computing environment," Wireless Personal Communications, vol. 122, no. 3, pp. 2639-2656, 2022.

[26] M. Haris and S. Zubair, "Mantaray modified multi-objective Harris hawk optimization algorithm expedites optimal load balancing in cloud computing," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 10, pp. 9696-9709, 2022.

A Focal Loss-based Multi-layer Perceptron for Diagnosis of Cardiovascular Risk in Athletes

Chuan Yang*
School of Public Utility Management, Chongqing Vocational College of Transportation, Chongqing 402247, China

***Abstract*—Cardiovascular diseases (CVDs) are a prevalent cause of heart failure around the world. This research was required in order to investigate potential approaches to treating the disease. The article presents a focal loss (FL)-based multi-layer perceptron called MLP-FL-CRD to diagnose cardiovascular risk in athletes. In 2012, 26,002 athletes were measured for their height, weight, age, sex, blood pressure, and pulse rate in a medical exam that had electrocardiography at rest. Outcomes were negative for the largest majority, leading to class imbalance. Training on imbalanced data hurts classifier performance. To address this, the study proposes a training approach based on focal loss, which effectively emphasizes minority class examples. Focal loss softens the influence of simplistic samples, enabling the model to concentrate on more intricate examples. It is useful in circumstances when there is a substantial class imbalance. Additionally, the paper highlights a challenge in the training phase, often characterized by the use of gradient-based learning methods like backpropagation. These methods exhibit several disadvantages, including sensitivity to initialization. The paper recommends the implementation of a mutual learning-based artificial bee colony (ML-ABC). This approach adjusts the primary weight by substituting the food resource candidate, which is selected due to superior fitness, with one based on a mutual learning factor between two individuals. The sample obtains great outcomes, outperforming other machine learning samples. Optimal values for important parameters are identified for the model based on experiments on the study dataset. Ablation studies that exclude FL and ML-ABC of the sample confirm the additive effect of, which is not negative and dependent, these factors on the sample's efficiency.**

Keywords*—*Cardiovascular diseases; multi-layer perceptron; focal loss; artificial bee colony; imbalanced classification

I. Introduction

As per the World Health Organization's discoveries, cardiovascular diseases are responsible for the most deaths across the globe, leading to 17.9 million fatalities every year [1]. Smoking, obesity, high blood pressure, and physical inactivity are the most significant risk elements. Even though the risk of cardiovascular diseases can be minimized through regular exercise, athletes competing or training with intensity and frequency remain at risk due to the intensity and frequency of their activities. Athletes are consistently monitored by sports physicians, who gain biomedical and personal information, along with executing electrocardiography screenings (ECG). According to the ECG results, people are classified as either a risk factor or not a risk factor. Those deemed at risk may not gain medical clearance for participation in sports and will be subject to further assessments. There is an unequal ratio between the two categories, as the N class contains a larger percentage of individuals, while the more remarkable P class is not very common. Generally speaking, a false negative (FN) may have severe consequences, potentially including a human fatality. In contrast, a false positive (FP) would lead to additional medical examinations and temporarily halt sports practice [2].

Classification is indeed a method of machine learning that can expect binary classification values, such as P or N, and it could have great use in the health world, particularly in diagnostics. Several machine learning techniques have already been implemented in medical diagnosis systems to assist decision-making. These techniques may estimate health risks from huge datasets. Wong et al. [3] utilized Bayesian networks to uncover past disease outbreaks and achieved remarkable outcomes on actual data taken from a database that contained seven years of medical records from an emergency region. The dataset was collected from unwell patients in a hospital, indicating that the research was aimed at epidemiological investigation rather than diagnostic purposes. This strategy could also be implemented in counterterrorism to discover biological attacks. Campbell et al. [4] utilized a kernel-based approach that effectively identified a rare disease from medical data. However, the size of the dataset was limited, and the proportion of noteworthy instances in the test set was considerably greater than the prevalence of the disease among the overall population. Fontaine et al. [5] investigated methods of data mining to enhance the clinical assessment of patients with brain disorders. Salam and McGrath [6] conducted a machine-learning strategy for dermatology. In this study, a classifier for multiple diseases improved the identification of skin infections. Sacchi et al. [7] applied a Naive Bayes classifier for glaucoma prediction. Because of the small and unbalanced dataset, resampling and bootstrapping were employed to train their model. Numerous authors [8, 9] have suggested comparing different classification methods in medical statistics to identify the significant benefits of one approach compared to others. There is still a dearth of studies using large datasets in regions where diseases have a low prevalence, but individuals may face increased risks due to heightened levels of pressure or stress. Moreover, there is a continuous debate about the necessity for affordable and efficient healthcare, as well as the more cautious utilization of medical tests.

In machine learning approaches, the strategy for extracting features is inflexible, resulting in poor generalization ability, rising time, and low precision [10]. With the emergence of profound learning methods in numerous usages [11], numerous investigators have employed those for categorization [12].

Deep learning can accurately learn high-level features due to its layered structure. Multi-layer perceptron is a universal approximation initially developed for nonlinear XOR and has been implemented effectively for diverse combinatory enhancement issues ever since [13]. MLP is widely utilized for a variety of tasks, including information processing, pattern recognition, classification issues, image processing, linear and nonlinear optimization issues, and real data prediction. MLP functions as a universal approximation where input signals propagate forward. The processing node related to a human neuron is the fundamental component of the ANN method. Each processing node gets a collection of intake deals, adds them, and after that, passes this sum through an activation function that determines the node's output value. In MLP, nodes comprise fully interconnected layers, with the exception that nodes within the same layer are not interconnected [14].

Medical classification faces significant challenges due to data imbalance, which can significantly lower performance because there are far more negative instances [15, 16]. Over-sampling, under-sampling, or a compound of both are used in the data-level strategy to mitigate the negative effects of imbalanced classification [17, 18]. Approaches at the algorithmic level give the minority class more weight [19]. Additionally, profound learning techniques can be used to solve the categorization balancing problem [20]. Huang et al. [21] formulate a process to identify distinguishing features of imbalanced data while upholding inter-cluster and inter-class margins. Yan et al. [22] suggested a technique using the bootstrapping method to balance data of convolutional networks across mini-batches.

Deep models are widely used in natural language processing, computer vision, and medical image analysis [23, 24]. The best solution for optimizing the neural network can be selected from a population of created models using population-based training. This approach is less likely to get stuck in local optima than traditional training methods [25, 26]. Indeed, a simple evolutionary algorithm was found to rival stochastic gradient descent for neural network training. Jaderberg et al. [27] applied population-based training to state-of-the-art models of deep RL, machine translation, and generative adversarial networks and demonstrated consistent improvements in accuracy, training time, and stability. In [28] and [29], effective training of the weights of neural networks was achieved using a differential evolution-based strategy and Artificial Bee Colony (ABC), respectively. The ABC algorithm can be improved by the mutual learning-based ABC [30], which changes the algorithm to use mutual learning between two selected position parameters instead of choosing the candidate food source with the highest fitness [31].

In response to the challenges outlined, this article introduces an FL-based MLP, designated as MLP-FL-CRD, for diagnosing cardiovascular risk in athletes within an imbalanced dataset. The proposed MLP-FL-CRD model contains an MLP for sick and healthy inputs that employ FL to class imbalance. An ML-ABC is utilized for weight initialization to identify a promising area within the study environment for initiating the BP method in the model. This process involves continuous evaluation of the understanding of athleticism, achieved through shared knowledge among current and nearby nutrition resources, to develop a more valuable nutrition resource. The MLP-FL-CRD model is assessed on a dataset comprising medical examinations of 26,002 athletes, demonstrating its superiority over other approaches. Furthermore, ablation studies are conducted to evaluate the relative contributions of the FL and pre-training strategies. Various alternative model component options, e.g., evolutionary algorithms and loss functions, are tested and compared in a series of experiments on the same dataset to investigate how to obtain the best results.

The main contributions of the proposed model are as follows:

- Innovative Use of FL for Class Imbalance: The model employs focal loss to effectively manage the class imbalance in the dataset. This approach is particularly beneficial in circumstances with substantial class disparities, as it reduces the impact of simpler, more common examples (negative outcomes in this case). This allows the model to focus more on complex, minority class examples (positive cases), which are crucial for accurate diagnosis in medical settings.
- Identification of Optimal Model Parameters: The research has also led to the identification of optimal values for key parameters in the model, which is crucial for its application in real-world scenarios. This ensures that the model can be fine-tuned for maximum efficiency and accuracy in diagnosing cardiovascular risk.
- Enhanced Training Approach with ML-ABC: The model introduces a ML-ABC for the training phase. This method innovatively addresses the sensitivity to initial weight settings inherent in gradient-based learning methods like backpropagation. By using ML-ABC, the model adapts the primary weights based on mutual learning factors between two individuals, leading to a more efficient training process and potentially better performance.

The organization of this paper is as follows. Section I is about the Introduction, Section II delves into related work. Method is proposed in Section III. Section IV presents the experimental results and proposes ideas for further endeavors and Section V concludes the conclusion.

II. Related Work

In the past few years, the health sector has witnessed notable progress in data analysis and the application of machine learning methods. These approaches have been extensively embraced and proven effective across a range of medical applications, especially within cardiac medicine. The burgeoning growth of medical data affords scholars a unique chance to devise and evaluate novel algorithms in this domain. CVDs continue to be a predominant cause of death in less developed countries [32, 33], and pinpointing risk elements and preliminary indicators of such illnesses is now a crucial research endeavor. The adoption of data analysis and machine learning strategies in this realm could substantially contribute to the timely detection and deterrence of cardiac diseases.

To date, research employing machine learning techniques has been advanced for CVD. The research conducted by Narain et al. [34] aimed to develop a sophisticated, machine-learning-driven CVD prognosis model to enhance the accuracy of the established Framingham risk score (FRS). Utilizing data from 689 subjects exhibiting CVD symptoms and a validation cohort from the Framingham study, the newly proposed model, employing a quantum neural network to discern CVD patterns, underwent experimental validation and comparison with the FRS. The model's proficiency in predicting CVD risk was ascertained to be 98.57%, substantially surpassing the 19.22% accuracy of the FRS and other current methods. The findings suggest the model could serve as a valuable asset for medical professionals in predicting CVD risk, thus aiding in the formulation of superior treatment strategies and promoting prompt diagnosis. Shah et al. [35] sought to construct a cardiovascular disease prediction model using machine learning tools. The data, comprising 303 records and 17 attributes, were sourced from the Cleveland heart disease database available in the UCI machine learning repository. The team employed several supervised classification techniques, including naive Bayes, decision tree, random forest, and k-nearest neighbor (KKN). The study's outcomes revealed that the KKN model demonstrated the highest predictive accuracy, reaching 90.8%, underscoring the promise of machine learning methods in forecasting cardiovascular disease and the importance of model and technique selection for best results. Drod et al. [36] embarked on a study to pinpoint crucial CVD risk factors among patients suffering from metabolic-associated fatty liver disease (MAFLD), utilizing machine learning methodologies. They performed blood biochemistry analyses and subclinical atherosclerosis assessments on a cohort of 191 individuals diagnosed with MAFLD. The research team crafted a model that incorporated ML techniques, including a multiple logistic regression classifier, univariate feature ranking, and principal component analysis (PCA). This model aimed to identify patients at heightened risk for CVD. The findings highlighted hypercholesterolemia, plaque scores, and the duration of diabetes as the most critical clinical indicators. Employing the ML approach, the study was able to effectively distinguish 85.11% of patients at high risk and 79.17% of those at low risk for CVD, achieving an Area Under the Curve (AUC) of 0.87. This underscores the efficacy of ML tools in identifying high-risk MAFLD patients for CVD using basic patient data. In research by Alotalibi et al. [37] the aim was to explore the effectiveness of various machine learning (ML) methodologies in forecasting heart failure incidents. Utilizing patient data sourced from the Cleveland Clinic Foundation, the research applied a variety of ML algorithms, including decision tree, logistic regression, random forest, naive Bayes, and support vector machine (SVM). Predictive models were developed using a rigorous 10-fold cross-validation approach. Among these, the decision tree algorithm was identified as the top performer, achieving a remarkable prediction accuracy of 93.19%, closely followed by SVM with an accuracy of 92.30%. This study highlights the significant potential of ML strategies in predicting heart failure, particularly underlining the decision tree algorithm's efficacy, making it a prime candidate for further research endeavors. Comparing various algorithms, Hasan and Bao [38] undertook a research project focused on identifying the most efficient feature selection technique for forecasting cardiovascular diseases. The study initially evaluated three well-known feature selection strategies—filter, wrapper, and embedded methods—and generated feature subsets using a standard "True" condition in a Boolean framework. This dual-phase selection procedure was then applied across various models such as random forest, support vector classifier (SVC), k-nearest neighbors, naive Bayes, and XGBoost, to assess their predictive accuracy and establish the best performing model. The study used the artificial neural network (ANN) as a reference point for these comparisons. Results from the investigation highlighted that the XGBoost classifier, combined with the wrapper feature selection method, was the most effective in predicting cardiovascular diseases. XGBoost recorded a 73.74% accuracy rate, closely trailed by the SVC with 73.18% and the ANN with 73.20%.

Research in the realm of cardiovascular diseases utilizing deep learning techniques has yielded promising results. Mohan et al. [39] developed a Multi-Task Deep and Wide Neural Network (MT-DWNN) designed for simultaneous multiple tasks, aiming to predict critical incidents during hospitalizations. This algorithm was evaluated using an extensive dataset covering 18 years, encompassing 35,101 instances of hospital admissions due to heart failure and 2,478 cases of renal failure at the Chinese PLA General Hospital. The MT-DWNN's ability to forecast renal complications (with an AUC of 0.9393) outperformed conventional approaches, surpassing the AUC scores of standalone deep neural networks (0.9370), the random forest model (0.9360), and logistic regression (which scored below 0.9233). The results of these experiments indicate that the MT-DWNN is highly effective in predicting renal issues in heart failure patients. In a separate study, Arslan and Karhan [40] introduced a duo of sophisticated deep neural networks, specifically aimed at accurately predicting the risk of coronary heart disease. Common prediction models often struggle with the inconsistencies typical in many real-world datasets. To overcome this, the researchers proposed a unique approach for assembling training data by dividing the original dataset into two parts: one with a general distribution and the other with a significant bias. They employed a two-phase data processing technique, wherein variable autoencoders initially split the training data into these two distinct categories. Following this, two separate deep neural network classifiers are trained on these datasets. The efficacy of this method was evident, as it achieved an AUC of 0.882 and an accuracy rate of 0.892, outshining conventional methods in several key metrics, including specificity, accuracy, precision, recall, and the F-measure (0.915).

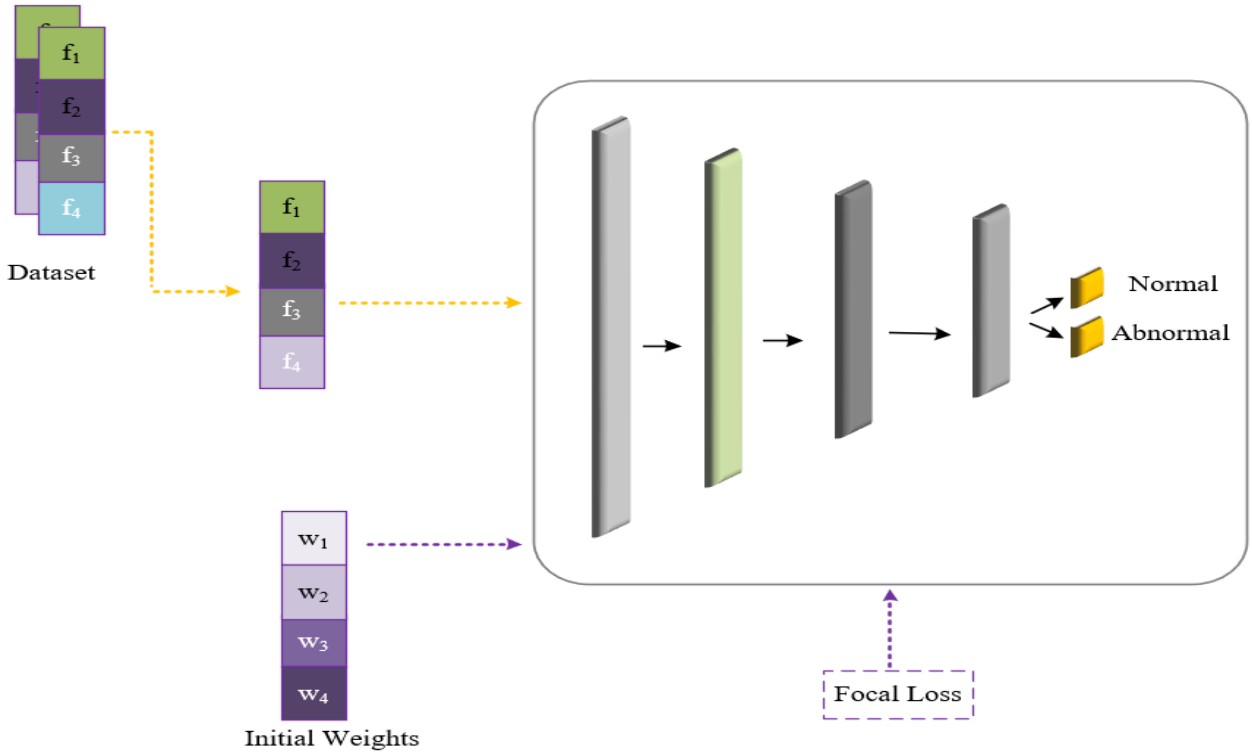


Fig. 1. Overview of the MLP-FL-CRD model.

III. Proposed Method

The structure of the MLP-FL-CRD model is shown in Fig. 1. MLP-FL-CRD is specifically engineered to augment the diagnosis of cardiovascular diseases, addressing key challenges such as imbalanced class distribution and the essential need for accurate initial weight determination. The integration of ML-ABC and RL within the framework effectively tackles these pivotal issues where conventional models often falter.

Traditional algorithms typically lack a structured approach for selecting initial weights, potentially impeding the learning process. This can lead to slower rates of convergence and the possibility of converging on less than optimal minima. Furthermore, prevalent models struggle with the issue of class imbalance, a common obstacle in cardiovascular diagnosis where significant events are infrequent. Such models usually exhibit bias towards the majority class, leading to the critical underrepresentation of minority classes. These minority classes are particularly vital to identify accurately in cardiovascular diagnostics. The proposed method, utilizing ML-ABC, introduces a thoughtfully curated and varied assortment of initial weights. This diversity aids the model in avoiding local minima and facilitates more effective convergence towards a comprehensive global solution.

Additionally, the RL aspect of the model is meticulously designed to provide greater rewards for accurately classifying the minority class, thus recalibrating the model's focus towards these crucial predictions. This marks a significant enhancement over traditional supervised learning techniques, which might lack adequate representative data for effective training across diverse classes. The flexible learning policy of RL fosters a more equitable exploration of the decision-making space. This leads to the formulation of strategies that prioritize the precise classification of lesser-represented classes. The adaptive capacity of RL within the model distinctly differentiates it from existing methodologies, rendering it capable of surmounting the intrinsic challenges typically encountered in conventional classification models, particularly in the realm of cardiovascular diagnostics.

A. Artificial Bee Colony Method

ABC [10] is a sophisticated optimization technique that simulates the foraging behavior of honey bees. Central to the ABC algorithm are four key components: employed bees, onlookers, scouts, and food sources. Employed bees are responsible for exploring nearby areas around an initial food source. Once they gather information, they return to the hive and share this knowledge with onlooker bees [41]. These onlooker bees, in turn, evaluate the potential of the food sources based on the information conveyed, including the likelihood of finding nectar. The decision-making process of onlooker bees involves assessing the probability of a food source's availability and its richness in nectar. Should a particular food source become depleted or no longer viable, the employed bee associated with that source undergoes a transformation. This bee becomes a scout, embarking on a random search for new and potentially more lucrative food sources [30]. This aspect of the algorithm exemplifies a dynamic optimization process, mirroring the adaptive and efficient foraging strategies of real-world honey bees. This bee-inspired algorithm is particularly effective in solving complex optimization problems due to its ability to explore and exploit resources. The ABC algorithm's balanced approach to exploration (via scouts) and exploitation (through employed

and onlooker bees) ensures a comprehensive search of the solution space, avoiding premature convergence on suboptimal solutions. This makes it an ideal choice for applications in various fields requiring robust optimization solutions, including data mining, engineering, and, as demonstrated in this study, healthcare analytics.

Eq. (1) presents a methodology for recalculating the position of an employed bee within the context of the ABC algorithm. This reformation of position is contingent upon the nectar quality associated with the new potential location. If the new position offers a higher nectar quality compared to the previous location, the bee is programmed to remember this new position and discard the previous one. This decision-making process is based on the principle of seeking the most rewarding nectar source, which in algorithmic terms translates to finding a more optimal solution. Conversely, if the nectar quality at the new position does not surpass that of the original location, the bee retains its prior position. This aspect of the algorithm ensures that beneficial positions are not forsaken for lesser or equivalent ones, thereby optimizing the search process. The emphasis on comparing and selecting positions based on nectar quality (or solution fitness) mirrors natural foraging behaviors and is a key factor in the ABC algorithm's ability to effectively navigate and exploit the solution space. This rule of position updating based on nectar quality highlights the algorithm's capacity for adaptive learning. It allows the algorithm to dynamically adjust its search strategy based on the evolving understanding of the solution landscape, enhancing its efficiency in locating and converging on optimal or near-optimal solutions.

$$v_i^j = x_i^j + \varphi_i^j (x_i^j - x_k^j) \tag{1}$$

where, j is the $j-$th place, and each answer x_i contains a size of D. D represents the principles to be enhanced, while k represents an arbitrary answer $(k \neq i)$. φ_i^j shows a number randomly selected on the scale of $[0,1]$. Modifying a single element of x_i, the potentially novel solution v_i can be realized.

In the framework of a D-dimensional optimization process, a key strategy involves selectively altering the value of a randomly chosen dimension. Following each iteration, the selection of an improved solution is based on its 'athleticism' worth, a term used to signify its fitness or suitability within the context of the problem being solved. According to Formula 1, the newly generated solution v_i^j is dependent primarily on x_i^j and x_k^j, which ensures that the new food source v_i^j remains unpredictable and dynamic. This element of variability is crucial in preventing the algorithm from stagnating at local optima and aids in exploring a broader range of potential solutions. The study's approach in considering the concept of 'athleticism' is drawn from a collective understanding derived from both current and nearby nutrition resources. By utilizing shared knowledge and insights gathered from these resources, the algorithm continuously seeks to develop a food source with a higher value. This method is in alignment with the principles of the ABC algorithm, where the objective is to discover and exploit resources that possess superior 'athleticism' worth, or in algorithmic terms, higher fitness values. Such an approach not only enhances the diversity of the solutions explored by the algorithm but also ensures that the process of optimization is dynamic and adaptive. By continuously updating the solution based on shared knowledge and the comparative worth of nearby resources, the algorithm can effectively navigate the solution space, moving towards more promising areas while avoiding less fruitful ones. This dynamic nature of solution generation and selection is integral to the success of the ABC algorithm in solving complex, multi-dimensional optimization problems.

$$v_i^j = \begin{cases} x_i^j + \varphi_i^j (x_k^j - x_i^j), & Fit_i < Fit_k \\ x_k^j + \varphi_i^j (x_i^j - x_k^j), & Fit_i \geq Fit_k \end{cases} \tag{2}$$

where, Fit_i and Fit_k represent the fitness values or the athleticism worth of the nearby and current food resources, respectively. The parameter φ_i^j is defined as a uniformly distributed random number within the range [0, F], where F, a positive value, is known as the mutual learning factor. This factor plays a crucial role in the algorithm by guiding the fitness values of newly generated solutions towards superior food sources. It does this by evaluating and contrasting the current food resources with those in proximity. The candidate solution undergoes modification depending on the quality of the current food sources. If the existing food sources are found to be sufficiently rewarding, the solution will be further refined based on these sources. Conversely, if the current sources are deemed inadequate, the solution will shift towards the nearby, potentially more promising food source. This dynamic allows for a balanced approach between exploration of new possibilities and exploitation of known good solutions. The mutual learning factor F is pivotal in regulating the extent of perturbation between the positions of different food sources. A non-negative value of F is crucial to ensure that the resultant changes lead to an improved solution. As F increases from zero, the impact of perturbation on the corresponding food source diminishes, implying that the fitness value of the alternative resource is almost equal to that of the current, superior resource. Conversely, a higher value of F can reduce the algorithm's ability to effectively explore and exploit, as it lessens the impact of contrasting different food resources. Hence, the choice of a suitable value for F is vital in preserving a balanced interplay between exploration and exploitation. This balance is key to the ABC algorithm's ability to effectively navigate and identify optimal or near-optimal solutions in intricate optimization scenarios. The proper calibration of F ensures that the algorithm is neither overly explorative, risking inefficiency, nor excessively exploitative, which might lead to premature convergence. This careful tuning is essential for the algorithm's success in addressing complex optimization challenges efficiently.

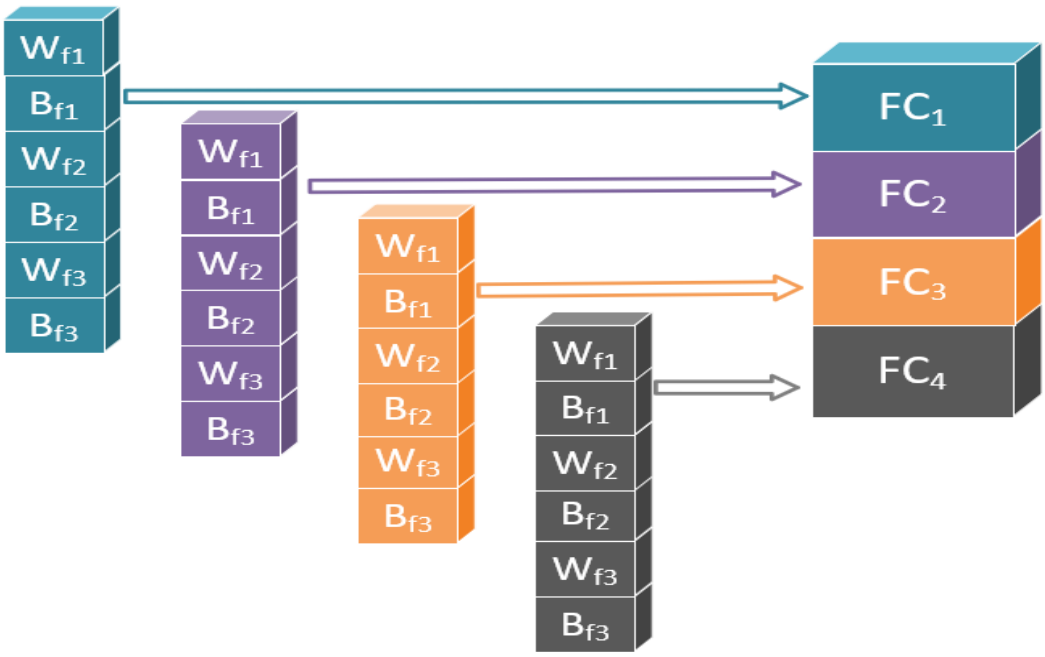


Fig. 2. Encoding approach in the offered method.

B. Model Architecture

1) Pre-Training: In this particular scenario, the management of the MLP's weights is executed through an ABC algorithm, which is enhanced by the principles of mutual learning. The encoding procedure is a critical step that involves arranging these weights into a vector, a process that symbolically represents the positioning of bees within the ABC paradigm [42, 43]. This strategic alignment of weights into a vector format is not just a mere arrangement; it is conceptualized as mirroring the bees' placements in the ABC algorithm, thus establishing a direct correlation between the weights' configuration and the algorithm's operational mechanics. Achieving the optimal configuration for this encoding is a complex task that demands meticulous attention to detail. Despite the inherent challenges, a series of methodical experiments were conducted to ascertain the most effective encoding strategy, ensuring that the weights are optimally aligned for the ABC algorithm's functioning. This experimental approach was crucial in refining the encoding process, thereby enhancing the overall efficiency of the algorithm. As illustrated in Fig. 2, the comprehensive approach to this encoding process is evident. Here, not just the weights but also the bias terms are meticulously arranged into a vector. This arrangement is more than a mere collection of numerical values; it forms the foundation of a candidate solution within the ambit of the proposed ABC algorithm. This candidate solution, representative of the bees' locations in the ABC model, is instrumental in the algorithm's problem-solving process, highlighting the synergy between the MLP's weight management and the ABC algorithm's operational framework.

For assessing the caliber of a candidate solution, the fitness function is delineated as

$$Fitness = \frac{1}{\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2} \quad (3)$$

where, N is the training examples number, with $\tilde{y}_i$ and y_i indicating the i -th sample-estimated and goal outcome, respectively.

2) Focal loss: In this research, the detection problem is defined within the framework of binary classification, where instances are divided into positive and negative classes. A significant challenge faced is the imbalance in the dataset, particularly manifested in the limited number of samples representing the negative class. This imbalance can adversely affect the model's ability to learn from the underrepresented class. To address this issue, the study employs focal loss (FL) [44], a modified version of the binary cross-entropy (CE) loss function. FL is specifically engineered to refocus the training process on the more complex, less represented samples, which are often crucial for accurate classification but are overshadowed in imbalanced datasets [45]. FL achieves this by adding a modulating factor to the traditional cross-entropy loss, which adjusts the contribution of each sample to the loss based on the ease or difficulty of classifying it. This ensures that the model does not become biased towards the majority class and pays adequate attention to the minority class samples, which are pivotal for a balanced and comprehensive learning process. Such an approach is particularly beneficial in scenarios where the minority class is not just numerically fewer but also qualitatively more challenging to predict. By effectively targeting these challenging samples, FL aids in enhancing the overall robustness and accuracy of the model in

dealing with binary classification tasks under imbalanced conditions.

CE is conceptualized as:

$$CE = \begin{cases} -log\ (p), & y = 1 \\ -log(1-p), & \text{otherwise} \end{cases} \quad (4)$$

where, $y \in \{-1,1\}$ represents the actual class label, while $p \in [0,1]$ denotes the model's predicted probability for the class that corresponds to the label $y = 1$. The predicted probability is

$$p_t = \begin{cases} -p, & y = 1 \\ 1-p, & \text{otherwise} \end{cases} \quad (5)$$

Hence,

$$CE\ (p,y) = CE\ (p_t) = -\log(p_t) \quad (6)$$

FL introduces a modulating factor to the standard cross-entropy loss, resulting in the following modification:

$$FL\ (p_t) = -\alpha_t(1-p_t)^{\gamma}\ \log(p_t) \quad (7)$$

where, $\gamma > 0$ is a positive parameter (notably, when γ=1, FL resembles the CE loss), and α lies within the range of 0 to 1, representing the inverse of the class frequency.

IV. Experimental Results

In 2012, healthcare practitioners at the Polyclinic for Occupational Health and Sports in Zagreb assembled a comprehensive dataset from 26,002 medical examinations. These examinations were conducted on athletes seeking medical clearance for participation in competitive sports. The collected data encompassed a wide range of vital health parameters for each athlete, including their sex, age, height, weight, resting pulse rate, as well as both diastolic and systolic blood pressure measurements. Additionally, resting electrocardiogram (ECG) data were meticulously recorded for each individual. The dataset revealed a significant imbalance in the classification of results, with a dominant majority, 91.2%, being categorized as N (negative), indicating no apparent risk or health concern that would preclude sports participation. Conversely, a smaller fraction, 8.8%, fell into the P (positive) category, suggesting potential health risks or conditions that required further medical evaluation and could potentially restrict the athlete's ability to engage in competitive sports. This disproportionate distribution between the N and P classifications underscores the challenges faced in medical diagnostics, particularly in accurately identifying and diagnosing conditions in smaller, potentially high-risk groups. The dataset, therefore, provides an invaluable resource for developing and testing medical diagnostic models that can effectively manage such imbalances, ensuring that high-risk cases are accurately identified and addressed, despite being numerically fewer in the dataset. This approach is crucial in enhancing the safety and health management of athletes, aligning with the overarching goal of ensuring their well-being and fitness for competitive sports participation.

The model put forth operates on a 64-bit Windows OS, supported by a robust 64 GB of RAM, and is further enhanced by a graphics processing unit (GPU) with a capacity of 64 GB. Table I presents the hyperparameters applied to the MLP-FL-CRD model.

The MLP-FL-CRD model is first trained and tested on the introduced dataset in parallel with 6 computer learning samples (SVM [46], Naïve Bayes [47], KNN [48], Random forests [49], Logistic Regression [50], and Decision tree [51]) and two smaller parts of the proposed model, i.e., Proposed+ random weights (which possesses a base architecture similar to the model but uses random weights instead for initialization), and Proposed+ random weights+ FL (which uses FL for classification). Table II shows the parameters applied to machine learning models.

TABLE I. Parameter Setting for the MLP-FL-CRD Model

Parameter	Value
Epoch	256
Batch size	128
Learning rate	0.03
Discount factor	0.4

TABLE II. Parameter Setting for Machine Learning Models

Algorithm	Parameter	Value
Naïve Bayes	α (Lidstone smoothing parameter)	0.5
KNN	k (Number of neighbors)	5
	Distance Metric	Euclidean (p=2), Manhattan (p=1)
SVM	Kernel	polynomial
	γ (Kernel coefficient)	0.5
Random Forests	Number of trees	20
	Max depth of tree	10
Logistic Regression	C (Inverse regularization strength)	0.3
	Solver	liblinear
Decision Tree	Criterion	entropy
	Max depth	10

The results are compared using standard performance metrics (see Table III), of which F-measure and geometric mean are the preferred metrics for imbalanced data [52]. The MLP-FL-CRD model outperforms other models, including the closest competitor, the Decision Tree, across all evaluation metrics. Specifically, this model achieves a reduction in error exceeding 59% and 34% for two primary metrics, namely the F-measure and G-means, respectively. A comparative analysis of the MLP-FL-CRD model against variants such as Proposed+random weights and Proposed+random weights+FL reveals a notable decrease in error rates, approximately 70%. This significant reduction underscores the effectiveness of the enhanced ABC and FL methodologies.

A. Impact of other Metaheuristics

The next experiment involved comparing the improved ABC algorithm with various metaheuristic optimization algorithms. In this experiment, the initial model parameters were obtained using different metaheuristics while retaining the other model components. Six algorithms were tested, including standard ABC [53], FA [54], BA [55], COA [56], DE [28], and GWO [57]. The default configurations are detailed in Table IV. The results obtained are presented in Table V. The findings indicated that the suggested ABC algorithm reduced the error by approximately 48% compared to the standard ABC algorithm. This result showed that the proposed model outperformed the standard one. Furthermore, the ABC algorithm delivered better results than others, including DE, GWO, and BA.

TABLE III. OUTCOMES OF DIVERSE CATEGORIZATION METHODS

	accuracy	recall	precision	F-measure	G-means
SVM	0.687± 0.051	0.594± 0.068	0.540± 0.125	0.566± 0.068	0.661± 0.005
Naïve Bayes	0.820± 0.159	0.743± 0.004	0.772± 0.105	0.757± 0.109	0.806± 0.045
KNN	0.790± 0.104	0.667± 0.005	0.704± 0.048	0.685± 0.215	0.754± 0.006
Random forests	0.695± 0.105	0.569± 0.226	0.554± 0.103	0.561± 0.100	0.658± 0.256
Logistic Regression	0.818± 0.145	0.781± 0.048	0.715± 0.106	0.747± 0.148	0.809± 0.125
Decision tree	0.844± 0.146	0.825± 0.105	0.765± 0.275	0.804± 0.205	0.830± 0.014
Proposed + random weights	0.800± 0.052	0.810± 0.100	0.791± 0.000	0.781± 0.152	0.820± 0.252
Proposed + random weights+FL	0.851± 0.014	0.860± 0.115	0.840± 0.259	0.840± 0.041	0.850± 0.082
MLP-FL-CRD	0.878± 0.028	0.892± 0.103	0.870± 0.021	0.871± 0.019	0.892± 0.032

TABLE IV. PARAMETER SETTING FOR METAHEURISTIC MODELS

Algorithm	Parameter	Value
DE	scaling ratio	0.6
	probability of crossover	0.8
ABC	limit	n_e × dimensionality
	n_o	50% of the colony
	n_e	50% of the colony
	n_s	1
FA	coefficient of light absorption	1
	initial attractiveness at r = 0	0.3
	proportional factor	0.4
BA	update constant for loudness	0.45
	update constant for rate of emission	0.45
	initial rate of pulse emission	0.002
COA	rate of alien solution discovery	0.30
GWO	no parameters	-

TABLE V. RESULTS OF VARIOUS METAHEURISTIC METHODS

	accuracy	recall	precision	F-measure	G-means
ABC	0.852± 0.148	0.840± 0.123	0.859± 0.051	0.841± 0.009	0.821± 0.412
DE	0.841± 0.158	0.831± 0.103	0.842± 0.251	0.830± 0.015	0.800± 0.025
FA	0.827± 0.015	0.815± 0.123	0.810± 0.261	0.819± 0.071	0.780± 0.014
BA	0.810± 0.032	0.800± 0.014	0.801± 0.071	0.805± 0.132	0.762± 0.125
COA	0.792± 0.123	0.772± 0.016	0.786± 0.274	0.772± 0.171	0.741± 0.156
GWO	0.740± 0.152	0.724± 0.015	0.740± 0.012	0.731± 0.223	0.690± 0.126

B. Impact of parameter F on the model

The performance of the suggested approach is heavily influenced by the mutual learning factor *F*, as expressed in Eq. (2). When *F* is set too low, the algorithm may not be able to take full advantage of the mutual learning process, and thus, the performance of the algorithm may suffer. However, if *F* is excessively increased, the algorithm may become too ambitious in its learning, resulting in overfitting and poor generalization. Therefore, it is essential to find the optimal value of *F* those balances between the benefits of mutual learning and the risks of overfitting. As shown in Figure 3, the performance of the algorithm improves significantly as *F* increases from 0.5 to 2.5. This is because a higher value of *F* allows the models to exchange more information, leading to better generalization and higher accuracy. However, when *F* is increased beyond 2.5, the performance of the algorithm starts to deteriorate. This is because the models become too aggressive in their learning and start to overfit the data. Overall, these results demonstrate that the mutual learning factor *F* is a critical parameter in the proposed approach, and its value should be carefully chosen to achieve optimal performance. A moderate value of *F* between 1.5 and 2.5 may be a good starting point, and the optimal value can be determined through experimentation and cross-validation.

C. Exploring the Number of MLP Layers

The article highlights that as the number of layers in a multi-layer perceptron (MLP) increase, the model's complexity increases, leading to a higher risk of overfitting. Conversely, having too few layers may limit the model's ability to represent essential features in the training data. In the proposed approach, six different values (1, 2, 4, 8, 10, 12) are tested as the number of layers in MLP to study its effect on the model's performance. Table VI displays the achieved results, which show a descending trend for the number of layers from 1 to 4 and an ascending trend for values from 4 to 12. This finding suggests that having four layers in MLP is the optimal value for achieving the best results.

D. Impact of Loss Function

There are several methods to address data imbalances in machine learning models, including arranging info-augmenting strategies and the choice of the dropping operation or LF. Along with the methods, the choice of the dropping operation or LF is particularly critical since it can help the model learn from the minority class. To test the effectiveness of different loss functions, five functions were chosen, including weighted cross-entropy (WCE) [58], balanced cross-entropy (BCE)] [59], Dice loss (DL) [60], Tversky loss (TL) [61], and Combo Loss (CL) [62]. The BCE and WCE loss functions are commonly used to treat both positive and negative examples equally. However, these loss functions may not be suitable for imbalanced datasets where the minority class needs to be emphasized. The DL and TL loss functions are more suitable for imbalanced datasets, as they perform better on the minority class. The CL function is a promising loss function that can benefit applications using unbalanced data. The CL method can decrease the importance of straightforward examples and emphasize learning intricate instances by modifying the weights of the loss function. To evaluate the effectiveness of these loss functions, experiments were conducted and the results are reported in Table VII. According to the findings, the performance of the CL function was better than that of the TL function. It reduced the error rate by 39% and 25%for the F-measure and accuracy metrics in order. However, the CL function operates 60% worse than the FL, which is a specialized loss function for binary classification tasks.

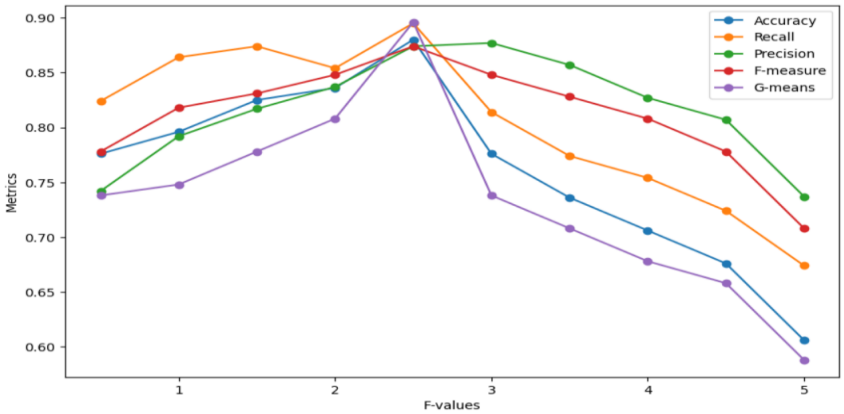


Fig. 3. The performance metrics of the MLP-FL-CRD model in relation to the *F* value of the proposed model.

TABLE VI. THE DIVERSE NUMBER OF MLP LAYERS VS. THE EFFICIENCY MEASUREMENTS PLOTTED

Number of layers	accuracy	recall	precision	F-measure	G-means
1	0.750± 0.019	0.772± 0.203	0.731± 0.025	0.732± 0.034	0.763± 0.102
2	0.862± 0.008	0.870± 0.153	0.852± 0.005	0.857± 0.029	0.871± 0.152
4	0.880± 0.028	0.895± 0.103	0.874± 0.021	0.874± 0.019	0.896± 0.032
8	0.842± 0.018	0.851± 0.101	0.840± 0.221	0.820± 0.010	0.852± 0.022
10	0.704± 0.125	0.720± 0.111	0.682± 0.201	0.700± 0.015	0.650± 0.032
12	0.480± 0.055	0.592± 0.021	0.490± 0.221	0.526± 0.002	0.420± 0.152

TABLE VII. OUTCOMES OF DIFFERENT LFS

	accuracy	recall	precision	F-measure	G-means
WCE	0.750± 0.026	0.742± 0.010	0.731± 0.119	0.736± 0.002	0.762± 0.032
BCE	0.801± 0.021	0.793± 0.045	0.772± 0.156	0.770± 0.012	0.811± 0.001
DL	0.812± 0.033	0.801± 0.025	0.780± 0.026	0.798± 0.006	0.820± 0.000
TL	0.821± 0.123	0.824± 0.012	0.800± 0.016	0.813± 0.036	0.840± 0.062
CL	0.861± 0.003	0.852± 0.215	0.840± 0.006	0.840± 0.042	0.863± 0.152

E. Discussion

This article proposed a novel approach to identifying athletes at risk of developing CVDs by utilizing a multi-layer perceptron model enhanced with focal loss and a ML-ABC optimization technique. This model specifically addresses the challenge of class imbalance in datasets through focal loss, which effectively de-emphasizes simpler cases to focus on more complex ones, thus improving diagnostic accuracy in scenarios with significant class disparities. Additionally, the article introduces the ML-ABC method as a solution to overcome the limitations of traditional gradient-based learning methods, such as sensitivity to initial weight settings. By adjusting candidate food sources based on mutual learning factors and considering initial weights, the model shows improved performance. The effectiveness of this approach is validated through extensive experiments and ablation studies, demonstrating its superiority over other models in accurately assessing the risk of CVDs among athletes.

The susceptibility of the MLP-FL-CRD model to class imbalance, even with the integration of FL, presents a notable constraint in accurately diagnosing cardiovascular risk among athletes. In the referenced 2012 dataset, which included data from 26,002 athletes, there was a pronounced imbalance, predominantly skewed towards negative outcomes. While the implementation of focal loss aims to accentuate the learning from minority class examples, the preponderance of negative results can still hinder the model's ability to generalize across varied scenarios effectively. This imbalance in the dataset could induce a predictive bias within the model, potentially leading to an underestimation of risk in cases that are actually positive but less frequent. Such a bias is particularly concerning in a medical context, where failing to identify at-risk individuals can have serious, if not fatal, consequences. The model's overexposure to negative outcomes might condition it to lean towards these predictions, potentially missing out on identifying athletes who are genuinely at risk of cardiovascular issues. To address this limitation, additional strategies may need to be considered. One approach could involve incorporating more sophisticated balancing techniques that go beyond focal loss, such as synthetic data generation methods like SMOTE (Synthetic Minority Over-sampling Technique) [63] or adaptive resampling [64]. These methods can help in creating a more balanced training environment for the model, thus enhancing its capacity to learn from both majority and minority classes more effectively [65]. Another potential solution lies in expanding and diversifying the dataset. Gathering more comprehensive data that includes a wider range of cardiovascular conditions and outcomes could help in creating a more representative dataset. This expanded dataset would not only provide a broader spectrum of cases for the model to learn from but also reduce the likelihood of predictive bias, thereby improving the model's accuracy and reliability in real-world applications.

The dependency of the model's performance on the initial weight settings is a critical aspect, particularly prevalent in gradient-based learning methods such as backpropagation. While the model employs a ML-ABC methodology to mitigate this issue by dynamically adjusting the primary weights according to fitness, the inherent reliance on initial weights remains a significant vulnerability. Inadequate calibration of these initial weights could lead to the model converging towards suboptimal solutions, thereby affecting its overall effectiveness and accuracy. The initial weight setting acts as the starting point for the learning process and heavily influences the trajectory of the model’s convergence. If these weights are not set in a manner that reflects the complexity and nuances of the data, the model may find itself trapped in local minima, or on a prolonged path towards the global optimum. This challenge is further amplified in the context of complex and high-dimensional data typically encountered in cardiovascular risk assessment. To enhance the resilience of the model against potential pitfalls associated with initial weight selection, exploring alternative strategies for initialization is imperative. One such approach could involve the use of advanced heuristics or algorithms that analyze the data distribution to determine more effective starting weights. Techniques like Xavier or He initialization [66], which consider the size of the network layers in setting the initial weights, could offer more reliable starting points for the model's training. Additionally, incorporating more robust weight optimization methods could further strengthen the model. Techniques like stochastic gradient descent with momentum or adaptive learning rate algorithms like Adam could provide more nuanced adjustments during the training process. These methods help in navigating the weight space more effectively, increasing the likelihood of the model finding a more optimal solution.

The success of the model in the specific context of the study dataset focusing on athletes' cardiovascular risk assessment does not automatically translate to its efficacy in other scenarios or datasets. This limitation in generalizability and adaptability poses a significant challenge, especially considering the diverse nature of CVDs and the varying characteristics of different patient populations. The model's parameters, which have been fine-tuned for this particular dataset, may not be directly applicable or optimal for other datasets that differ in demographics, prevalence of CVD types, or other clinical factors. When transitioning the model to different populations or conditions, it may encounter data that significantly deviates from the characteristics of the original dataset. This deviation can result in reduced accuracy and reliability, as the model’s learned patterns and parameters may

not align well with the new data. Consequently, substantial re-tuning of the model's parameters becomes necessary, along with rigorous validation processes to ensure its efficacy in the new context. This re-tuning process can be resource-intensive and time-consuming, requiring a thorough understanding of the new dataset's attributes and underlying distributions. To address these challenges, developing more flexible learning algorithms that can easily adapt to varying data characteristics is essential. Such algorithms should be capable of identifying and adjusting to the nuances of different datasets without extensive manual intervention. This flexibility can be achieved through approaches like meta-learning, where the model learns to quickly adapt to new tasks using only a small amount of data, or through the development of models that are inherently more robust to changes in data distribution. Moreover, incorporating transfer learning techniques can significantly enhance the adaptability of the model. Transfer learning allows a model trained on one task to apply its learned knowledge to a different but related task. By leveraging pre-trained models or transferring knowledge from the original dataset to new ones, the model can achieve better performance with less need for re-tuning. This approach can be particularly beneficial in medical applications like CVD risk assessment, where similarities exist across different datasets, even though they may vary in specific characteristics.

V. Conclusion

CVDs remain a significant global health issue, and identifying individuals at risk of developing CVDs is crucial for early intervention and effective treatment. To this end, the article presents a model based on a multi-layer perceptron that employs focal loss to identify athletes who may be at risk of developing CVDs. Focal loss is a useful technique that reduces the significance of straightforward instances, enabling the model to concentrate on more difficult instances. This technique is particularly useful when dealing with a large disparity between classes. Usually, the training process of the model relies on gradient-based learning methods like backpropagation. These techniques have several limitations, including initialization sensitivity. The article suggests a solution for this problem, which involves utilizing ML-ABC. This approach involves modifying the candidate food source with higher fitness between two individuals using a mutual learning factor. This method takes into account the initial weights of the model and can improve its performance. The offered sample performs better than the rest of the samples, achieving excellent results. Experiments conducted on the study dataset help to determine the ideal values of the critical parameters in the model. Ablation studies further confirm the positive impact of the proposed components on model performance.

Despite these promising results, the article acknowledges the need for further research to test the efficacy of this model in non-athletic and older populations. Additionally, it is necessary to determine an appropriate classification performance measure that balances medical risk and welfare. This research is crucial to ensure that the proposed data mining methods can be effectively applied to improve healthcare policies and reduce unnecessary examinations. The potential impact of this research is significant. By identifying individuals at risk of developing CVD early, healthcare professionals can intervene with targeted prevention strategies and treatments. This could ultimately lead to improved health outcomes for individuals and reduced healthcare costs for society.

Furthermore, the methodologies integrated into this model hold potential for application beyond the realm of sports medicine, offering opportunities to enhance disease detection and management across various healthcare sectors. The adaptability of these techniques could play a pivotal role in refining diagnostic processes and treatment strategies for a range of medical conditions, thereby contributing to the overall advancement of healthcare practices. To summarize, the research detailed in this article introduces a robust model adept at identifying athletes who are at an increased risk of developing CVDs. Its effectiveness, as demonstrated in the specific context of sports healthcare, offers valuable insights into CVD risk assessment. However, it is imperative to extend this research to encompass a more diverse range of populations. Such expansion is essential to fully ascertain the model's efficacy across different demographic and health profiles. Developing and fine-tuning classification performance metrics tailored to varied populations will also be crucial in this regard. While the model shows considerable promise, its broader applicability and effectiveness in general healthcare settings remain areas for future investigation. Pursuing this line of research is vital for realizing the model’s full potential in enhancing early detection and intervention strategies for CVDs. Ultimately, the advancements in early identification and treatment strategies for CVDs, as suggested by this research, could lead to improved patient health outcomes and a reduction in healthcare costs. This underscores the significance of this research as a meaningful contribution to the field of medical diagnostics and patient care.

References

[1] D. Barbieri et al., "Predicting cardiovascular risk in Athletes: Resampling improves classification performance," International Journal of Environmental Research and Public Health, vol. 17, no. 21, p. 7923, 2020.

[2] I. Wronka, E. Suliga, and R. Pawliñska-Chmara, "Evaluation of lifestyle of underweight, normal weight and overweight young women," Collegium antropologicum, vol. 37, no. 2, pp. 359-365, 2013.

[3] W.-K. Wong, A. W. Moore, G. F. Cooper, and M. M. Wagner, "Bayesian network anomaly pattern detection for disease outbreaks," in Proceedings of the 20th International Conference on Machine Learning (ICML-03), 2003, pp. 808-815.

[4] C. Campbell and K. Bennett, "A linear programming approach to novelty detection," Advances in neural information processing systems, vol. 13, 2000.

[5] J.-F. Fontaine, J. Priller, E. Spruth, C. Perez-Iratxeta, and M. A. Andrade-Navarro, "Assessment of curated phenotype mining in neuropsychiatric disorder literature," Methods, vol. 74, pp. 90-96, 2015.

[6] A. Salam and J. A. McGrath, "Diagnosis by numbers: defining skin disease pathogenesis through collated gene signatures," Journal of Investigative Dermatology, vol. 135, no. 1, pp. 17-19, 2015.

[7] L. Sacchi, A. Tucker, S. Counsell, D. Garway-Heath, and S. Swift, "Improving predictive models of glaucoma severity by incorporating quality indicators," Artificial intelligence in medicine, vol. 60, no. 2, pp. 103-112, 2014.

[8] F. Masood, J. Masood, H. Zahir, K. Driss, N. Mehmood, and H. Farooq, "Novel approach to evaluate classification algorithms and feature selection filter algorithms using medical data," Journal of Computational and Cognitive Engineering, vol. 2, no. 1, pp. 57-67, 2023.

[9] E. H. Houssein, M. E. Hosney, W. M. Mohamed, A. A. Ali, and E. M. Younis, "Fuzzy-based hunger games search algorithm for global optimization and feature selection using medical data," Neural Computing and Applications, vol. 35, no. 7, pp. 5251-5275, 2023.

[10] S. V. Moravvej et al., "RLMD-PA: A reinforcement learning-based myocarditis diagnosis combined with a population-based algorithm for pretraining weights," Contrast Media & Molecular Imaging, vol. 2022, 2022.

[11] W. Wang et al., "Medical image classification using deep learning," Deep learning in healthcare: paradigms and applications, pp. 33-51, 2020.

[12] S. V. Moravvej, S. J. Mousavirad, D. Oliva, G. Schaefer, and Z. Sobhaninia, "An Improved DE Algorithm to Optimise the Learning Process of a BERT-based Plagiarism Detection Model," in 2022 IEEE Congress on Evolutionary Computation (CEC), 2022: IEEE, pp. 1-7.

[13] S. Zhang, C. Tjortjis, X. Zeng, H. Qiao, I. Buchan, and J. Keane, "Comparing Data Mining Methods with Logistic Regression."

[14] H. Zareiamand, A. Darroudi, I. Mohammadi, S. V. Moravvej, S. Danaei, and R. Alizadehsani, "Cardiac Magnetic Resonance Imaging (CMRI) Applications in Patients with Chest Pain in the Emergency Department: A Narrative Review," Diagnostics, vol. 13, no. 16, p. 2667, 2023.

[15] M. Bahadori, M. Soltani, M. Soleimani, and M. Bahadori, "Statistical Modeling in Healthcare: Shaping the Future of Medical Research and Healthcare Delivery," in AI and IoT-Based Technologies for Precision Medicine: IGI Global, 2023, pp. 431-446.

[16] L. Hong et al., "GAN‐LSTM‐3D: An efficient method for lung tumour 3D reconstruction enhanced by attention‐based LSTM," CAAI Transactions on Intelligence Technology, 2023.

[17] M. Soleimani, Z. Forouzanfar, M. Soltani, and M. J. Harandi, "Imbalanced Multiclass Medical Data Classification based on Learning Automata and Neural Network," EAI Endorsed Transactions on AI and Robotics, vol. 2, 2023.

[18] S. V. Moravvej, A. Mirzaei, and M. Safayani, "Biomedical text summarization using conditional generative adversarial network (CGAN)," arXiv preprint arXiv:2110.11870, 2021.

[19] Z. Ahmed and S. Das, "A Comparative Analysis on Recent Methods for Addressing Imbalance Classification," SN Computer Science, vol. 5, no. 1, pp. 1-18, 2024.

[20] S. Danaei et al., "Myocarditis Diagnosis: A Method using Mutual Learning-Based ABC and Reinforcement Learning," in 2022 IEEE 22nd International Symposium on Computational Intelligence and Informatics and 8th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics (CINTI-MACRo), 2022: IEEE, pp. 000265-000270.

[21] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5375-5384.

[22] Y. Yan, M. Chen, M.-L. Shyu, and S.-C. Chen, "Deep learning for imbalanced multimedia data classification," in 2015 IEEE international symposium on multimedia (ISM), 2015: IEEE, pp. 483-488.

[23] S. Moravvej, M. Maleki Kahaki, M. Salimi Sartakhti, and M. Joodaki, "Efficient GAN-based method for extractive summarization," Journal of Electrical and Computer Engineering Innovations (JECEI), vol. 10, no. 2, pp. 287-298, 2022.

[24] M. Marani, M. Soltani, M. Bahadori, M. Soleimani, and A. Moshayedi, "The Role of Biometric in Banking: A Review," EAI Endorsed Transactions on AI and Robotics, vol. 2, no. 1, 2023.

[25] M. S. Sartakhti, M. J. M. Kahaki, S. V. Moravvej, M. javadi Joortani, and A. Bagheri, "Persian language model based on BiLSTM model on COVID-19 corpus," in 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA), 2021: IEEE, pp. 1-5.

[26] S. V. Moravvej, S. J. Mousavirad, D. Oliva, and F. Mohammadi, "A Novel Plagiarism Detection Approach Combining BERT-based Word Embedding, Attention-based LSTMs and an Improved Differential Evolution Algorithm," arXiv preprint arXiv:2305.02374, 2023.

[27] M. Jaderberg et al., "Population based training of neural networks," arXiv preprint arXiv:1711.09846, 2017.

[28] S. J. Mousavirad, D. Oliva, S. Hinojosa, and G. Schaefer, "Differential evolution-based neural network training incorporating a centroid-based strategy and dynamic opposition-based learning," in 2021 IEEE Congress on Evolutionary Computation (CEC), 2021: IEEE, pp. 1233-1240.

[29] D. Karaboga, B. Akay, and C. Ozturk, "Artificial bee colony (ABC) optimization algorithm for training feed-forward neural networks," in Modeling Decisions for Artificial Intelligence: 4th International Conference, MDAI 2007, Kitakyushu, Japan, August 16-18, 2007. Proceedings 4, 2007: Springer, pp. 318-329.

[30] H. Gharagozlou, J. Mohammadzadeh, A. Bastanfard, and S. S. Ghidary, "RLAS-BIABC: A reinforcement learning-based answer selection using the bert model boosted by an improved ABC algorithm," Computational Intelligence and Neuroscience, vol. 2022, 2022.

[31] S. V. Moravvej, M. Joodaki, M. J. M. Kahaki, and M. S. Sartakhti, "A method based on an attention mechanism to measure the similarity of two sentences," in 2021 7th International Conference on Web Research (ICWR), 2021: IEEE, pp. 238-242.

[32] D. Waigi, D. S. Choudhary, D. P. Fulzele, and D. Mishra, "Predicting the risk of heart disease using advanced machine learning approach," Eur. J. Mol. Clin. Med, vol. 7, no. 7, pp. 1638-1645, 2020.

[33] K. Vanisree and J. Singaraju, "Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks," International Journal of computer applications, vol. 19, no. 6, pp. 6-12, 2011.

[34] A. Narin, Y. İşler, and M. Özer, "Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability," in 2016 Medical Technologies National Congress (TIPTEKNO), 2016: IEEE, pp. 1-4.

[35] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," SN Computer Science, vol. 1, pp. 1-6, 2020.

[36] K. Drożdż et al., "Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach," Cardiovascular Diabetology, vol. 21, no. 1, p. 240, 2022.

[37] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," International Journal of Advanced Computer Science and Applications, vol. 10, no. 6, 2019.

[38] N. Hasan and Y. Bao, "Comparing different feature selection algorithms for cardiovascular disease prediction," Health and Technology, vol. 11, pp. 49-62, 2021.

[39] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE access, vol. 7, pp. 81542-81554, 2019.

[40] Ö. Arslan and M. Karhan, "Effect of Hilbert-Huang transform on classification of PCG signals using machine learning," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 10, pp. 9915-9925, 2022.

[41] S. Vakilian, S. V. Moravvej, and A. Fanian, "Using the artificial bee colony (ABC) algorithm in collaboration with the fog nodes in the Internet of Things three-layer architecture," in 2021 29th Iranian Conference on Electrical Engineering (ICEE), 2021: IEEE, pp. 509-513.

[42] S. Vakilian, S. V. Moravvej, and A. Fanian, "Using the cuckoo algorithm to optimizing the response time and energy consumption cost of fog nodes by considering collaboration in the fog layer," in 2021 5th International Conference on Internet of Things and Applications (IoT), 2021: IEEE, pp. 1-5.

[43] S. V. Moravvej, M. J. M. Kahaki, M. S. Sartakhti, and A. Mirzaei, "A method based on attention mechanism using bidirectional long-short term memory (BLSTM) for question answering," in 2021 29th Iranian Conference on Electrical Engineering (ICEE), 2021: IEEE, pp. 460-464.

[44] D. Sarkar, A. Narang, and S. Rai, "Fed-focal loss for imbalanced data classification in federated learning," arXiv preprint arXiv:2011.06283, 2020.

[45] S. K. Prabhakar, H. Rajaguru, and D.-O. Won, "Performance Analysis of Hybrid Deep Learning Models with Attention Mechanism Positioning and Focal Loss for Text Classification," Scientific Programming, vol. 2021, pp. 1-12, 2021.

[46] M. A. de Almeida, "DATA MINING: DETERMINAC AO DE AGRUPAMENTOS EM GRANDES BASES DE DADOS," Universidade Federal do Rio de Janeiro, 2013.

[47] G. I. Webb, E. Keogh, and R. Miikkulainen, "Naïve Bayes," Encyclopedia of machine learning, vol. 15, pp. 713-714, 2010.

[48] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings, 2003: Springer, pp. 986-996.

[49] L. Breiman, "Random forests," Machine learning, vol. 45, pp. 5-32, 2001.

[50] M. P. LaValley, "Logistic regression," Circulation, vol. 117, no. 18, pp. 2395-2399, 2008.

[51] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," Journal of Chemometrics: A Journal of the Chemometrics Society, vol. 18, no. 6, pp. 275-285, 2004.

[52] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," J Inf Eng Appl, vol. 3, no. 10, 2013.

[53] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," Applied soft computing, vol. 8, no. 1, pp. 687-697, 2008.

[54] X.-S. Yang, "Firefly algorithm, stochastic test functions and design optimisation," International journal of bio-inspired computation, vol. 2, no. 2, pp. 78-84, 2010.

[55] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," Nature inspired cooperative strategies for optimization (NICSO 2010), pp. 65-74, 2010.

[56] X.-S. Yang and S. Deb, "Cuckoo search via Lévy flights," in 2009 World congress on nature & biologically inspired computing (NaBIC), 2009: Ieee, pp. 210-214.

[57] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," Advances in engineering software, vol. 69, pp. 46-61, 2014.

[58] Ö. Özdemir and E. B. Sönmez, "Weighted cross-entropy for unbalanced data with application on covid x-ray images," in 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 2020: IEEE, pp. 1-6.

[59] F. Huang, J. Li, and X. Zhu, "Balanced Symmetric Cross Entropy for Large Scale Imbalanced and Noisy Data," arXiv preprint arXiv:2007.01618, 2020.

[60] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced NLP tasks," arXiv preprint arXiv:1911.02855, 2019.

[61] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in Machine Learning in Medical Imaging: 8th International Workshop, MLMI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Proceedings 8, 2017: Springer, pp. 379-387.

[62] S. A. Taghanaki et al., "Combo loss: Handling input and output imbalance in multi-organ segmentation," Computerized Medical Imaging and Graphics, vol. 75, pp. 24-33, 2019.

[63] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.

[64] C. Kuptametee, Z.-H. Michalopoulou, and N. Aunsri, "Adaptive genetic algorithm-based particle herding scheme for mitigating particle impoverishment," Measurement, vol. 214, p. 112785, 2023.

[65] S. V. Moravvej, S. J. Mousavirad, M. H. Moghadam, and M. Saadatmand, "An LSTM-based plagiarism detection via attention mechanism and a population-based approach for pre-training parameters with imbalanced classes," in Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part III 28, 2021: Springer, pp. 690-701.

[66] J. Ye, Z. Zhu, F. Liu, R. Shokri, and V. Cevher, "Initialization Matters: Privacy-Utility Analysis of Overparameterized Neural Networks," arXiv preprint arXiv:2310.20579, 2023.

Fuzzy Neural Network Algorithm Application in User Behavior Portrait Construction

Peisen Song[1*], Bengcheng Yu[2], Chen Chen[3]
School of Economics and Management, China University of Mining and Technology, Xuzhou, 221000, China[1, 3]
School of Information Engineering, Xuzhou College of Industrial Technology, Xuzhou, 221000, China[1, 2]
School of Business Administration, Xuzhou College of Industrial Technology, Xuzhou, 221000, China[3]

***Abstract*—With the increasing number of online users, constructing user behavior profiles has received widespread attention from relevant scholars. In order to construct user behavior profiles more accurately, the research first designed an adaptive fuzzy neural network algorithm based on the momentum gradient descent method. It uses momentum gradient descent to optimize and learn the parameters adjusted by error backpropagation algorithm and least squares estimation method and optimizes the structure of the fuzzy neural network through subtraction clustering. Finally, the improved algorithm is applied to the construction of user behavior profiles. The results showed that in error analysis, the error range of the improved algorithm was within [-0.10, 0.10], and the accuracy was relatively high. In indicator calculation, the improved algorithm had a recall rate of 0.07 and 0.09 higher than the other two algorithms, an accuracy rate of 0.03 and 0.07 higher than the other two algorithms, and an F1 score of 0.07 and 0.08 higher than the other two algorithms, indicating good overall performance. In the ROC curve, the average detection rate of the designed user behavior profiling model was 0.065 and 0.155 higher than the other two models, respectively, with higher detection accuracy. These results demonstrated the effectiveness of improved algorithms and design models, providing certain reference value for the development of related fields.**

Keywords*—User behavior profiling; momentum gradient descent method; adaptive fuzzy neural network; error backpropagation algorithm; least squares estimation method; subtractive clustering

I. Introduction

In today's increasingly data-driven society, user behavior profiling is an important research object in data analysis [1]. User behavior profile construction can help enterprises gain a deeper understanding of users and develop more personalized and refined service and marketing strategies. In the process of constructing user behavior profiles, traditional user behavior analysis methods often struggle to handle such large-scale and complex data. Meanwhile, user behavior data often contains a large amount of uncertainty and ambiguity, making it a challenge to accurately analyze and predict user behavior. Fuzzy Neural Network (FNN) is an important modeling tool, which combines the adaptive learning ability of neural networks and the ability of fuzzy logic to handle uncertain and fuzzy information and can effectively learn and process nonlinear and complex data patterns [2]. Therefore, FNN is widely regarded by researchers as a powerful tool for dealing with various data problems in user behavior profiling, such as the nonlinearity and dynamic changes of user behavior, as well as the fuzzy relationship between user attributes and behaviors. However, although the theory and technology of FNN have developed to a considerable extent, its application in constructing user behavior profiles is still in the exploratory stage. On the one hand, due to the complexity of user behavior, applying FNN to practical user behavior profiling still faces many challenges; On the other hand, compared to other machine learning and data mining technologies, the superiority and adaptability of FNN in constructing user behavior profiles have not been fully demonstrated [3].

In this context, the study first utilizes the Back Propagation (BP) and Least Squares Estimation (LSE) to adjust the network parameters of FNN and then utilizes the Gradient Descent with Momentum (GDM) method to optimize and learn the adjusted parameters. The structure of FNN is optimized using the Subtractive Clustering Method (SCM) to shorten training time, and an MGD-ANFIS algorithm is designed for the model construction of user behavior profiling, in order to have a positive driving effect on the theory and practice of FNN. Compared to traditional methods, this algorithm can better process and parse user behavior data, thereby more accurately constructing user behavior models. The innovation of this study lies in the use of the BP algorithm and LSE algorithm to optimize FNN, which can more accurately characterize user behavior characteristics and provide strong support for enterprises to develop more refined service marketing strategies. The value of this study lies in providing a new type of user behavior analysis method, which has higher accuracy and predictive ability compared to traditional analysis methods and can better meet the needs of modern data-driven society for user behavior analysis.

The research content consists of six sections. Section I introduce the background of the research and propose methods. Section II is a review of online user behavior research at home and abroad, summarizing and summarizing existing research, and pointing out the shortcomings of existing research. Section III mainly constructs a network user profile model using MGD-ANFIS. Section III (A) is the design of the MGD-ANFIS algorithm, which provides a detailed introduction to the design ideas and implementation process of the MGD-ANFIS algorithm and Section III (B) is based on the MGD-ANFIS algorithm and constructs a user profile model. Section V and Section VI provides the conclusion and acknowledgment respectively.

II. Related Works

With the acceleration of the Internet process, network user behavior has been digitized, and constructing user behavior profiles through complex algorithms is currently a hot research topic. Kumar S and other researchers designed a semi-local algorithm based on the correction degree centrality exclusion ratio to maximize influence through user behavior. It used the correction degree centrality exclusion ratio idea to ensure minimal overlap between the regions affected by selected diffusion nodes. Results showed that the algorithm output value outperformed other methods [4]. Zhao and other scholars designed a spatiotemporal gated network method to recommend interest points to network users. It can use interest point context prediction to assist the next interest point through joint learning and jointly train interest point context prediction and next interest point recommendation. The results showed that this method had high accuracy [5]. Wu et al. designed a recommendation algorithm to predict user behavior using anonymous sessions. The algorithm constructed sequences, captured them using graph neural networks, and combined session representations that met conditions through attention networks. The results showed that this algorithm outperformed other methods [6]. Kumar designed a speed learning-based classifier method to predict children's behavior based on their current emotions. The probability model was introduced into the deep learning classifier and multiple sample emotions were used for prediction. The results showed that the method had high recognition rate and prediction accuracy [7]. Chen and other researchers designed a graph convolutional network method based on linear residual to model the interaction behavior between users and the network. This method can alleviate the over-smooth problem in graph convolutional aggregation operations of sparse user and network project interaction data, and the results showed that this method was highly efficient [8]. Adam et al. designed an artificial intelligence-based chat system for real-time communication with users in an e-commerce environment. Through random online experiments, they empirically tested the impact of verbal anthropomorphic design prompts and entry techniques on user request compliance. The results showed that the system had good interaction effects with users [9].

Zhang and other scholars designed a multi-scale application programming interface graph sequence model to detect the dynamic behavior of users using malicious software. It concatenates graph features from different time periods and graph scales to detect whether the software is malicious. The results showed a good performance [10]. Zhang and other researchers designed a network attack detection method that combines traffic calculation and deep learning to detect unknown attacks in high-speed networks. It utilized sliding window flow data processing to achieve real-time detection and improved classification accuracy through deep trust networks and support vector machines. The results showed that this method had high efficiency and accuracy [11]. Boone et al. designed a data-driven technology using big data technology and the Internet of Things to better execute user management and meet user needs. It can collect and analyze large amounts of data in real time, and results showed high technology accuracy [12]. Ullah and other researchers designed an Apache web server intelligent intrusion detection system using machine learning methods to enhance the security of communication between suppliers and users. It utilized naive Bayesian machine learning algorithms for training, and results showed that this system had a high validation accuracy [13]. Chen et al. designed an attention evaluation method based on multimodal data and multi-scene modeling to evaluate users' psychological states and provide early warnings. The method analyzed the relationship between emotional data and attention in-depth and corrected labels with emotional data. Results showed a high prediction efficiency [14]. Scholars such as Cui designed a combined model using the time correlation coefficient and improved K-means clustering with cuckoo search to help users obtain real-time information. Through K-means clustering, similar users were gathered together and their behavior was analyzed. Results showed a high model accuracy [15].

In summary, many scholars have improved their understanding and predictive ability of user behavior through different algorithms and models, utilizing data on online user behavior. At the same time, they have also provided new solutions for areas such as network security, e-commerce interactivity, and mental health warning. However, these methods still have certain shortcomings in terms of algorithm generalization ability and model robustness. Therefore, the study utilizes GDM to optimize and learn the network parameters after BP and LSE optimization and then uses SCM to optimize the network structure, and design the MGD-ANFIS algorithm to construct a user behavior portrait model.

III. Method

The first section of this chapter is to improve FNN and design the MGD-ANFIS algorithm. The second section is to construct a user behavior profiling model using the MGD-ANFIS algorithm.

A. MGD-ANFIS Algorithm Design

FNN is a hybrid model that combines fuzzy logic and neural networks [16-17]. Its goal is to handle fuzzy and uncertain problems through the reasoning ability of fuzzy logic and the learning ability of neural networks. In traditional neural networks, both input and output are fixed values, while in the real world; many problems have ambiguity and uncertainty. FNN can better handle these problems by introducing the concepts of fuzzy sets and fuzzy reasoning. The basic structure of FNN includes the input layer, hidden layer, and output layer. Its inputs and outputs can be fuzzy sets, rather than just fixed values. Fuzzy sets are mathematical tools used to represent fuzziness and uncertainty, which can describe the degree of membership of a value within a certain range. When training FNN, fuzzy inference, and fuzzy set methods are usually used to define the objective function and error function of the network. The BP algorithm can update the weights and biases of the network to minimize the error function [18]. FNN has extensive applications in fields such as fuzzy control, pattern recognition, and decision support systems. The FNN structure is shown in Fig. 1.

In Fig. 1, the first four layers are the precursor network,

the first layer is the input layer, the second layer is the fuzzification layer, the third layer is the fuzzy rule calculation layer, and the fourth layer is the normalization layer. The last two layers are the post network, the fifth layer is the fuzzy rule output layer, and the sixth layer is the output single node layer. FNN can handle fuzzy and uncertain inputs, providing more flexible and robust decision-making and reasoning capabilities. However, due to the complexity and computational overhead of FNN, its training and inference process may be more time-consuming than traditional neural networks [19]. Therefore, the study used the BP algorithm and LSE algorithm to adjust the parameters of the FNN's antecedent and consequent networks to minimize errors. At the same time, GDM was used to optimize and learn the adjusted parameters, and SCM was used to optimize the structure of the FNN to shorten training time. MGD-ANFIS algorithm was designed. Firstly, define a fuzzy set, and the calculation formula is shown in Eq. (1).

$$A = \left\{ (x, \mu_A(x)) \middle| x \in X \right\} \tag{1}$$

In Eq. (1), A represents a fuzzy set, x represents any feature, X represents a set of all features, and $\mu_A(x)$ represents the membership function. In the input layer, components of the input vector are directly connected to the nodes. The commonly used membership function shapes are shown in Fig. 2.

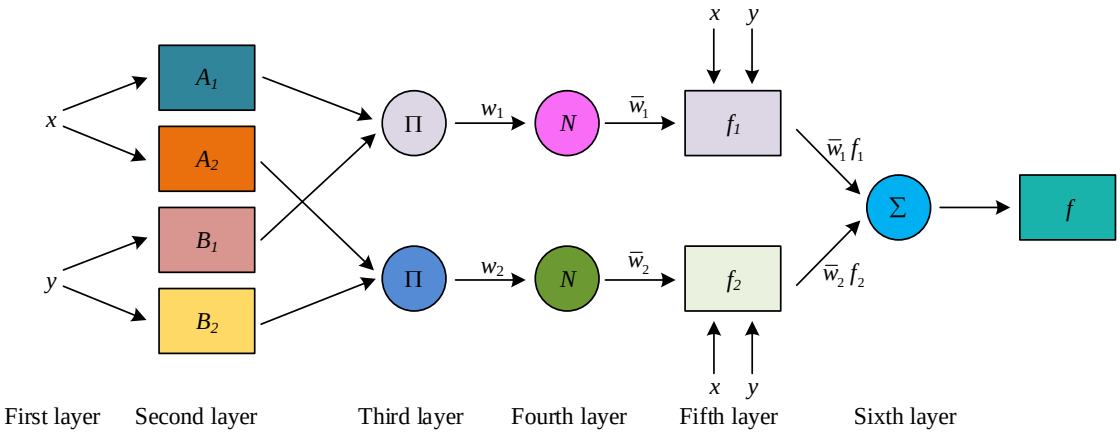


Fig. 1. Fuzzy neural network structure diagram.

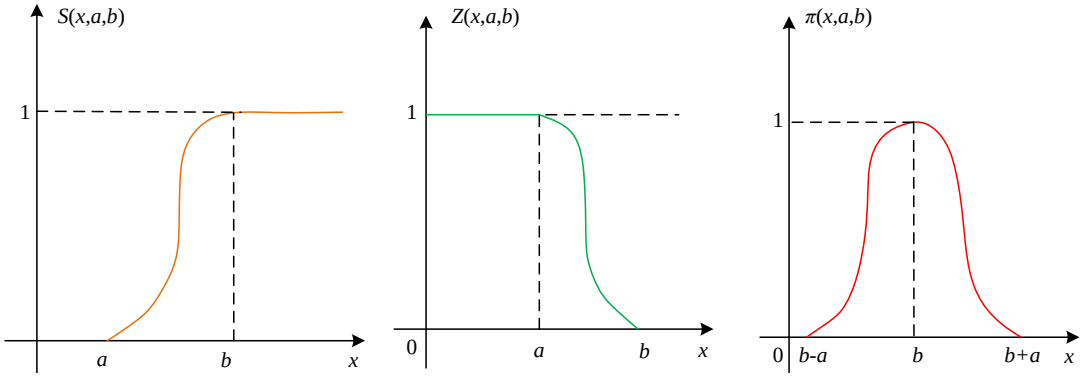


Fig. 2. Common membership function shapes.

In Fig. 2, common membership functions are divided into S-type, Z-type, and bell-type. The study adopts a Gaussian function with a bell shape as the membership function, and its expression is shown in Eq. (2).

$$f(x, \varepsilon, c) = e^{\frac{(x-c)^2}{2\varepsilon^2}} \tag{2}$$

In Eq. (2), c represents the membership function center value, and ε represents the membership function width. The next step is to map the input data onto a fuzzy aggregation through a fuzzification layer and describe the degree of membership of the input data through a membership function. The specific calculation method is shown in Eq. (3).

$$O_{1,i} = \begin{cases} \mu_{A_i}(x_1), i = 1,2 \\ \mu_{B_{i-2}}(x_2), i = 3,4 \end{cases} \tag{3}$$

In Eq. (3), $O_{1,i}$ represents the degree of membership of the input variable, x_1 and x_2 represent node i's input, and A_i and B_{i-2} represent two different fuzzy sets. The next step is to use the BP algorithm to adjust the antecedent parameters, where the momentum gradient descent method is used for BP propagation, as shown in Eq. (4).

$$\begin{cases} \dfrac{\partial E}{\partial w_{iq}} = -\delta_k^4 \times N_h^3 \\ \delta_h^3 = \delta_h^4 \times \dfrac{\sum_{s=1,s\neq h}^{m} N_h^3}{(\sum_{n=1}^{m} N_h^3)^2} \quad ,h=1,2,...,m, m=\prod_{j=1}^{n} m_j \\ \delta_{ij}^2 = \delta_h^3 \times s_{ij} \times e^{\frac{(x_j - c_{ij})^2}{\varepsilon_{ij}^2}} \end{cases} \tag{4}$$

In Eq. (4), ∂ represents the derivative, w_{iq} represents the point where the membership function value is 1, N_h^3 represents fuzzy rule calculation layer input value, δ represents the variance, s_{ij} represents the number of rules, k represents a certain rule, i and j represent nodes, and e represents natural constants. The specific expression form of the updated parameters is shown in Eq. (5).

$$\begin{cases} V_{dc} = \lambda \times dc + (1 - V_{dc}) \times dc, c = c + \alpha \times dc \\ V_{db} = \lambda \times db + (1 - V_{db}) \times db, b = b + \alpha \times db \\ V_{dw} = \lambda \times + (1 - V_{dw}) \times dw, w = w + \alpha \times dw \end{cases} \tag{5}$$

In Eq. (5), b represents the bias term, dc, db, and dw represent the differentiation of the parameters, V_{dc}, V_{db}, and V_{dw} represent the exponentially weighted average of each parameter, λ represents the momentum coefficient, and α represents the learning rate. Then the LSE algorithm can adjust consequent parameters, and the basic function of the LSE algorithm is shown in Eq. (6).

$$f(x) = a_1\varphi_1(x) + a_2\varphi_2(x) + ... + a_m\varphi_m(x) \tag{6}$$

In Eq. (6), $\varphi_m(x)$, $m=1,2,...,m$ represents a set of linearly independent functions, and a_m represents the undetermined coefficients. The calculation method for optimizing parameters using the least squares method is shown in Eq. (7).

$$\min F(x) = \sum_{i=1}^{m} f_i^2(x) \tag{7}$$

In Eq. (7), min represents the minimum value and $F(x)$ represents the objective function. The next step of the SCM algorithm is to cluster the input feature values to find the clustering center, thereby determining fuzzy rules and membership functions. The density index calculation method for feature data is shown in Eq. (8).

$$D_i = \sum_{j=1}^{n} \exp\left(\frac{\left\|x_i - x_j\right\|^2}{(\gamma_\alpha / 2)^2}\right) \tag{8}$$

In Eq. (8), D_i represents the density index, exp represents the exponential function with a base, and γ_α represents a positive number that can define a neighborhood of feature point x_i. The density index calculation method for each feature point is shown in Eq. (9).

$$\begin{cases} D_i = D_i - D_{cl} \times \exp\left[-\left(\dfrac{\left\|x_i - x_j\right\|^2}{(\gamma_\beta / 2)^2}\right)\right] \\ \gamma_\beta = k\gamma_\alpha \end{cases} \tag{9}$$

In Eq. (9), D_{cl} represents the density index corresponding to each cluster center x_{cl}, while γ_β and k both represent a positive number. The SCM clustering process is shown in Fig. 3.

In Fig. 3, the parameters are first initialized, then the density indicators of each sample point are calculated. Next, the density indicators of the remaining sample points are corrected, and whether the termination condition is met can be finally determined. If it is met, the density indicator can be output. Otherwise, recalculate. Fuzzy rules are calculated after SCM clustering, as shown in Eq. (10).

$$O_{2,i} = \mu_{Ai}(x_1)\mu_{Bi}(x_2), i = 1,2 \tag{10}$$

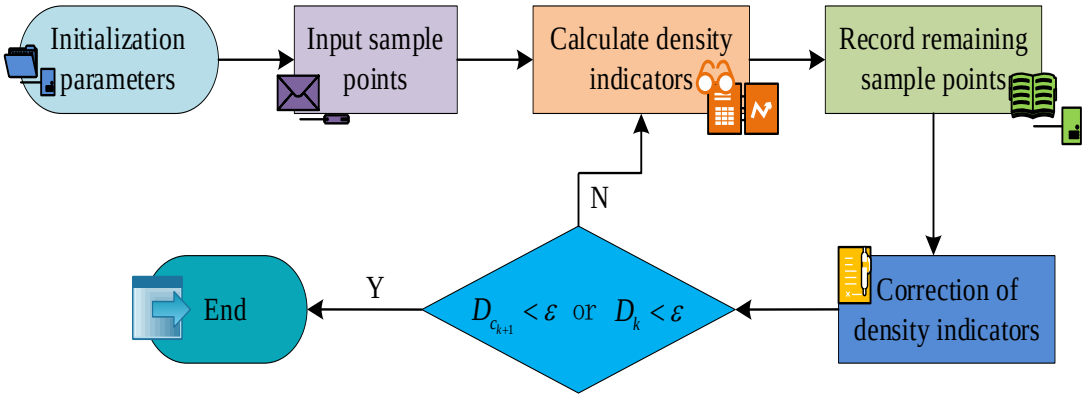


Fig. 3. SCM clustering process.

In Eq. (10), $O_{2,i}$ represents the strength of the fuzzy rule. The next step is to normalize, and the calculation method is shown in Eq. (11).

$$y = \frac{x - MinValue}{MaxValue - MinValue} \tag{11}$$

In Eq. (11), x represents the original value, y represents the converted value, $MaxValue$ represents the maximum feature value, and $MinValue$ represents the minimum feature value. The normalization results are shown in Eq. (12).

$$\begin{cases} O_{3,i} = \dfrac{w_i}{w_1 + w_2}, i = 1,2 \\ w_i = O_{2,i} \end{cases} \tag{12}$$

In Eq. (12), $O_{3,i}$ represents the normalized fuzzy rule, and w represents the numerical value of the fuzzy rule. The next step is to output the fuzzy rule, as shown in Eq. (13).

$$O_{4,i} = \bar{w}_i f_i = \bar{w}_i (p_i x_i + q_i x_2 + r_i), i = 1,2 \tag{13}$$

In Eq. (13), f_i represents the output function, and p_i, q_i, and r_i represent the node parameters. The total output can be obtained from the node output as shown in Eq. (14).

$$O_{5,i} = \sum \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{14}$$

In Eq. (14), $O_{5,i}$ represents the total output of MGD-ANFIS. When determining the parameters of the current component network, the output expression is shown in Eq. (15).

$$O'_{5,i} = (\bar{w}_1 x) p_1 + (\bar{w}_1 y) q_1 + \bar{w}_1 r_1 + (\bar{w}_2 x) p_2 + (\bar{w}_2 y) q_2 + \bar{w}_1 r_2 \tag{15}$$

In Eq. (15), $O'_{5,i}$ represents the final output value.

B. Construction of Network User Behavior Portrait Model

The construction of a network user behavior profiling model aims to identify unknown intrusions and profile user behavior. The MGD-ANFIS algorithm is used to construct a user behavior profiling model. This model can efficiently collect network user behavior data, and through filtering and compression, recombine feature vectors to generate records with various meanings to accurately identify malicious users and provide early warnings. The specific framework of the model is shown in Fig. 4.

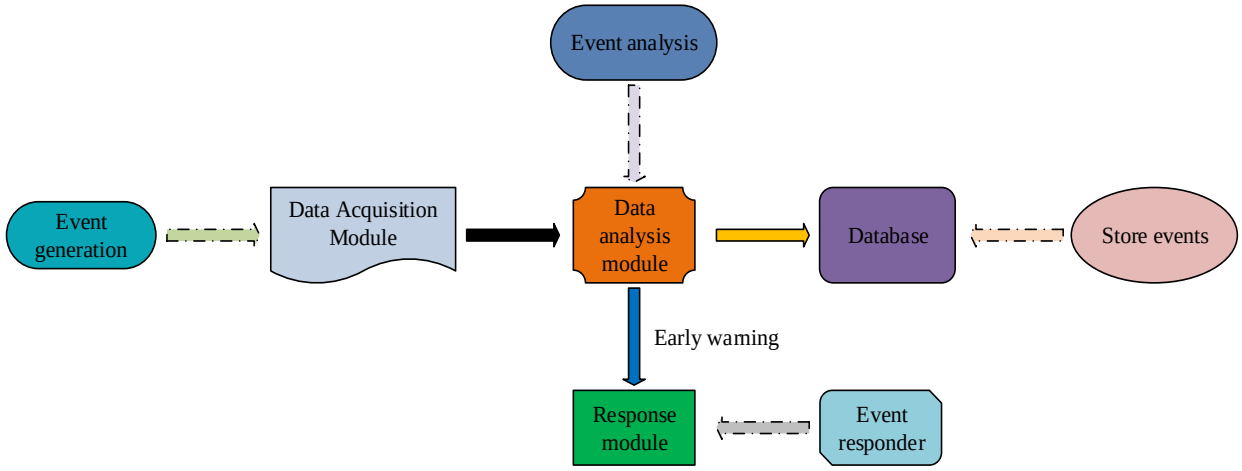


Fig. 4. User behavior portrait model framework.

In Fig. 4, the model includes four parts: data collection, data analysis, response module, and database. The data collection module is to build a user behavior database, and the integrity of the obtained database is related to the accuracy of user behavior judgment. It refers to the collection and processing of user behavior data on the network, and the use of the MGD-ANFIS algorithm for modeling and learning, by adjusting network structure and parameters to adapt to different user behavior characteristics. Data analysis is the most important component of user behavior profiling models, which obtains data from the data collection module and analyzes it, including data cleaning, data preprocessing, and feature extraction. Data cleaning refers to denoising, deduplication, recombination, and restoration of collected data to ensure data quality. Data preprocessing involves normalizing the data, and feature extraction involves extracting an appropriate number of features from the original data to distinguish whether user behavior is normal. The response module is to record, prevent, and alarm users with abnormal behavior, and can also be expanded based on the current user's input or environmental variables. The database mainly records users with abnormal behavior. Meanwhile, in user behavior profiling models, statistical feature quantity is an important factor affecting the performance of user behavior detection [20]. The research aims to determine whether a user's behavior is normal or malicious through a small number

of statistical features, and statistical features need to be designed for different types of attacks. The current common types of attacks include attacking through network vulnerabilities, exploiting network protocol flaws, and using illegal and irregular operations to enter the system. Therefore, the study designs feature quantities from three aspects: content, time traffic, and host traffic, as shown in Fig. 5.

In Fig. 5, feature statistics are in the data restoration section of data collection, which includes basic feature quantities, content feature quantities, time traffic characteristics, and host traffic characteristics. The basic feature quantity mainly counts the connection time, bytes, and network protocol types between users and the cloud. The content feature quantity mainly counts the number of access privacy and login failures. The time traffic characteristic mainly counts the connections to the same host and server. The host traffic characteristic mainly counts the connections to the same host and server in a relatively small number of connections. After designing each module, parameter settings are required. The parameters studied in this study mainly include input layer nodes, membership function, number of fuzzy subsets, and number of output layer nodes, training frequency, and radius of subtractive clustering. Finally, the obtained samples are trained and tested, and the specific process is shown in Fig. 6.

In Fig. 6, the training process and testing process correspond to the set training data and testing data, respectively. During the training process, the data consists of users with normal behavior and users with abnormal behavior, and the expected output is sent to the data analysis module as its input. Before training, it is necessary to initialize the parameters, LSE algorithm is used to identify the subsequent parameters, and continuously adjust precursor network parameters through the BP algorithm to minimize the difference between the predicted and actual output. During the detection process, data composition is consistent with the training process, but the test data is directly sent to the data analysis module as its input. After analysis, the detection rate and false alarm rate are calculated separately and compared with the results of the training data, to evaluate user behavior profile model performance based on the MGD-ANFIS algorithm.

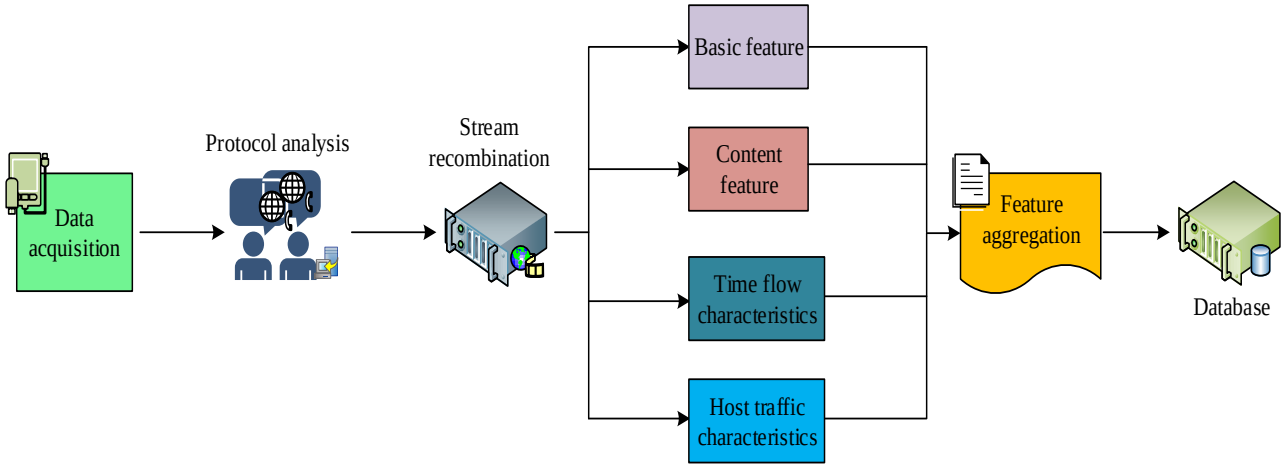


Fig. 5. Overall composition of characteristic quantities.

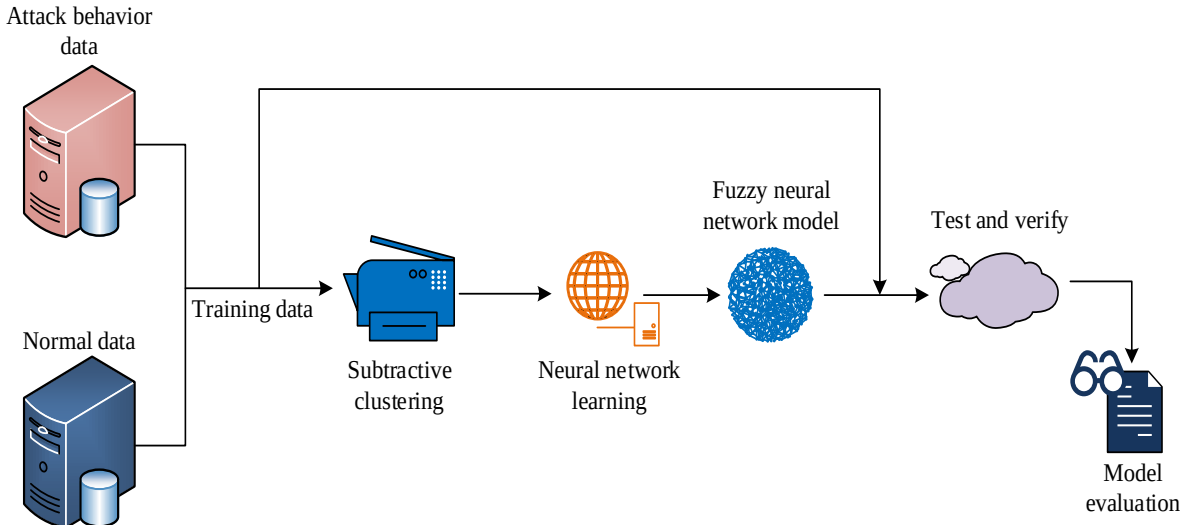


Fig. 6. The process of obtaining feature statistics.

IV. Results and Discussion

Section III (A) of this chapter analyzes the performance of the designed MGD-ANFIS algorithm, and the second section analyzes the actual application effect of the user behavior portrait model designed on the basis of the MGD-ANFIS algorithm.

A. MGD-ANFIS Algorithm Performance Analysis

To verify the designed MGD-ANFIS's performance, this study first generated 1000 pairs of datasets using the Sphere test function, and divided them into training and testing sets in 4:1. The maximum number of iterations was set to 500, and simulation comparisons were made with the MGD-ANFIS algorithm and ANFIS algorithm, respectively. The optimization results are shown in Fig. 7.

From Fig. 7(a), it can be seen that the optimization results of the ANFIS algorithm have a low fit with the test function. In Fig. 7(b), the optimization trend of the MGD-ANFIS algorithm was basically consistent with the results of the test function, with a high degree of fit. The above results indicated that the MGD-ANFIS algorithm had high accuracy and proved its effectiveness. The next step was to calculate the recall, accuracy, and F1 score of the MGD-ANFIS algorithm separately, and compare them with the ANFIS algorithm and FNN algorithm. The results are shown in Fig. 8.

In Fig. 8, the recall, accuracy, and F1 score of the MGD-ANFIS algorithm are 0.23, 0.99, and 0.20, respectively. The recall, accuracy, and F1 score of the ANFIS algorithm are 0.17, 0.96, and 0.13, respectively. The recall, accuracy, and F1 score of the FNN algorithm are 0.15, 0.92, and 0.12, respectively. Analysis shows that the MGD-ANFIS algorithm had a recall rate of 0.07 and 0.09 higher than the other two algorithms, an accuracy rate of 0.03 and 0.07 higher than the other two algorithms, and an F1 score of 0.07 and 0.08 higher than the other two algorithms. The above results demonstrate that the MGD-ANFIS algorithm had good overall performance. Finally, the MGD-ANFIS algorithm was used to perform error analysis on the training and testing sets, and compared with the ANFIS and FNN algorithms. The results are shown in Fig. 9.

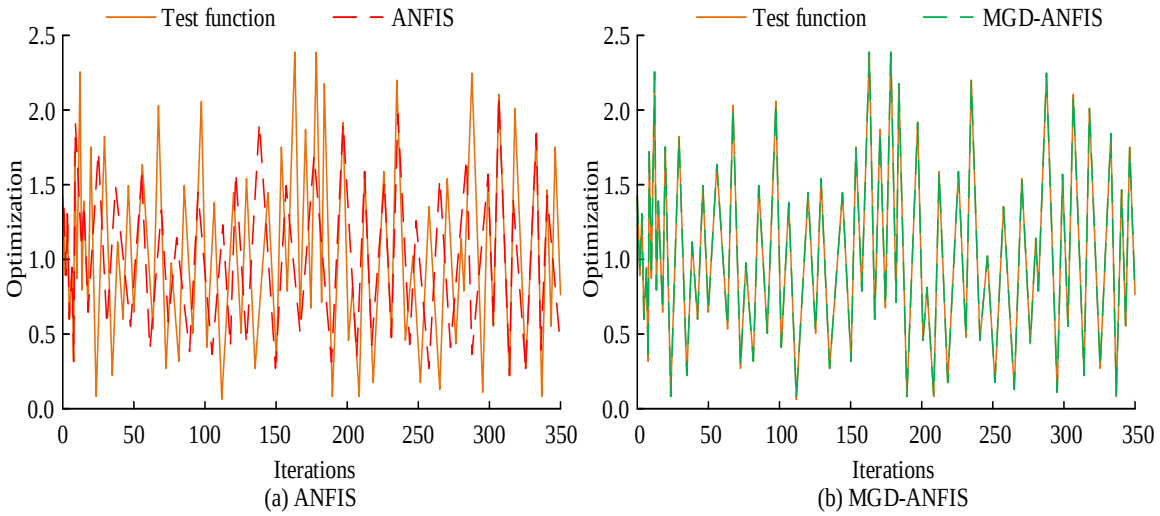


Fig. 7. MGD-ANFIS and ANFIS simulation comparison.

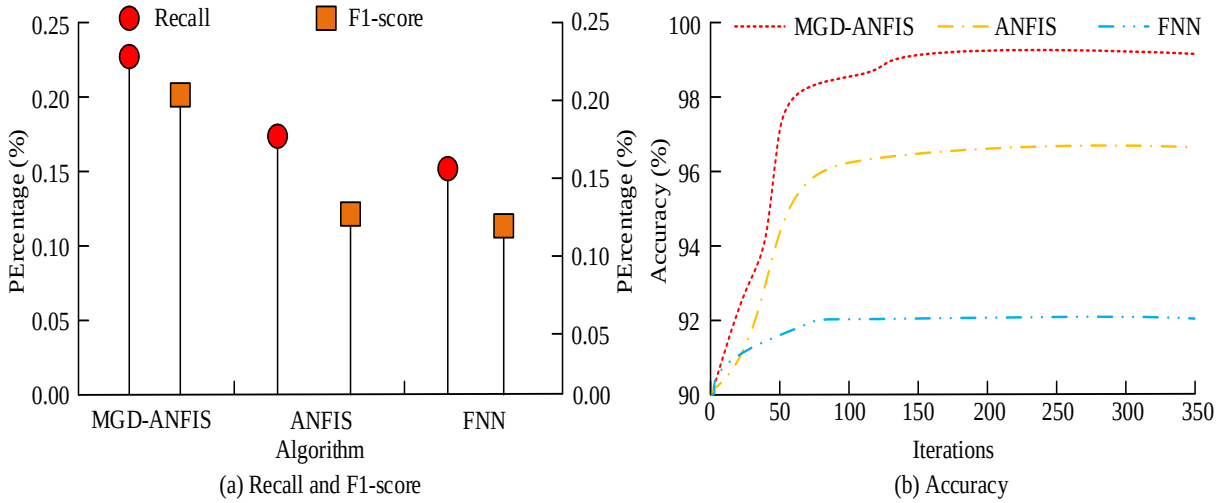


Fig. 8. Recall rate, accuracy rate, and f1 score of different algorithms.

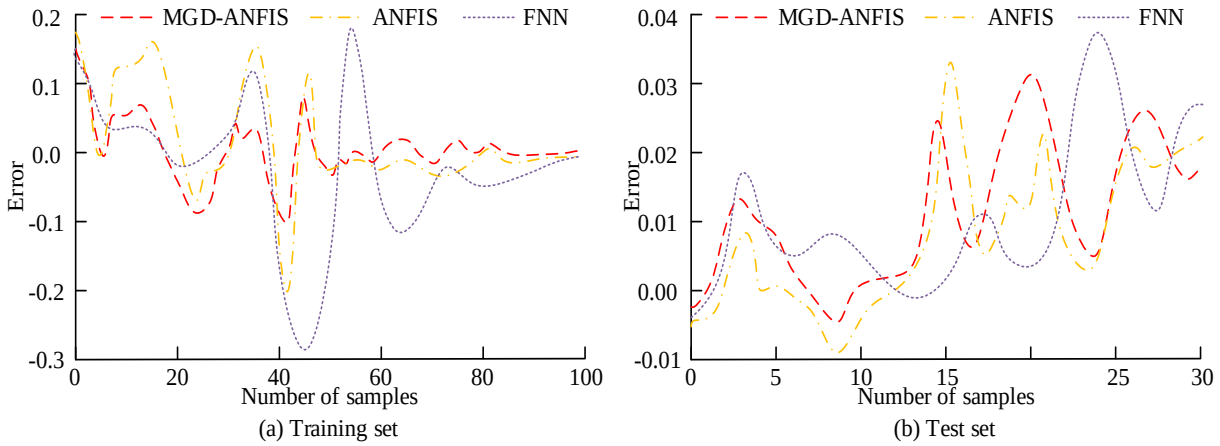


Fig. 9. Error of different algorithms on training and testing sets.

In Fig. 9 (a), in the training set error analysis, the error range of the MGD-ANFIS algorithm is between -0.10 and 0.10, the error range of ANFIS is between -0.20 and 0.18, and the error range of FNN is between -0.30 and 0.20. In Figure 9(b), in test set error analysis, the error range of the MGD-ANFIS algorithm is between -0.005 and 0.03, the error range of ANFIS is between -0.008 and 0.035, and the error range of FNN is between -0.01 and 0.04. Analysis shows that in the error analysis of the training and testing sets, the error range of the MGD-ANFIS algorithm was smaller than the other two algorithms, further indicating its high accuracy and demonstrating its good performance.

B. *Effect Analysis of User Behavior Portrait Model*

To test the performance of the designed user behavior profiling model, the study first conducted simulation experiments using the KDD99 dataset. 5 sets of data were selected with 4 training data and 1 testing data, which contained malicious user attack behavior. In training set 1, the number of normal and abnormal data was equal, and in training set 2 and the test set, the number of normal and abnormal data was also equal. Training set 2 was smaller than the training set 1 and greater than the test set. In training set 3, there was more normal data than abnormal data, while in training set 4, there was less normal data. In the network configuration, the output layer node was set to 1, with an output of 1 indicating abnormal behavior and an output of 0 indicating normal behavior. The allowable error was set to 0.2, and the clustering radius was set to 0.5. It was recommended to train the network for 50 iterations. Each training set was mixed with the test set to detect malicious user attack behavior, and the results are shown below.

TABLE I. MALICIOUS USER ATTACK BEHAVIOR IN TRAINING AND TESTING SETS

Dataset	Test error	Training Error	Training false alarm rate	Training detection rate	Test false alarm rate	Test detection rate
Training 1	0.1297	0.2673	99.2	4.9	96.3	6.3
Training 2	0.1527	0.2628	99.4	4.5	95.9	5.5
Training 3	0.2401	0.2654	98.3	4.7	94.6	6.7
Training 4	0.1399	0.2493	99.3	5.2	95.8	6.9

In Table I, the four training sets' errors are smaller than those of the test set, and the detection rate and false alarm rate are both higher than those of the test set. However, overall, different combinations of training and testing sets had higher detection rates and lower false positives and errors, indicating that the designed model had higher adaptability. The next step was to calculate the decision values for normal user behavior, vulnerability-based attacks (attack behavior 1), and network protocol defect-based attacks (attack behavior 2), as shown in Fig. 10.

In Fig. 10, users with normal behavior have a relatively small fluctuation in the judgment curve, with a maximum judgment value of 1.00, a minimum judgment value of 0.95, and an average judgment value of approximately 0.98. Attack behavior 1 involved exploiting system vulnerabilities by sending requests to the host, with a maximum decision value of 1.00, a minimum decision value of 0.58, and an average decision value of approximately 0.79. Attack behavior 2 utilized flaws in network protocols to attack, which was significantly different from normal user modes. Its maximum decision value was only 0.80, the minimum decision value was 0.10, and the average decision value was 0.45. Overall, the designed user behavior profiling model effectively distinguished between users with normal and abnormal behavior, and further proved its effectiveness. Finally, the detection rate of user behavior profiling models based on different algorithms was tested using ROC curves, and the results are shown in Fig. 11.

In Fig. 11, the designed MGD-ANFIS-based user behavior profiling model has a maximum detection rate of 1.000, a

minimum of 0.950, and an average of 0.975. The maximum of the ANFIS-based user behavior profiling model was 0.98, the minimum was 0.84, and the average was 0.91. The maximum of the user behavior profiling model based on FNN was 0.98, the minimum was 0.66, and the average was 0.82. Analysis shows that the average detection rate of the MGD-ANFIS-based user behavior profiling model was 0.065 and 0.155 higher than the other two models, respectively, proving its high detection accuracy.

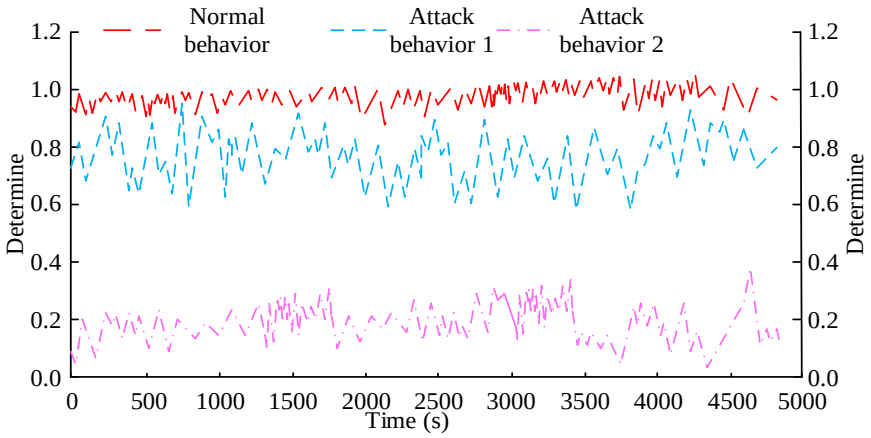


Fig. 10. Judgment values for three behaviors.

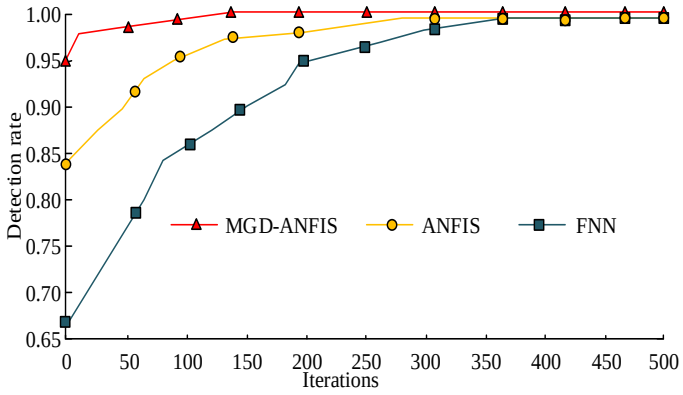


Fig. 11. Detection rate of user behavior profiling models using different algorithms.

V. Conclusion

With the popularization of mobile devices and the advancement of network technology, user behavior data have been generated in the network. How to better distinguish user behavior has become a focus of research by relevant personnel. In order to better detect user behavior, the study first adjusts the network parameters of FNN using the BP algorithm and LSE algorithm, then optimizes and learns the adjusted parameters using GDM, and optimizes the structure of FNN through SCM to shorten training time. MGD-ANFIS algorithm is designed, and finally, MGD-ANFIS is applied to user behavior profiling and a model is constructed. The results showed that in the simulation comparison, the trend of the optimization results of the MGD-ANFIS algorithm and the fitting degree of the test function were higher than those of the ANFIS algorithm, indicating its high accuracy and proving its effectiveness. In the calculation of recall, accuracy, and F1 score, the three indicator values of the MGD-ANFIS algorithm were 0.20, 0.99, and 0.20, respectively. The three indicator values of the ANFIS algorithm were 0.13, 0.96, and 0.13, respectively. The three indicator values of FNN were 0.11, 0.92, and 0.12, respectively. The three indicator values of the MGD-ANFIS algorithm were all higher than other algorithms, proving its good comprehensive performance. In the simulation experiment of the KDD99 dataset, different combinations of training and testing sets had higher detection rates and smaller false positives and errors, indicating that the designed user profile model had high adaptability. In the calculation of decision values, the average decision value for users with normal behavior was about 0.98, the average decision value for attack behavior 1 was 0.79, and the average decision value for attack behavior 2 was 0.45, proving the effectiveness of the designed model. The study only analyzed vulnerability-based attacks and protocol defect-based attacks, which had a certain impact on behavior judgment. Further exploration will be conducted in related aspects in the future. The study will conduct a more in-depth analysis of attack behavior, including the attacker's behavior patterns, attack time, and frequency, in order to better understand the attacker's motivation and strategy. Meanwhile, when constructing user behavior profiles, the quality and completeness of data have a significant impact on the accuracy and reliability of the results. Therefore, in future

research, it is necessary to consider the quality and completeness of data more comprehensively.

VI. Acknowledgment

The research is supported by: Jiangsu Qinglan Project QLGG-2022-03; Doctor Program of Xuzhou College of Industrial Technology XGY2021EA01; Industrial R&D projects of Xuzhou College of Industrial Technology XGY2022CXZ06.

References

[1] Vitiello M, Walk S, Helic D, Chang V, Guetl C. User Behavioral Patterns and Early Dropouts Detection: Improved Users Profiling through Analysis of Successive Offering of MOOC. J. Univers. Comput. Sci., 2018, 24(8): 1131-1150.

[2] Fei J, Wang Z, Liang X, Feng Z, Xue Y. Fractional sliding-mode control for microgyroscope based on multilayer recurrent fuzzy neural network. IEEE transactions on fuzzy systems, 2021, 30(6): 1712-1721.

[3] Garud K S, Jayaraj S, Lee M Y. A review on modeling of solar photovoltaic systems using artificial neural networks, fuzzy logic, genetic algorithm and hybrid models. International Journal of Energy Research, 2021, 45(1): 6-35.

[4] Kumar S, Lohia D, Pratap D, Krishna A, Panda B S. MDER: modified degree with exclusion ratio algorithm for influence maximisation in social networks. Computing, 2022, 104(2): 359-382.

[5] Zhao P, Luo A, Liu Y, Xu J, Li Z, Zhuang F, Zhou X. Where to go next: A spatio-temporal gated network for next poi recommendation. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(5): 2512-2524.

[6] Wu S, Tang Y, Zhu Y, Wang L, Xie, X S, Tan T. ession-based recommendation with graph neural networks//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(1): 346-353.

[7] Kumar D T S. Construction of hybrid deep learning model for predicting children behavior based on their emotional reaction. Journal of Information Technology and Digital World, 2021, 3(1): 29-43.

[8] Chen L, Wu L, Hong R, Zhang K, Wang M. Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(1): 27-34.

[9] Adam M, Wessel M, Benlian A. AI-based chatbots in customer service and their effects on user compliance. Electronic Markets, 2021, 31(2): 427-445.

[10] Zhang Z, Li Y, Wang W, Song H, Dong H. Malware detection with dynamic evolving graph convolutional networks. International Journal of Intelligent Systems, 2022, 37(10): 7261-7280.

[11] Zhang H, Li Y, Lv Z, Sangaiah A K, Huang T. A real-time and ubiquitous network attack detection based on deep belief network and support vector machine. IEEE/CAA Journal of Automatica Sinica, 2020, 7(3): 790-799.

[12] Boone T, Ganeshan R, Jain A, Sanders N R. Forecasting sales in the supply chain: Consumer analytics in the big data era. International Journal of Forecasting, 2019, 35(1): 170-180.

[13] Ullah M U, Hassan A, Asif M, Farooq M S, Saleem M. Intelligent Intrusion Detection System for Apache Web Server Empowered with Machine Learning Approaches. International Journal of Computational and Innovative Sciences, 2022, 1(1): 21-27.

[14] Chen M, Cao Y, Wang R, Li Y, Wu D, Liu Z. DeepFocus: Deep encoding brainwaves and emotions with multi-scenario behavior analytics for human attention enhancement. IEEE Network, 2019, 33(6): 70-77.

[15] Cui Z, Xu X, Fei X, Cai X, Cao Y, Zhang W, Chen J. Personalized recommendation system based on collaborative filtering for IoT scenarios. IEEE Transactions on Services Computing, 2020, 13(4): 685-695.

[16] Debnath S. Fuzzy quadripartitioned neutrosophic soft matrix theory and its decision-making approach. Journal of Computational and Cognitive Engineering, 2022, 1(2): 88-93.

[17] Saeed M, Ahmad M R, & Rahman A U. Refined Pythagorean Fuzzy Sets: Properties, Set-Theoretic Operations and Axiomatic Results. Journal of Computational and Cognitive Engineering, 2022, 2(1), 10–16.

[18] Hou S, Chu Y, Fei J. Adaptive type-2 fuzzy neural network inherited terminal sliding mode control for power quality improvement. IEEE transactions on industrial informatics, 2021, 17(11): 7564-7574.

[19] Yao Q. Adaptive fuzzy neural network control for a space manipulator in the presence of output constraints and input nonlinearities. Advances in Space Research, 2021, 67(6): 1830-1843.

[20] Khraisat A, Gondal I, Vamplew P, Kamruzzaman J. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity, 2019, 2(1): 1-22.

Automated Classification of Multiclass Brain Tumor MRI Images using Enhanced Deep Learning Technique

Faiz Ainur Razi[1], Alhadi Bustamam[2*], Arnida L. Latifah[3], Shandar Ahmad[4]
Department of Mathematics, Universitas Indonesia, Depok, Indonesia[1, 2]
Data Science Center, Universitas Indonesia, Depok, Indonesia[2]
Research Center for Computing, National Research and Innovation Agency, Bogor, Indonesia[3]
School of Computational and Integrative Science, Jawaharlal Nehru University, New Delhi, India[4]

***Abstract*—The brain is a vital organ, and the brain tumor is one of the most dangerous types of tumors in the world. Neuroimaging is an interesting and important discussion in diagnosing central nervous system tumors. Brain tumors have several types, namely meningioma, glioma, pituitary, schwannoma, and neurocytoma. A radiologist uses magnetic resonance imaging (MRI) to detect brain tumors because of its advantages over computed tomography. However, classifying multiclass MRI is difficult and takes a long time. This study proposes an automated classification of multiclass brain tumors using enhanced deep learning techniques. Various models are used in this research, namely VGG16, NasNet-Mobile, InceptionV3, ResNet50, and EfficientNet. For EfficientNet, we applied EfficientNet-B0–B7. From the experiments, EfficientNet-B2 is the superior, with the highest level of training accuracy of 99.90%, testing accuracy of 99.55%, precision of 99.50%, recall of 99.67%, and F1-Score of 99.58% with a training time of 15 minutes. The development of this automatic classification can assist radiologists in classifying brain tumor types more efficiently.**

***Keywords*—Brain tumor; enhanced deep learning; MRI; multiclass; neuroimaging**

I. Introduction

The central nervous system (CNS), composed of the brain and spinal cord, controls all major biological systems. It consists of supporting cells (glial cells) and nerve cells (neurons) that communicate with each other and the rest of the body by sending and receiving impulses through the nerves. Magnetic resonance imaging (MRI) is an imaging technique that shows accurate anatomical images of the human body and provides valuable data for biomedical research and clinical diagnosis. For example, MRI images help diagnose brain tumors, abnormal cell growths that form an odd segment compared with normal cells. As a vital organ in the human body for speaking, thinking, and receiving environmental responses [1], any disturbance in the brain will also affect other organs. Based on the growth speed, brain tumors are classified into benign and malignant. Benign brain tumors can be cured with surgery, but malignant brain tumors are the deadliest of the cancers and can cause instant death [2]–[4]. Meanwhile, brain tumors can either be primary and secondary (i.e., metastatic). Primary tumors originate from the brain or the nerves of the brain. Metastatic brain tumors, conversely, are caused by cancer cells that spread to the brain from other parts of the body. Clinical studies show that 30%–50% of all patients with brain metastases develop multiple lesions, depending on the type of primary cancer [5], [6].

There are three types of primary brain tumors: meningioma, glioma, and pituitary tumor. Meningiomas arise from arachnoid cells in the brain and account for 37.6% of all adult primary brain tumors. The disease accounts for approximately 35,000 new cases annually, making it the most common type of intracranial tumor in the United States [7], [8]. Gliomas are found in the cerebral pedicle and spinal cord, with symptoms such as vomiting, headache, and discomfort. Glioma tumors represent nearly 30% of primary brain tumors and 80% of all malignant ones. Based on their histopathological appearance, gliomas are traditionally classified by the World Health Organization as grades I and II (low-grade glioma), grade III (anaplastic), and grade IV (glioblastoma) [9], [10]. The pituitary is a complex organ consisting of neuroendocrine cells that secrete hormones from the adenohypophysis; posterior pituitary lobe, which is modified glia; axonal extensions of hypothalamic neurons, which secrete hormones into the bloodstream; and stromal cells, which include blood vessels, nerves, meninges, bones, and other connective tissue elements. Pituitary tumors, which arise from anterior pituitary cells and are called pituitary adenomas, are generally benign and rare (about 0.2%), showing craniospinal or systemic metastases [11]–[13]. Schwannoma (neurilemmoma) is benign neoplasms derived from tumorigenic schwann cells that protect nerve cells [14]. This condition is caused by a loss of function mutation of the neurofibromatosis type 2 (NF2) tumor gene [15]. Then, apart from that, there is also a type of brain tumor called neurocytoma. This tumor is a rare brain tumor according to the world health organization (WHO). Neurocytoma arising from the ventricle accounts for 0.1% - 0.5% of all primary brains. Moreover, these tumors rarely arise from the brain parenchyma [16].

MRI is essential in detecting brain tumors, early tumors, or CNS disorders, simultaneously seeing the response to treatment. MRI has become a part of routine clinical practice, capturing various anatomical and physiological processes [17], [18]. MRI data must be analyzed, considering that brain

*Corresponding Author, alhadi@sci.ui.ac.id

tumors generally consist of distinct structural and functional areas [19]; however, analysis can take quite a long time for accurate results. Moreover, classifying brain images of the normal brain, meningioma, glioma, pituitary, schwannoma, and neurocytoma may take longer since identifying brain images is difficult. Therefore, an automated classification of brain tumors is required to ease and speed up the analysis.

Several deep learning-based models, such as CNN, VGG (visual geometry group)-16, NasNet, and support vector machine, have been used to classify features and show reliable results [20]. Most improvements of the neural networks focused on optimizing network width, depth, and resolution. Meanwhile, the EfficientNet proposed by Tan & Le (2019) combined or collaborated those three factors. Balancing the three factors aims to obtain an optimal model at a certain complexity [21]. This study proposes an enhanced deep learning model, the EfficientNet model, to classify MRI images of brain tumors into six classes: normal brain, meningioma, glioma, pituitary, schwannoma, and neurocytoma. The EfficientNet model is expected to improve classification accuracy with less computation cost. For comparison study, four deep learning-based models namely VGG16, NasNet-Mobile, InceptionV3, and ResNet50 will also be implemented to classify MRI images.

The structure for the rest of the paper is as follows. Section II provides previous works about brain tumor classifications using various models. Section III describes the data and methods used in this study. The results of the proposed brain tumor classifications are presented in Section IV. Section V concludes the paper..

II. Related Work

D. Filatov and G. N. A. H. Yar (2022) classified brain tumors into four classes, with the highest accuracy in the EfficientNet-B0 model of 87.67%, while the ResNet50 model obtained an accuracy of 72.82%. Similarly, M. A. Gómez-Guzmán et al. (2023) used the EfficientNet-B0 model and obtained a higher accuracy of 90.88%. R. Jha, V. Bhattacharjee, and A. Mustafi (2022) used TrFEMNet and obtained an accuracy of 99.39% for two classes and 78.05% for four classes of brain tumors. With the same number of classes, A. Kowshir et al. (2023) used the ResNet50 model and obtained an accuracy of 96.67%.

Using the brain MRI dataset to classify Alzheimer's, the accuracy level obtained in the Hazarika et al. study (2022) was 86.75% and 86.25% for the NasNet-A and NasNet-C models, respectively. Research conducted by [22] classified Alzheimer's using the 3D-Hog feature. Research conducted by S. R. Sowrirajan et al., (2023) using a three-class dataset, obtained accurate results for the VGG16-NADE model with an augmentation of 96.01%, and VGG16 without augmentation of 92.33%. This accuracy is the highest compared with other models in their study. The previous research summary can be seen in Table I, which shows that the highest accuracy rate is 99.39% for two classes of brain tumors. Even though the accuracy is already high, the classification is only implemented for two classes. Meanwhile, the classification of brain tumors with more than four classes has not been investigated in detail yet.

TABLE I. Comparative Research for Brain Classification

Reference	Model	Result
[23] D. Filatov and G. N. A. H. Yar, (2022)	EfficientNet-B7 EfficientNet-B0 ResNet50	84.19% 87.67% 72.82%
[24] R. Jha et al., (2022)	TrFEMNet	2 classes: 99.39% 4 classes: 78.05%
[25] R. A. Hazarika et al., (2022)	NasNet-A NasNet-C	86.75% 86.25%
[26] M. A. Gómez-Guzmán et al., (2023)	InceptionV3 EfficientNet-B0 Generic CNN	97.12% 90.88% 81.08%
[20] A. Kowshir et al., (2023)	InceptionV3 ResNet50 Xception	94.71% 96.67% 91.18%
[27] S. R. Sowrirajan et al., (2023)	VGG16	96.01%

In this study, the classification of brain tumors into six classes, normal brain, meningioma, glioma, pituitary, schwannoma, and neurocytoma will be proposed using five types of models, VGG16, NasNet-Mobile, InceptionV3, ResNet50, and EfficientNet. Classification of six classes will provide a better level of training and validation accuracy with shorter training time using enhanced deep learning techniques.

III. Research Method

A. Dataset

Fig. 1 shows the samples of the MRI images dataset, consisting of the normal brain, meningioma, glioma, pituitary, schwannoma, and neurocytoma tumor. Meningioma is the most common CNS or primary tumor. Meningioma tumors grow from the meninges, the tissues surrounding and protecting the brain just below the skull [28].

Glioma-type tumors arise from glial cells and are intraparenchymal tumors [29], [30]. An example of glioma tumor MRI results can be seen in Fig. 1 (c). The pituitary gland is a complex organ composed of the hormone-secreting neuroendocrine cells of the pituitary gland. The MRI results for this type of tumor can be seen in Fig. 1 (d). In Fig. 1 (e) is an image of a schwannoma-type tumor and the MRI result of a neurocytoma-type tumor shown in Fig. 1 (f).

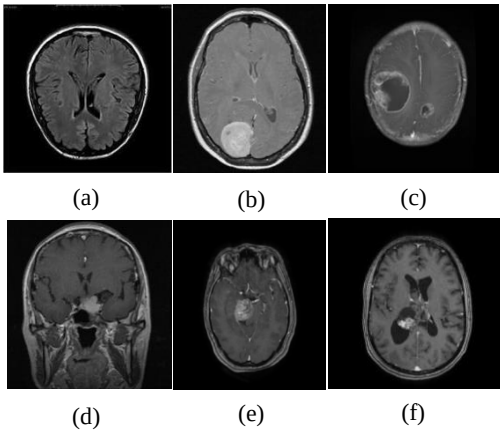


Fig. 1. MRI result (a) normal brain, (b) meningioma, (c) glioma, (d) pituitary, (e) schwannoma, (f) neurocytoma.

The dataset used in this study is retrieved from [31] and [32], and there are six types of images with a total of 7519 images. The number of details of the dataset can be seen in Table II.

TABLE II. DATASET DISTRIBUTION

Phase	Train	Test	Total
Normal	1595	205	1800
Glioma	1321	167	1488
Meningioma	1339	159	1498
Pituitary	1457	163	1620
Schwannoma	463	88	551
Neurocytoma	457	105	562
Total	**6632**	**887**	**7519**
	88.2%	11.8%	

B. VGG16 Architecture

VGG is one of the CNN architectures. VGG16 has five convolution blocks with 13 convolution layers and 3 fully connected. VGG is more capable of processing small datasets and has better recognition efficiency [33], [34]. During the training process, a loss function is used to measure the error between predicted and actual values. The following formula was applied for the cross-entropy loss function for experiments with VGG16:

$$L_{CE} = -\sum_{i=1}^{n} t_i \log p_i, \quad \text{for } n \text{ classes} \tag{1}$$

Where t_i is truth label and p_i denotes the softmax loss functions for i^{th} class. The VGG16 architecture can be seen in Fig. 2.

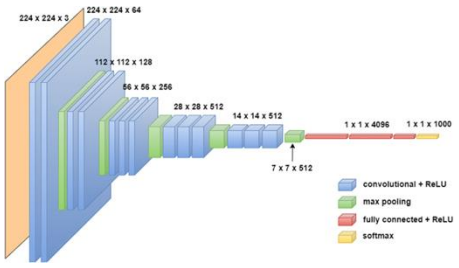


Fig. 2. VGG16 architecture [35].

C. NasNet-Mobile

NasNet is one of the CNN architectures consisting of basic building blocks optimized using reinforcement learning [36]–[38]. There are two blocks that must be considered, namely the child block and the parental block. The child block serves to adjust the network based on changes in effectiveness, while the parental block serves to evaluate the effectiveness of the child block.

NasNet development defines a high-performance building block in image set categorization (CIFAR-10). The block is generalized to a wider dataset so that it can achieve a higher classification capacity [39]. The illustration for this model can be seen in Fig. 3.

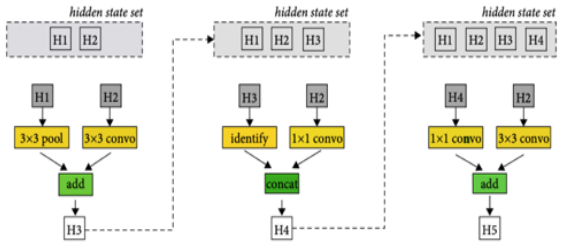


Fig. 3. NasNet architecture [55].

D. InceptionV3

InceptionV3 is the latest version of the InceptionV1 model. The InceptionV3 model has a wider network than InceptionV1 and V2. Training takes longer and is very difficult to complete. This problem is solved using transfer learning techniques [40], [41].

The structure of InceptionV3 can be seen in Fig. 4. In InceptionV3, the probability of each label $k \in \{1, \ldots, K\}$ can be determined by

$$Q(k|z) = \frac{exp(y_k)}{\sum_i^K exp(y_i)} \tag{2}$$

where, y denotes the nonnormalized log probability. The ground truth distribution on labels $p(k|z)$ is normalized by $\sum_k p(k|z) = 1$.

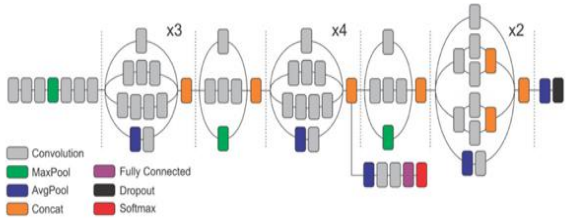


Fig. 4. InceptionV3 structure [54].

E. ResNet50

The ResNet50 model is a CNN model with a 50-layer residual network partitioned into five parts. The first part contains a convolutional layer for input preprocessing. Part 2–5 contain the bottleneck components. This model was first introduced by Microsoft in 2015 [42]. Residual building blocks can be shown in the following formula:

$$y = F(x) + x \tag{3}$$

where, $F(x)$ is the residual function, x is the input, and y is the output parameter of the residual function. ResNet50 architecture can be seen in Table III [43].

F. EfficientNet

EfficientNet has eight different architectures, namely EfficientNet-B0–B7 with the basic model being B0 obtained from neural architecture search (NAS) and B1–B7, which is an additional model with an extension of the basic model [44]. In NAS [45] to get the optimal architecture, a controller is needed to maximize the expected results represented by $J(\theta_c)$:

$$J(\theta_c) = E_{P(\alpha_{1:T};\theta_c)}[R] \tag{4}$$

TABLE III. RESNET50 ARCHITECTURE

Layer Name	Output Size	Layer
Conv1	112 x 112	7 x 7, 64, stride 2
Conv2_x	56 x 56	3 x 3 max pool, stride 2
		$\begin{bmatrix} 1\,x\,1, 64 \\ 3\,x\,3, 64 \\ 1\,x\,1, 256 \end{bmatrix} x\,3$
Conv3_x	28 x 28	$\begin{bmatrix} 1\,x\,1, 128 \\ 3\,x\,3, 128 \\ 1\,x\,1, 512 \end{bmatrix} x\,4$
Conv4_x	14 x 14	$\begin{bmatrix} 1\,x\,1, 256 \\ 3\,x\,3, 256 \\ 1\,x\,1, 1024 \end{bmatrix} x\,6$
Conv5_x	7 x 7	$\begin{bmatrix} 1\,x\,1, 512 \\ 3\,x\,3, 512 \\ 1\,x\,1, 2048 \end{bmatrix} x\,3$
	1 x 1	Average pool, 1000-d fc, softmax

List of action $\alpha_{1:T}$ is used to design the child network architecture. This child network will achieve accuracy R at the time of dataset convergence. By taking advantage of the R accuracy, it can be used as a reward signal for training controllers or as reinforcement learning. The reward signal R is not differentiable, so a gradient method is required to iteratively update θ_c by using the reinforce rule [46]:

$$\nabla_{\theta_c} J(\theta_c) = \sum_{t=1}^{T} E_{P(\alpha_{1:T};\theta_c)}\left[\nabla_{\theta_c} \log P\left(a_t \middle| a_{(t-1):1}; \theta_c\right) R\right] \quad (5)$$

The empirical approximation of the quantity (4) is:

$$\frac{1}{m}\sum_{k=1}^{m}\sum_{t=1}^{T} \nabla_{\theta_c} \log P\left(a_t \middle| a_{(t-1):1}; \theta_c\right) R_k \quad (6)$$

where, m is the number of different architectures containing example controllers in the stack. On the other hand, T is the number of hyperparameters expected by the controller to design the neural network architecture.

The validation accuracy obtained by the k^{th} neural network architecture after being trained on the training dataset is R_k. To reduce variance, the baseline function is used:

$$\frac{1}{m}\sum_{k=1}^{m}\sum_{t=1}^{T} \nabla_{\theta_c} \log P\left(a_t \middle| a_{(t-1):1}; \theta_c\right)(R_k - b) \quad (7)$$

EfficientNet is based on NAS technology as a simple, scalable, and generalizable benchmark network. In increasing the resolution and complexity of the network structure, the architecture of the EfficientNet model can be seen in Table IV [47].

By combining the three factors in the architecture, the coefficient calculation formula is as follows [48], [49]:

$$\begin{cases} depth: d = \alpha^{\phi} \\ width: w = \beta^{\phi} \\ resolution: r = \gamma^{\phi} \end{cases} \quad (8)$$

TABLE IV. ARCHITECTURE OF EFFICIENT NET

Description	Layer	Input Resolution	Channel
EfficientNet-B0	240	224x224	1280
EfficientNet-B1	342	240x240	1280
EfficientNet-B2	342	260x260	1408
EfficientNet-B3	387	300x300	1536
EfficientNet-B4	477	380x380	1792
EfficientNet-B5	579	456x456	2048
EfficientNet-B6	669	528x528	2304
EfficientNet-B7	816	600x600	2560

where, $\alpha \geq 1, \beta \geq 1, \gamma \geq 1$ and $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ with $w, d, and\ r$ can be used to scale network width, depth, and resolution coefficients. While value ϕ can be used to determine the number of effective resource extension models. The constants $\alpha, \beta, and\ \gamma$ are used to allocate these resources into three-dimensional network depth, width, and resolutions.

In the proposed model, the preprocessing stage will carry out a cropping process with an image size of 150×150. This aims to eliminate noise or delete unnecessary image information. The results of the cropping process can be seen in Fig. 5. After the cropping process is carried out, the new data will be saved in a new directory. Then the data will be augmented. This data augmentation functions to suppress overfitting when data is run through artificial data augmentation techniques [50].

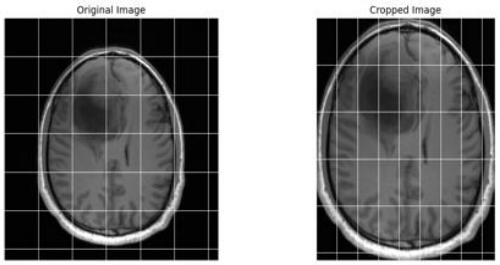


Fig. 5. Original and cropped image.

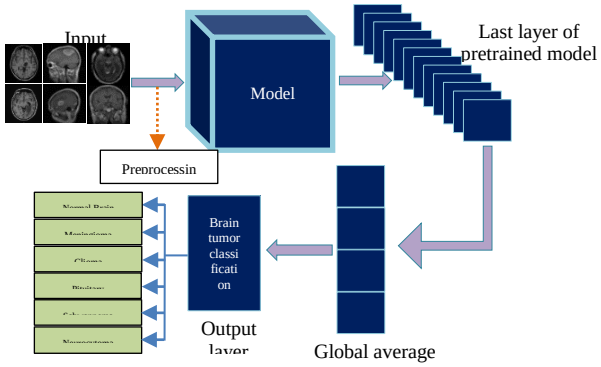


Fig. 6. Proposed model for brain tumor classification.

The complete model proposed in this study is shown in Fig. 6. The input model is the MRI image, which results in the image's class; it is detected as a normal brain or one of the brain tumors: meningioma, glioma, pituitary, schwannoma, or neurocytoma. Furthermore, the structure of the EfficientNet model is illustrated in Fig. 7.

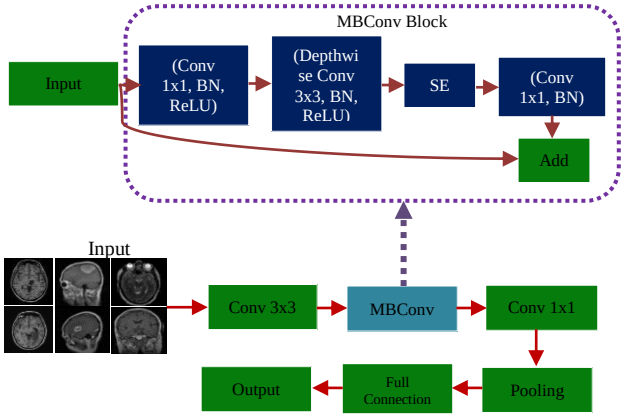


Fig. 7. The structure of the efficientnet model.

G. Model Implementation

A depth wise separable convolution (DSC) layer is used to build the base of the MobileNet block. Therefore, this hierarchical structure is also called mobile convolution (MB Conv) [51]. DSC consists of two parts: depth convolution (DWC) and point convolution (PWC). The combined convolution process between DWC and PWC is shown in Fig. 8. The goal of revealing model parameters while preserving output quality was achieved.

H. Model Evaluation

Confusion matrix provides a combination of class and actual predictions. This makes it possible to define various multiclass performance metrics as shown in Fig. 6 [52], [53]. The multiclass confusion matrix presented in Fig. 9 has dimensions of $N \times N$, where N is the number of different class labels $C_0, C_1, \ldots, C_N$. From the confusion matrix, we can compute the classification metrics: accuracy, recall, precision, and the F1-score by formulas presented in Table V.

TABLE V. PERFORMANCE METRICS FOR MULTICLASS CLASSIFICATION

Metric	Formula
Accuracy	$\frac{\sum_{i=1}^{N} TP(C_i)}{\sum_{i=1}^{N}\sum_{j=1}^{N} C_{i,j}}$
Recall of class C_i $(TPR(C_i))$	$\frac{TP(C_i)}{TP(C_i)+FN(C_i)}$
Precision of class C_i $(PPV(C_i))$	$\frac{TP(C_i)}{TP(C_i)+FP(C_i)}$
$F_1 -$ Score of class C_i	$2 \cdot \frac{TPR(C_i)\cdot PPV(C_i)}{TPR(C_i)+PPV(C_i)}$

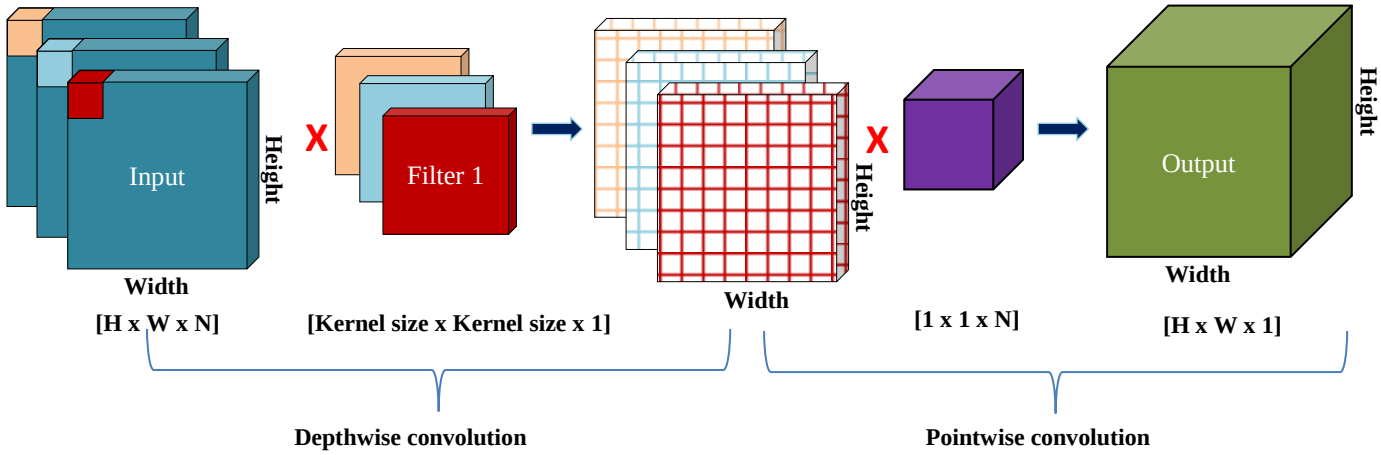


Fig. 8. Depthwise and pointwise convolution illustration.

		Predicted Class			
		C_1	C_2	…	C_N
Actual Class	C_1	$C_{1,1}$	FP	…	$C_{1,N}$
	C_2	FN	TP	…	FN
	…	…	…	…	…
	C_N	$C_{N,1}$	FP	…	$C_{N,N}$

Fig. 9. Multiclass classification.

IV. RESULT

This research conducted all experiments in the Google Colab application, with the graphic processing unit backend Google Compute Engine Python 3 A100. The RAM was 83.5 GB, disk 166.8 GB, model name Intel(R) Xeon(R) CPU@2.20GHz. The detailed parameters that have been optimized of the model experiments are shown in Table VI.

Except for the number of epochs, all models used the same parameters and loss function.

TABLE VI. PARAMETERS USED IN THE PROPOSED BRAIN TUMOR CLASSIFICATION FRAMEWORK

Hyper-parameter	VGG16	NasNet-Mobile	InceptionV3	ResNet50	EfficientNet
Optimizer	Adam	Adam	Adam	Adam	Adam
Batch Size	32	32	32	32	32
Epoch	80	40	20	20	20
Learning Rate	0.0001	0.0001	0.0001	0.0001	0.0001
Verbose	1	1	1	1	1
Loss Function	Flatten Cross-Entropy				

First, we will show the results of four models' experiments: VGG16, NasNet-Mobile, InceptionV3 and ResNet50. Initially, the performance of the training and validation of four models are presented through the loss and accuracy functions in Fig. 10. Meanwhile, Fig. 11 shows the corresponding confusion matrix that we can see that InceptionV3 only gives few false results.

Except in the VGG16 model, there is a jump in both loss and accuracy graphs of the models (see Fig. 10) (b-d). Nevertheless, as the epoch is larger the accuracy tends to improve, and the loss progressively reduces. In the accuracy graph, the initial validation accuracy is very low, 0.2 in VGG16 and NasNet-Mobile, and 0.4 in ResNet50. However, it increases to approximately 95% after the epoch is larger than 20 in the model of VGG16 and NasNet-Mobile. In InceptionV3 and ResNet50, the accuracy is already consistently more than 95% after 10 epochs. The highest accuracy of each model as shown in Table VII is achieved with epoch 24, 40, 13, and 7 for VGG16, NasNet-Mobile, InceptionV3 and ResNet50, respectively. Based on the performance results of the four models presented in Table VII, InceptionV3 achieves the best with only 13 epochs. Further, we will investigate how EfficientNet performs compared to InceptionV3, in particular.

Like the four other models, the experiments using eight varieties of EfficientNet (B0-B7) will also be compared through the loss and accuracy trend functions, confusion matrix, and the model performance metrics. Fig. 12 depicts the loss and accuracy graph of all EfficientNet models over 20 epochs. As expected, the performance of the loss and accuracy of the training are slightly better than the validation. All EfficientNet models tend to increase their performance as the epoch grows. Of the eight EfficientNet models, their performances do not differ significantly. All accuracies are mostly perfect, approximately 99%.

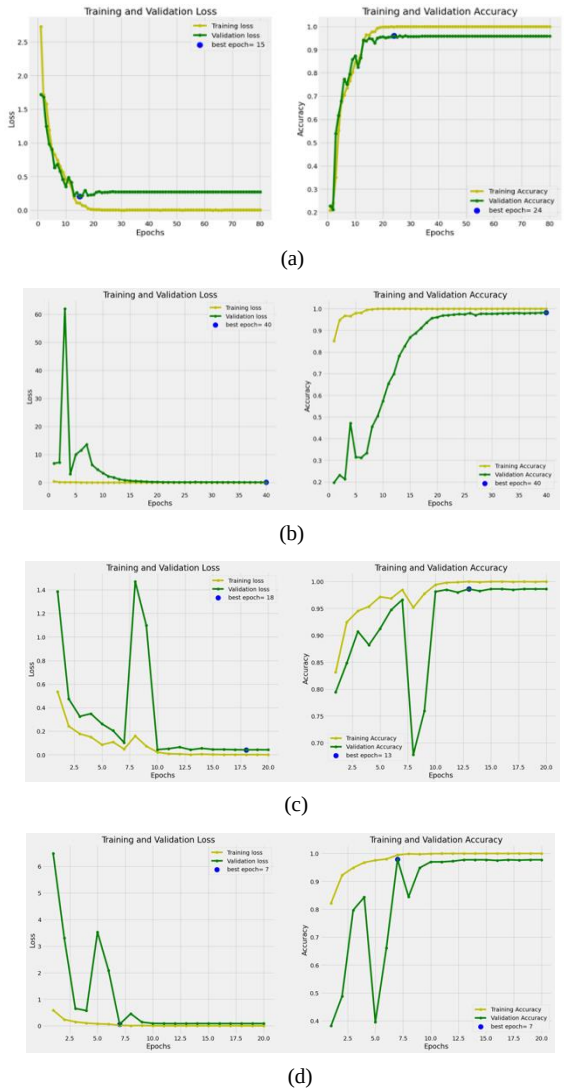


Fig. 10. The loss function and accuracy of the experiment using (a) VGG16, (b) NasNet-Mobile, (c) InceptionV3, and (d) ResNet50.

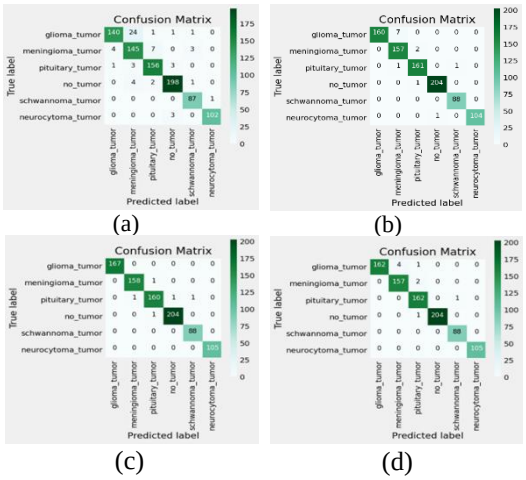


Fig. 11. The confusion matrix of experiment (a) VGG16, (b) NasNet-Mobile, (c) InceptionV3, and (d) ResNet50.

TABLE VII. PERFORMANCE MODEL

Model	Class	Precision (%)	Recall (%)	F1-Score (%)
VGG16	Glioma	97.00	84.00	90.03
	Meningioma	82.00	91.00	86.27
	Pituitary	94.00	96.00	94.99
	Normal	97.00	97.00	97.00
	Schwannoma	95.00	99.00	96.96
	Neurocytoma	99.00	97.00	97.99
	Average	**94.00**	**94.00**	94.00
NasNet-Mobile	Glioma	100.00	96.00	97.96
	Meningioma	95.00	99.00	96.96
	Pituitary	98.00	99.00	98.50
	Normal	100.00	100.00	100.00
	Schwannoma	99.00	100.00	99.50
	Neurocytoma	100.00	99.00	99.50
	Average	**98.67**	**98.83**	**98.75**
InceptionV3	Glioma	100.00	100.00	100.00
	Meningioma	99.00	99.00	99.00
	Pituitary	99.00	98.00	98.50
	Normal	100.00	100.00	100.00
	Schwannoma	99.00	100.00	99.50
	Neurocytoma	100.00	100.00	100.00
	Average	**99.50**	**99.50**	**99.50**
ResNet50	Glioma	100.00	97.00	98.48
	Meningioma	98.00	99.00	98.50
	Pituitary	98.00	99.00	98.50
	Normal	100.00	100.00	100.00
	Schwannoma	99.00	100.00	99.50
	Neurocytoma	100.00	100.00	100.00
	Average	**99.17**	**99.17**	**99.16**

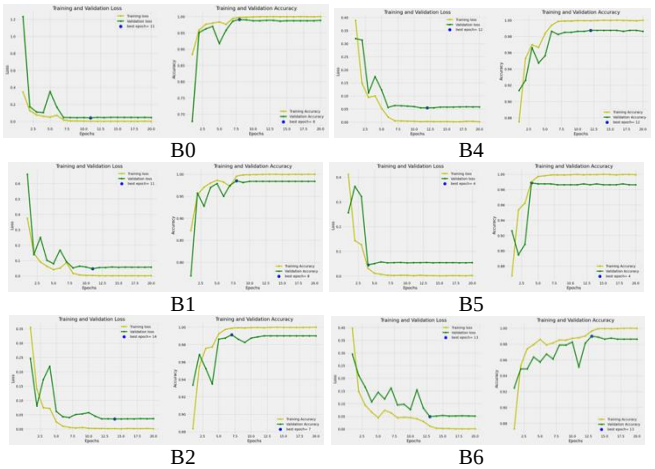


Fig. 12. The loss function and accuracy of the experiment using EfficientNet-B0-B7.

From the confusion matrix described in Fig. 13, we observe that all models only deliver a few faulty classifications. The false positives seem to be a little more than the false negatives results. There is no false negative result in EfficientNet-B1 and there is only one false negative in EfficientNet-B2, B3, and B5 while there are two false negative results in InceptionV3.

Further detail of the model performance of EfficientNet can be seen from the performance metrics of precision, recall, F1-score in Table VIII. The metrics of all EfficientNet are shown to be more than 99%.

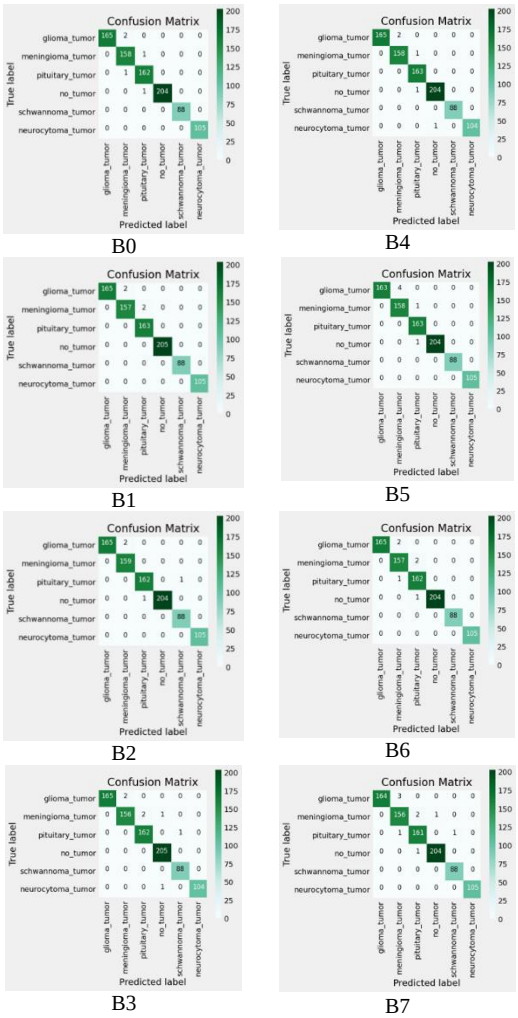


Fig. 13. The confusion matrix of the experiment using EfficientNet-B0-B7.

TABLE VIII. PERFORMANCE OF THE EFFICIENTNET MODEL

Model	Class	Precision (%)	Recall (%)	F1-Score (%)
EfficientNet-B0	Glioma	100.00	99.00	99.50
	Meningioma	98.00	99.00	98.50
	Pituitary	99.00	99.00	99.00
	Normal	100.00	100.00	100.00
	Schwannoma	100.00	100.00	100.00
	Neurocytoma	100.00	100.00	100.00

Model	Class	Precision (%)	Recall (%)	F1-Score (%)
	Average	**99.50**	**99.50**	**99.50**
EfficientNet-B1	Glioma	100.00	99.00	99.50
	Meningioma	99.00	99.00	99.00
	Pituitary	99.00	100.00	99.50
	Normal	100.00	100.00	100.00
	Schwannoma	100.00	100.00	100.00
	Neurocytoma	100.00	100.00	100.00
	Average	**99.67**	**99.67**	**99.67**
EfficientNet-B2	Glioma	100.00	99.00	99.50
	Meningioma	99.00	100.00	99.50
	Pituitary	99.00	99.00	99.00
	Normal	100.00	100.00	100.00
	Schwannoma	99.00	100.00	99.50
	Neurocytoma	100.00	100.00	100.00
	Average	**99.50**	**99.67**	**99.58**
EfficientNet-B3	Glioma	100.00	99.00	99.50
	Meningioma	99.00	98.00	98.50
	Pituitary	99.00	99.00	99.00
	Normal	99.00	100.00	99.50
	Schwannoma	99.00	100.00	99.50
	Neurocytoma	100.00	99.00	99.50
	Average	**99.33**	**99.17**	**99.25**
EfficientNet-B4	Glioma	100.00	99.00	99.50
	Meningioma	99.00	99.00	99.00
	Pituitary	99.00	100.00	99.50
	Normal	100.00	100.00	100.00
	Schwannoma	100.00	100.00	100.00
	Neurocytoma	100.00	99.00	99.50
	Average	**99.67**	**99.50**	**99.58**
EfficientNet-B5	Glioma	100.00	98.00	98.99
	Meningioma	98.00	99.00	98.50
	Pituitary	99.00	100.00	99.50
	Normal	100.00	100.00	100.00
	Schwannoma	100.00	100.00	100.00
	Neurocytoma	100.00	100.00	100.00
	Average	**99.50**	**99.50**	**99.50**
EfficientNet-B6	Glioma	100.00	99.00	99.50
	Meningioma	98.00	99.00	98.50
	Pituitary	98.00	99.00	98.50
	Normal	100.00	100.00	100.00
	Schwannoma	100.00	100.00	100.00
	Neurocytoma	100.00	100.00	100.00
	Average	**99.33**	**99.50**	**99.42**
EfficientNet-B7	Glioma	100.00	98.00	98.99
	Meningioma	97.00	98.00	97.50
	Pituitary	98.00	99.00	98.50
	Normal	100.00	100.00	100.00
	Schwannoma	99.00	100.00	99.50
	Neurocytoma	100.00	100.00	100.00
	Average	**99.00**	**99.17**	**99.08**

Next, we compare the EfficientNet model with four previous models. The models' performances are evaluated by considering the model accuracy and the computational time, the time consuming for training. The model comparisons are presented in Table IX. Based on the model accuracy in both the training and the validation data, the EfficientNet can outperform NasNet-Mobile, VGG16, InceptionV3, and ResNet50. EfficientNet-B2 achieves the highest accuracy of 99.9% in training and 99.55% in validation. The model with the least accuracy is VGG16, with a training accuracy of 97.2% and validation accuracy of 93.35%. This model also had the longest training time of 84 minutes with 80 epochs. The selection of more epochs is due to the stability of the chart. Compared with other proposed models, VGG16 requires a long epoch to be stable. In terms of the computation time, the proposed EfficientNet model requires vary time, from 10 minute in EfficientNet-B0 up to 60 minutes in EfficientNet-B7. Meanwhile, the best model, EfficientNet-B2 requires 15 minutes for the training. It is almost double the time of InceptionV3 model which only takes eight minutes.

TABLE IX. ACCURACY AND COMPUTATIONAL TIME OF MODEL

Model	Train (%)	Validation (%)	Time (m)
EfficientNet-B0	99.89	99.44	10
EfficientNet-B1	99.84	99.55	14
EfficientNet-B2	**99.90**	**99.55**	**15**
EfficientNet-B3	99.80	99.21	19
EfficientNet-B4	99.86	99.44	25
EfficientNet-B5	99.86	99.32	35
EfficientNet-B6	99.86	99.32	46
EfficientNet-B7	99.85	98.99	60
VGG16	97.20	93.35	84
NasNet-Mobile	99.82	98.53	26
ResNet50	99.77	98.99	14
InceptionV3	**99.86**	**99.44**	**8**

V. CONCLUSION

This study applies eight enhanced EfficientNet models, namely EfficientNet-B0-B7, which is based on the concept of CNN. According to existing literature, EfficientNet is a deep learning model that modifies the model so that computational efficiency produces the best results. With its efficiency advantage, we use the model to build an automated classification of brain MRI images into six classes: normal, meningioma, glioma, pituitary, schwannoma, and

neurocytoma. In this study, the EfficientNet models are also compared to the previous models, namely VGG16, NasNet-Mobile, InceptionV3, and ResNet50. All varieties of EfficientNet perform high accuracy and the EfficientNet-B2 model is superior. The EfficientNet-B2 model achieves the highest training accuracy of 99.9% and validation accuracy of 99.55%. However, it takes a slightly longer time to do the training. It requires 15 minutes, while InceptionV3 only needs eight minutes to achieve a training accuracy of 99.86%. Both EfficientNet-B2 and InceptionV3 models are best options in classifying brain MRI images efficiently and accurately.

ACKNOWLEDGMENT

This research is funded by Directorate of Research and Development, Universitas Indonesia under research grant PUTI 2023-2024 with contract number NKB-723/UN2.RST/HKP.05.00/2023.

REFERENCES

[1] K. S. Ayomide, T. Noranis, M. Aris, and M. Zolkepli, "Improving Brain Tumor Segmentation in MRI Images through Enhanced Convolutional Neural Networks," vol. 14, no. 4, pp. 670–678, 2023.

[2] X. Gu, Z. Shen, J. Xue, Y. Fan, and T. Ni, "Brain Tumor MR Image Classification Using Convolutional Dictionary Learning With Local Constraint," Front. Neurosci., vol. 15, no. May, pp. 1–12, 2021, doi: 10.3389/fnins.2021.679847.

[3] C. Ge, I. Y. H. Gu, A. S. Jakola, and J. Yang, "Deep semi-supervised learning for brain tumor classification," BMC Med. Imaging, vol. 20, no. 1, pp. 1–11, 2020, doi: 10.1186/s12880-020-00485-0.

[4] F. A. Razi, A. Bustamam, and A. L. Latifah, "Development of Efficient Brain Tumor Classification on MRI Image Results Using EfficientNet," 2023 Int. Semin. Intell. Technol. Its Appl., pp. 575–580, 2023, doi: 10.1109/ISITIA59021.2023.10221186.

[5] M. A. Proescholdt et al., "The management of brain metastases—systematic review of neurosurgical aspects," Cancers (Basel)., vol. 13, no. 7, pp. 1–17, 2021, doi: 10.3390/cancers13071616.

[6] H. Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA. Cancer J. Clin., vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[7] K. Huntoon, A. M. S. Toland, and S. Dahiya, "Meningioma: A Review of Clinicopathological and Molecular Aspects," Front. Oncol., vol. 10, no. October, pp. 1–14, 2020, doi: 10.3389/fonc.2020.579599.

[8] J. Driver et al., "A molecularly integrated grade for meningioma," Neuro. Oncol., vol. 24, no. 5, pp. 796–808, 2022, doi: 10.1093/neuonc/noab213.

[9] M. M. Shaver et al., "Optimizing neuro-oncology imaging: A review of deep learning approaches for glioma imaging," Cancers (Basel)., vol. 11, no. 6, pp. 1–14, 2019, doi: 10.3390/cancers11060829.

[10] L. Rong, N. Li, and Z. Zhang, "Emerging therapies for glioblastoma: current state and future directions," J. Exp. Clin. Cancer Res., vol. 41, no. 1, pp. 1–18, 2022, doi: 10.1186/s13046-022-02349-7.

[11] S. L. Asa, O. Mete, A. Perry, and R. Y. Osamura, "Overview of the 2022 WHO Classification of Pituitary Tumors," Endocr. Pathol., vol. 33, no. 1, pp. 6–26, 2022, doi: 10.1007/s12022-022-09703-7.

[12] L. Kasuki and G. Raverot, "Definition and diagnosis of aggressive pituitary tumors," Rev. Endocr. Metab. Disord., vol. 21, no. 2, pp. 203–208, 2020, doi: 10.1007/s11154-019-09531-x.

[13] S. L. Asa et al., "From pituitary adenoma to pituitary neuroendocrine tumor (pitnet): An international pituitary pathology club proposal," Endocr. Relat. Cancer, vol. 24, no. 4, pp. C5–C8, 2017, doi: 10.1530/ERC-17-0004.

[14] J. Gosk, O. Gutkowska, M. Urban, W. Wnukiewicz, P. Reichert, and P. Ziółkowski, "Results of surgical treatment of schwannomas arising from extremities," Biomed Res. Int., vol. 2015, 2015, doi: 10.1155/2015/547926.

[15] D. L. Helbing, A. Schulz, and H. Morrison, "Pathomechanisms in schwannoma development and progression," Oncogene, vol. 39, no. 32, pp. 5421–5429, 2020, doi: 10.1038/s41388-020-1374-5.

[16] R. Mohamed et al., "Clinicopathological features and treatment outcome of central neurocytoma: a single institute experience," Egypt. J. Neurol. Psychiatry Neurosurg., vol. 58, no. 1, 2022, doi: 10.1186/s41983-022-00540-3.

[17] J. Kalpathy-Cramer, E. R. Gerstner, K. E. Emblem, O. Andronesi, and B. Rosen, "Advanced Magnetic Resonance Imaging of the Physical Processes in Human Glioblastoma," NIH-PA Author Manuscr., vol. 74, no. 17, pp. 4622–4637, 2015, doi: 10.1158/0008-5472.CAN-14-0383.Advanced.

[18] A. I. Neugut et al., "Magnetic Resonance Imaging-Based Screening for Asymptomatic Brain Tumors: A Review," Oncologist, vol. 24, no. 3, pp. 375–384, 2019, doi: 10.1634/theoncologist.2018-0177.

[19] D. Aquino, A. Gioppo, G. Finocchiaro, M. G. Bruzzone, and V. Cuccarini, "MRI in Glioma Immunotherapy: Evidence, Pitfalls, and Perspectives," J. Immunol. Res., vol. 2017, 2017, doi: 10.1155/2017/5813951.

[20] A. Kowshir, H. Imam, S. Yesmin, and I. Mahmud, "Tumor-Net : convolutional neural network modeling for classifying brain tumors from MRI images," vol. 9, no. 2, pp. 148–160, 2023.

[21] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 36th Int. Conf. Mach. Learn. ICML 2019, vol. 2019-June, pp. 10691–10700, 2019.

[22] D. Sarwinda and A. Bustamam, "3D-HOG Features-Based Classification using MRI Images to Early Diagnosis of Alzheimer's Disease," Proc. - 17th IEEE/ACIS Int. Conf. Comput. Inf. Sci. ICIS 2018, pp. 457–462, 2018, doi: 10.1109/ICIS.2018.8466524.

[23] D. Filatov and G. N. A. H. Yar, "Brain Tumor Diagnosis and Classification via Pre-Trained Convolutional Neural Networks," pp. 0–5, 2022, [Online]. Available: http://arxiv.org/abs/2208.00768.

[24] R. Jha, V. Bhattacharjee, and A. Mustafi, "Transfer Learning with Feature Extraction Modules for Improved Classifier Performance on Medical Image Data," Sci. Program., vol. 2022, 2022, doi: 10.1155/2022/4983174.

[25] R. A. Hazarika, D. Kandar, and A. K. Maji, "An experimental analysis of different Deep Learning based Models for Alzheimer's Disease classification using Brain Magnetic Resonance Images," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 10, pp. 8576–8598, 2022, doi: 10.1016/j.jksuci.2021.09.003.

[26] M. A. Gómez-Guzmán et al., "Classifying Brain Tumors on Magnetic Resonance Imaging by Using Convolutional Neural Networks," Electron., vol. 12, no. 4, pp. 1–22, 2023, doi: 10.3390/electronics12040955.

[27] S. R. Sowrirajan, S. Balasubramanian, and R. S. P. Raj, "MRI Brain Tumor Classification Using a Hybrid VGG16-NADE Model," Brazilian Arch. Biol. Technol., vol. 66, no. Mcc, 2023, doi: 10.1590/1678-4324-2023220071.

[28] C. Ogasawara, B. D. Philbrick, and D. C. Adamson, "Meningioma: A review of epidemiology, pathology, diagnosis, treatment, and future directions," Biomedicines, vol. 9, no. 3, 2021, doi: 10.3390/biomedicines9030319.

[29] A. Lasocki, M. Anjari, S. Örs Kokurcan, and S. C. Thust, "Conventional MRI features of adult diffuse glioma molecular subtypes: a systematic review," Neuroradiology, vol. 63, no. 3, pp. 353–362, 2021, doi: 10.1007/s00234-020-02532-7.

[30] Q. D. Buchlak, N. Esmaili, J. C. Leveque, C. Bennett, F. Farrokhi, and M. Piccardi, "Machine learning applications to neuroimaging for glioma detection and classification: An artificial intelligence augmented systematic review," J. Clin. Neurosci., vol. 89, pp. 177–198, 2021, doi: 10.1016/j.jocn.2021.04.043.

[31] M. Nickparvar, "Brain Tumor MRI Dataset," 2021. https://www.kaggle.com/datasets/masoudnickparvar (accessed May 18, 2023).

[32] F. Feltrin, "Brain Tumor MRI Images 44 Classes," 2023. https://www.kaggle.com/datasets/fernando2rad (accessed Jun. 01, 2023).

[33] F. Zhao, B. Zhang, Z. Zhang, X. Zhang, and C. Wei, "Classification and detection method of Blood lancet based on VGG16 network," 2021 IEEE Int. Conf. Mechatronics Autom. ICMA 2021, pp. 849–853, 2021, doi: 10.1109/ICMA52036.2021.9512686.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–14, 2015.

[35] C. Xia, M. Wang, Y. Fan, Z. Yang, and X. Du, "A hierarchical autoencoder and temporal convolutional neural network reduced-order model for the turbulent wake of a three-dimensional bluff body," no. August 2015, 2023, doi: 10.1063/5.0137285.

[36] F. Saxen, P. Werner, S. Handrich, E. Othman, L. Dinges, and A. Al-Hamadi, "Face attribute detection with mobilenetv2 and nasnet-mobile," Int. Symp. Image Signal Process. Anal. ISPA, vol. 2019-Septe, pp. 176–180, 2019, doi: 10.1109/ISPA.2019.8868585.

[37] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 8697–8710, 2018, doi: 10.1109/CVPR.2018.00907.

[38] P. M. Shah et al., "DC-GAN-based synthetic X-ray images augmentation for increasing the performance of EfficientNet for COVID-19 detection," no. August 2021, pp. 1–13, 2022, doi: 10.1111/exsy.12823.

[39] F. Martínez, F. Martínez, and E. Jacinto, "Performance evaluation of the NASnet convolutional network in the automatic identification of COVID-19," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 10, no. 2, pp. 662–667, 2020, doi: 10.18517/ijaseit.10.2.11446.

[40] M. Mujahid, F. Rustam, R. Álvarez, J. Luis Vidal Mazón, I. de la T. Díez, and I. Ashraf, "Pneumonia Classification from X-ray Images with Inception-V3 and Convolutional Neural Network," Diagnostics, vol. 12, no. 5, pp. 1–16, 2022, doi: 10.3390/diagnostics12051280.

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 2818–2826, 2016, doi: 10.1109/CVPR.2016.308.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition Kaiming," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, vol. 45, no. 8, pp. 770–778, doi: 10.1002/chin.200650130.

[43] R. Zhang, Y. Zhu, Z. Ge, H. Mu, D. Qi, and H. Ni, "Transfer Learning for Leaf Small Dataset Using Improved ResNet50 Network with Mixed Activation Functions," Forests, vol. 13, no. 12, 2022, doi: 10.3390/f13122072.

[44] X. Ma, X. Li, A. Luo, J. Zhang, and H. Li, "Galaxy image classification using hierarchical data learning with weighted sampling and label smoothing," Mon. Not. R. Astron. Soc., vol. 519, no. 3, pp. 4765–4779, 2023, doi: 10.1093/mnras/stac3770.

[45] B. Zoph and Q. V Le, "Neural Architecture Search With Reinforcement Learning," Iclr 2017, vol. 32, pp. 1–16, 2017.

[46] R. J. Willia, "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning," Mach. Learn., vol. 8, no. 3, pp. 229–256, 1992, doi: 10.1023/A:1022672621406.

[47] C. M. Bhuma and R. Kongara, "Childhood medulloblastoma classification using efficientnets," 2020 IEEE Bombay Sect. Signat. Conf. IBSSC 2020, pp. 64–68, 2020, doi: 10.1109/IBSSC51096.2020.9332175.

[48] J. Liu, M. Wang, L. Bao, and X. Li, "EfficientNet based recognition of maize diseases by leaf image classification," J. Phys. Conf. Ser., vol. 1693, no. 1, 2020, doi: 10.1088/1742-6596/1693/1/012148.

[49] Y. Tadepalli, M. Kollati, S. Kuraparthi, and P. Kora, "EfficientNet-B0 based monocular dense-depth map estimation," Trait. du Signal, vol. 38, no. 5, pp. 1485–1493, 2021, doi: 10.18280/ts.380524.

[50] G. S. Tandel, A. Tiwari, and O. G. Kakde, "Performance enhancement of MRI-based brain tumor classification using suitable segmentation method and deep learning-based ensemble algorithm," Biomed. Signal Process. Control, vol. 78, no. March, p. 104018, 2022, doi: 10.1016/j.bspc.2022.104018.

[51] B. Hu, J. Tang, J. Wu, and J. Qing, "An Attention EfficientNet-Based Strategy for Bearing Fault Diagnosis under Strong Noise," Sensors, vol. 22, no. 17, pp. 1–19, 2022, doi: 10.3390/s22176570.

[52] P. Machart and L. Ralaivola, "Confusion Matrix Stability Bounds for Multiclass Classification," 2012, [Online]. Available: http://arxiv.org/abs/1202.6221.

[53] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, "Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem," Technologies, vol. 9, no. 4, 2021, doi: 10.3390/technologies9040081.

[54] S. F. Stefenon, K. C. Yow, A. Nied, and L. H. Meyer, "Classification of distribution power grid structures using inception v3 deep neural network," Electr. Eng., vol. 104, no. 6, pp. 4557–4569, 2022, doi: 10.1007/s00202-022-01641-1.

[55] S. V. Kogilavani et al., "COVID-19 Detection Based on Lung Ct Scan Using Deep Learning Techniques," Comput. Math. Methods Med., vol. 2022, 2022, doi: 10.1155/2022/7672196.

Nature-Inspired Optimization for Virtual Machine Allocation in Cloud Computing: Current Methods and Future Directions

Xiaoqing YANG*
Henan Technical, College of Construction, Zhengzhou 450000, China

***Abstract*—An expanding range of services is offered by cloud data centers. The execution of application tasks is facilitated by assigning (VMs) Virtual Machines to (PMs) Physical Machines. Speaking of VM allocation in the cloud service center, two key factors are taken into consideration: quality of service (QoS) and energy consumption. The cloud service center aims to optimize these aspects while allocating VMs. On the other hand, cloud users have their priorities and focus on their specific requirements, particularly throughput and reliability. User requirements are considered by the cloud service center, resulting in VM allocation that meets QoS targets and optimizes energy consumption. Cloud service centers must, therefore, find a balance between QoS and energy efficiency while considering the user's requirements. To achieve this, various optimization algorithms and techniques must be employed. The objective is to find the best allocation of VMs to PMs. Due to the NP-hardness of the VM allocation problem, nature-inspired meta-heuristic algorithms have become commonly used to solve it. However, there are no comprehensive and in-depth review papers on this specific area. This paper aims to bridge a knowledge gap by providing an understanding of the significance of metaheuristic methods to address the VM allocation issue effectively. It not only highlights the role played by these algorithms but also examines the existing methods, provides comprehensive comparisons of strategies based on key parameters, and concludes with valuable recommendations for future research.**

***Keywords*—Cloud computing; virtualization; virtual machine allocation; optimization**

I. INTRODUCTION

Cloud computing, characterized by on-demand services utilizing virtualized computing resources and streamlined software and hardware maintenance [1], has significantly shifted organizational paradigms from private infrastructure to cloud-based platforms [2]. However, the rapid growth of cloud data centers has brought about challenges, notably in energy consumption and environmental impact due to the extensive deployment of computing resources [3]. Service provisioning in cloud computing revolves around Service Level Agreements (SLAs), offering a spectrum of services encompassing hardware/software rental, resource management, and workload distribution [4]. The versatility of cloud services, encompassing Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS), and the comprehensive concept of Everything as a Service (XaaS), optimizes IT infrastructure for enhanced service delivery [5]. The pervasive adoption of cloud computing has amplified concerns regarding the substantial energy consumption inherent in data centers, accentuated by underutilized servers leading to inefficient energy utilization [6]. Addressing this challenge necessitates optimizing server utilization in data centers while ensuring seamless service delivery [7]. However, the increasing diversity of resource types in data center architectures poses a challenge in improving resource efficiency, especially in dispersed and heterogeneous environments owned by major cloud providers. The fundamental problem centers on efficiently allocating resources based on cloud user requests while adhering to SLAs.

The convergence of Internet of Things (IoT), fuzzy logic, Machine Learning (ML), Deep Learning (DL), Neural Networks (NNs), and meta-heuristic algorithms shapes efficient cloud resource allocation. The proliferation of IoT has generated vast data streams, requiring sophisticated allocation mechanisms [8, 9]. Fuzzy logic, integrating imprecise or uncertain data, enhances decision-making in allocating resources, considering ambiguous parameters [10, 11]. ML comprising supervised and unsupervised learning paradigms, aids in predicting resource demands and pattern recognition for optimized allocations [12-14]. DL, a subset of ML, with its complex neural networks, enables automatic feature extraction, fostering accurate resource predictions, and allocation decision-making [15, 16]. NNs, mimicking human brain functions, offer robust solutions for dynamic resource allocation challenges by learning from patterns and behaviors in cloud environments [17].

By employing clustering techniques, cloud systems can categorize and group entities with similar attributes or behaviors, allowing for more efficient resource allocation strategies. This approach enables the identification of patterns and similarities among diverse entities, such as VMs or user requests, facilitating the allocation of resources based on common characteristics [18]. Moreover, meta-heuristic algorithms, drawing inspiration from natural phenomena, provide efficient search strategies in complex solution spaces, optimizing cloud resource allocation by addressing scalability, dynamicity, and diverse user requirements [19]. This convergence is pivotal in enhancing the adaptability, accuracy, and efficiency of cloud resource allocation, catering to the burgeoning demands of modern cloud infrastructures while enabling dynamic, scalable, and optimized resource

provisioning, underpinning the evolution and sustainability of cloud computing ecosystems [20].

The primary questions addressed in this study revolve around the optimization of energy consumption and resource allocation complexities within cloud computing. Firstly, how can server utilization be improved in data centers to mitigate energy wastage? Secondly, what effective strategies can manage the diverse resource types and demand patterns inherent in cloud systems? Finally, how can resources be allocated efficiently while meeting the diverse SLA requirements of cloud users? These key inquiries guide the exploration of efficient resource allocation methodologies and user-centric approaches within the realm of cloud computing. This review paper aims to address the gap in existing literature by comprehensively exploring solutions for efficient resource allocation in cloud computing. It assesses methodologies to optimize energy consumption while meeting SLAs, investigates resource allocation complexities, and suggests user-centric strategies. By offering comparative insights and recommendations, this study aims to provide a robust understanding of cloud resource allocation for future research directions.

II. BACKGROUND

Various methods have been developed to optimize the allocation of VMs on physical machines in cloud data centers. Fig. 1 presents an overview of virtual machine allocation strategies. Using multidimensional resources in an unbalanced manner can enhance resource utilization and lead to abnormalities. A balanced use of multidimensional resources refers to the use of resources in each dimension proportional to their total amount. Virtual machines may run out of resources if placed regardless of this characteristic. Placement decisions can take into account various costs, including virtual machines, physical machines, cooling, data centers, and traffic, as shown in Fig. 2. The cost of cloud services is influenced by multiple factors. One aspect focuses on reducing costs for cloud users, specifically, the expense associated with virtual machines. Simultaneously, other factors aim to decrease costs for cloud service providers. Virtual machine allocation involves assigning virtual machines to suitable physical machines. To address the complexity of the NP-hard problems involved in this allocation, meta-heuristic algorithms are employed. Virtualization technology is critical for cloud computing, and the task of assigning virtual machines to physical machines is referred to as the virtual machine allocation problem, a known NP-hard challenge.

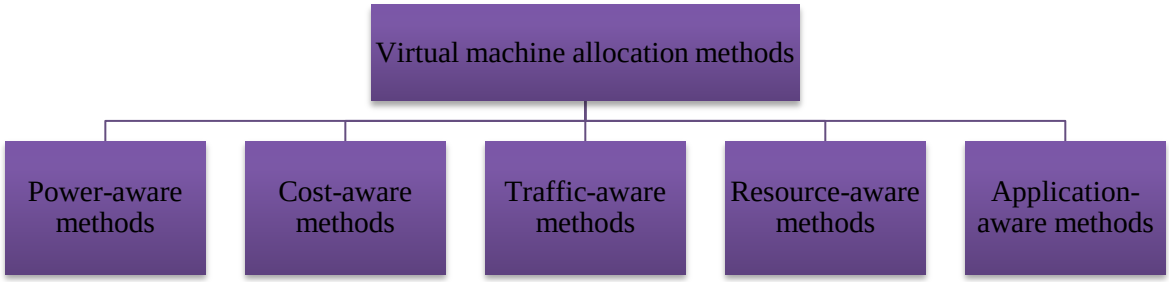


Fig. 1. Taxonomy of virtual machine allocation methods

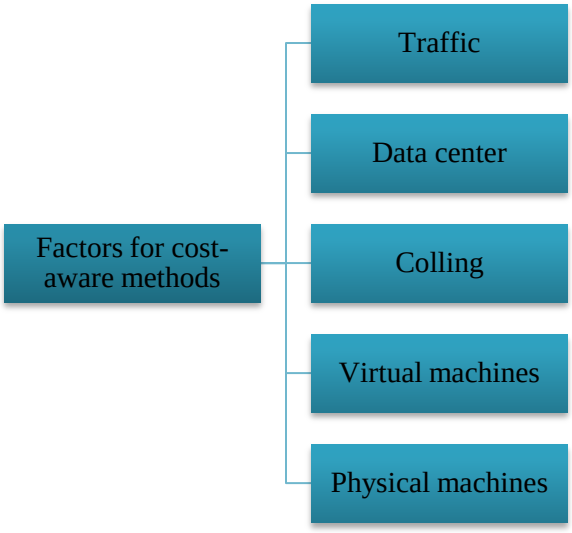


Fig. 2. Considered factors for cost-aware methods.

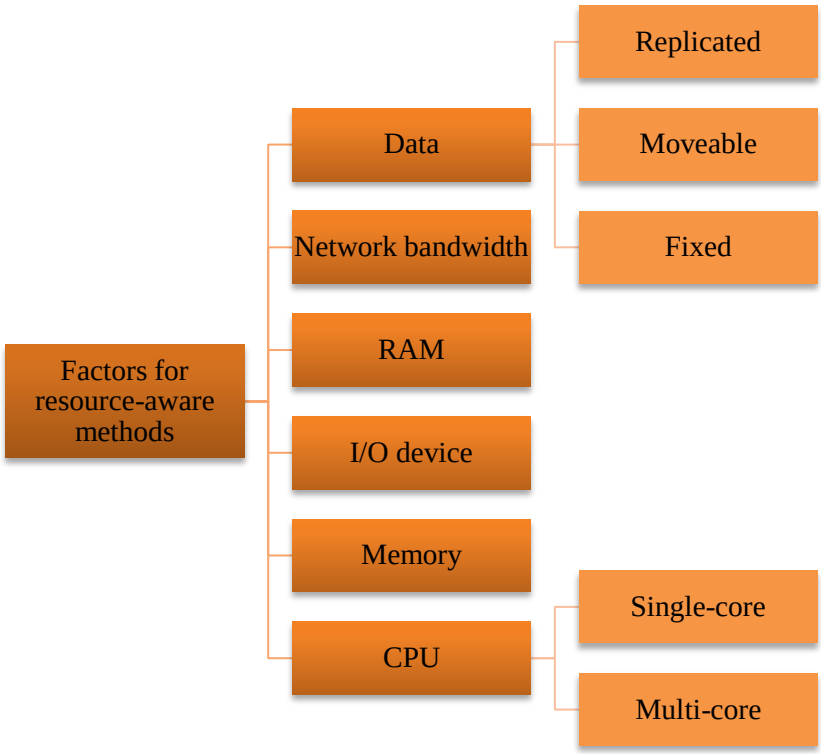


Fig. 3. Factors for resource-aware methods.

By utilizing virtualization technology extensively, data centers can optimize their resource management. The use of virtualization technology in cloud data centers optimizes the use of resources and minimizes operating costs. Cloud computing uses demand-driven resources such as VMs, to simplify the processing of complex duties. In addition, some VM allocation strategies attempt to place virtual machines at a minimum cost by considering various cost parameters. In this regard, as illustrated in Fig. 3, virtual machines or virtual data centers may be assigned different resources. Virtual machines are primarily equipped with a virtual CPU, which may either be single-core or multi-core. Virtual machines have access to crucial virtual resources, including network bandwidth and data. To optimize resource allocation for virtual machine applications, it is beneficial to place interacting virtual machines in close proximity. Additionally, the management of data, whether moveable, fixed, or replicated, depends on factors such as size and security policy.

Fig. 4 depicts the importance of considering traffic-related parameters, specifically the significant role of bandwidth, in improving the performance and efficiency of virtual machine allocation. Most approaches consider traffic between physical machines, traffic between interacting virtual machines, and traffic between virtual machines and their data repositories. Furthermore, as shown in Fig. 5, in the case of multiple-site clouds or geographically distributed clouds, a variety of factors will affect the location of the data center. Virtual machine allocation techniques, as shown in Fig. 5, depend on several factors, such as physical machine power consumption, number of physical machines, switches, and switch ports. Data centers are constructed based on specific topologies, taking into account these considerations.

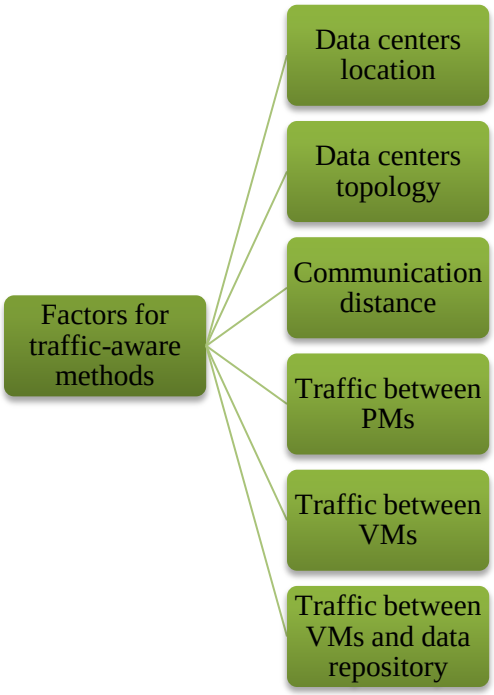


Fig. 4. Factors for traffic-aware methods.

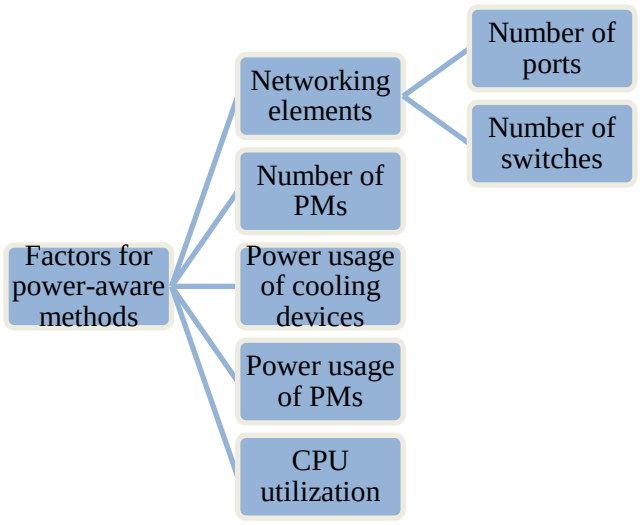


Fig. 5. Factors for power-aware methods.

Fig. 6 categorizes virtual machine allocation methods based on the availability of sites and clouds for virtual machine deployment. Both the single-cloud and multi-cloud models have the capability to utilize one or multiple sites to provide services. In the single-cloud model, services are offered from a single cloud provider's infrastructure, which may be composed of multiple sites or data centers. On the other hand, the multi-cloud model involves multiple cloud providers, each with their own sites or data centers, to deliver services. Various factors, including resources, location, performance, and cost, determine the selection of one or more sites. The aim is to ensure efficient service delivery, high availability, and scalability based on the specific needs of the applications or workloads. On the other hand, Fig. 7 classifies virtual machine allocation strategies in terms of load-aware parameters. These parameters are utilized by virtual machine allocation techniques to forecast and optimize virtual machine utilization.

Fig. 8 demonstrates the categorization of data-aware factors within a cloud data center, highlighting the significance of data characteristics in data-aware virtual machine allocation methods. Virtual machine allocation strategies are classified as dynamic and static. Virtual machine lifetime is the primary factor in static virtual machine allocation. The suitability of a physical machine to host a virtual machine is determined by its performance and longevity. During dynamic virtual machine allocation, the initial virtual machine assignment on PMs within a cloud data center can be modified based on certain factors, such as system load. Dynamic virtual machine assignment strategies can be categorized as reactive or proactive. During reactive virtual machine allocation, the initial placement changes when a particular undesirable state is reached. Modification of the initial placement of virtual machines is often necessitated by various factors such as SLA violations, load balancing, power consumption, and performance. Proactive virtual machine allocation methods proactively adjust the initial placement in advance of reaching a particular state or meeting specific demands. Virtual machine allocation techniques typically fall into one of two categories: allocation within a single cloud or allocation within a federated cloud.

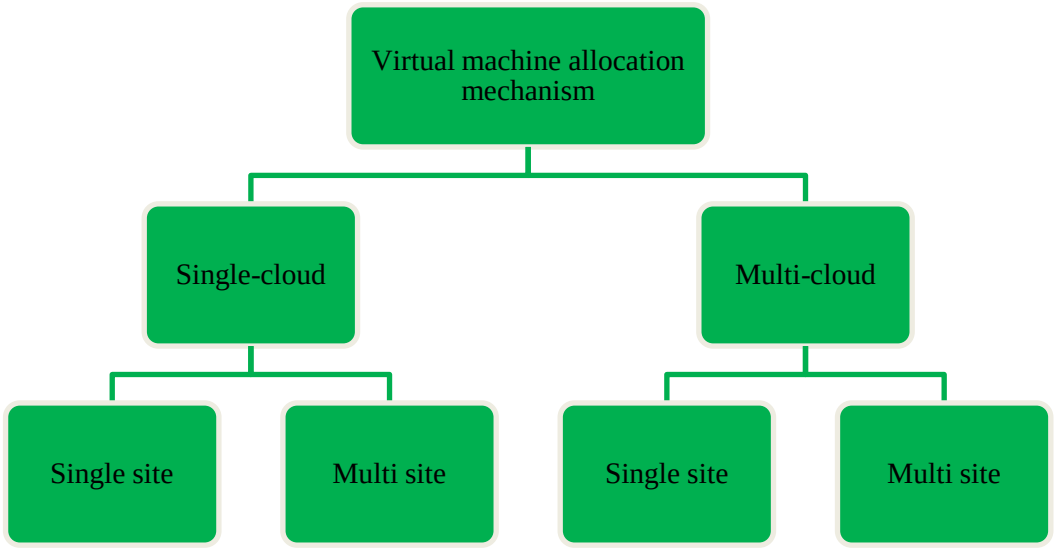


Fig. 6. Taxonomy of methods based on the number of clouds.

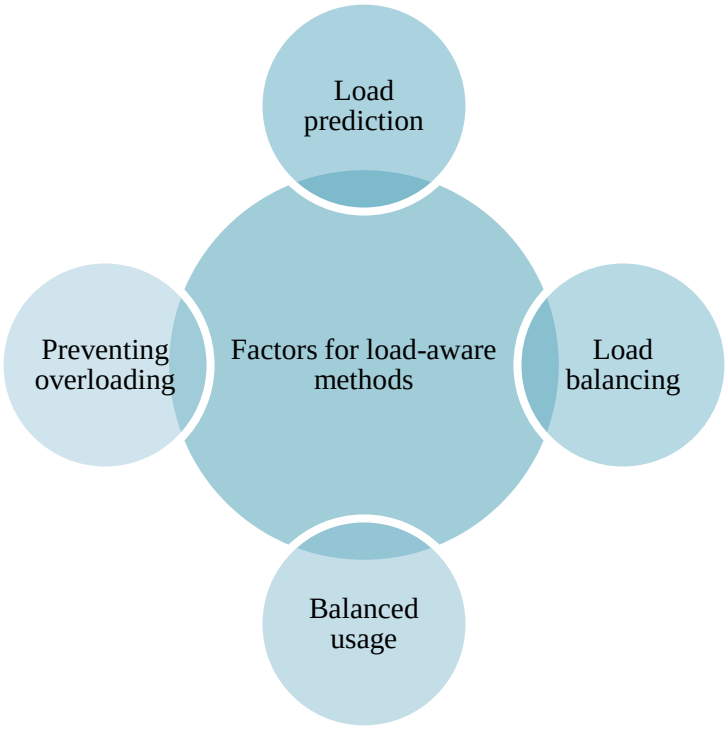


Fig. 7. Factors for load-aware methods.

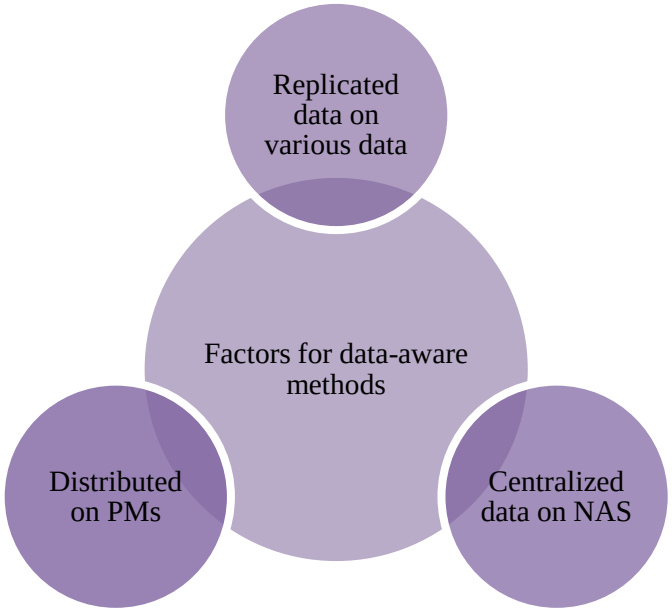


Fig. 8. Factors for data-aware methods.

Traffic in the cloud data center can be categorized into the inter-data center and intra-data center. These factors can be improved by optimizing the placement of virtual machines on physical machines. Additionally, achieving optimal allocation of virtual machines brings about significant enhancements such as increased availability, improved accessibility, higher performance, enhanced load balancing, reduced SLA violations, lowered costs in cloud data centers, decreased power consumption, and optimized resource utilization.

III. Review of the Selected Meta-Heuristic Algorithms

This section offers a detailed assessment of the selected virtual machine allocation methods, emphasizing their essential features, benefits, and drawbacks. We have conducted an analysis of nature-inspired meta-heuristic algorithms across various categories, including SA, PSO, HS, GA, FA, CSA, ABC, BBO, ACO, CSO, GRASP, SCA, PSOA, SFLA, and

TS. We examine the main characteristics and performance of these algorithms, allowing for a comprehensive understanding of their applicability to virtual machine allocation.

A. Ant Colony Optimization Algorithm

The ACO algorithm imitates how ants search for food. It has gained significant importance in addressing the virtual machine allocation problem in cloud computing. ACO employs a population of artificial ants that iteratively build solutions by depositing pheromone trails on a graph representation. These trails guide the ants to explore and exploit potential areas of the solution space. In the context of virtual machine allocation, ACO can effectively optimize resource allocation and load balancing and minimize energy consumption. By using the pheromone trails to represent the desirability of allocating virtual machines to specific hosts, ACO can intelligently guide the allocation process [21]. Due to its ability to find near-optimal solutions, adapt to dynamic environments, and deal with large-scale optimization problems, ACO provides a valuable tool for addressing VM allocation challenges in cloud computing.

Esnault, et al. [22] suggested a decentralized dynamic VM allocation method based on a Peer-to-Peer (P2P) network of PMs. The system utilizes a dynamic structure where neighborhood data is randomly exchanged among the PMs, eliminating the need for global system knowledge. This approach enables regular VM allocation within the local area, ensuring scalability even as the number of PMs increases. The neighborhood-randomized characteristics of this approach contribute to the convergence of the allocation, resulting in a performance that closely resembles a centralized one. Experimental results on the Grid'5000 testbeds demonstrate that this approach achieves global packing efficiency comparable to a centralized system, leading to increased scalability, resource utilization, and overall performance. This technique has the potential to disregard high SLA violations and create more PMs and migration overhead, which is important to keep in mind.

Feller, et al. [23] offered a novel dynamic VM allocation method to enhance resource utilization by reducing the number of running PMs according to the available resources. This is a distributed mechanism to solve two sub-problems: VM allocation optimization and PM status monitoring. The ACO algorithm has been applied to discover the best VM allocation solutions. Finally, the authors have shown that this technique on the real workload effects has improved performance and high scalability, reduced energy consumption and resource utilization, and fewer SLA violations. However, this method ignores high migration time and high migration overhead. Farahnakian, et al. [24] have suggested compatibility and integration of the ACO algorithm with optimal and balanced computing resources integrated with vector algebra. This technique minimizes resource wastage and energy consumption in virtual data centers by using the ACO-based server consolidation mechanism. Furthermore, this technique has low energy consumption, low time complexity, and high resource utilization. But this technique suffers from a high overhead VMs placement and low QoS.

Ferdaus, et al. [25] have proposed distributed system architecture to achieve a dynamic VM allocation method. The proposed method finds an optimal solution for a specified objective function. Energy consumption in data centers has decreased by assigning VMs to fewer active PMs while preserving QoS requirements. The technique incorporates inactive PMs and immigrants into a multi-objective model. Furthermore, the comparison has shown that this method reduced power consumption, increased performance, fewer SLA violations, and high resource utilization. However, this method ignores low scalability and high overhead migration.

Farahnakian, et al. [26] proposed a dynamic load balancing method based on the average load of the cloud data center. This technique has an algorithm for dynamic load balancing and VM allocation to decrease cloud data centers' energy utilization while preserving the desired QoS. In this method, the ACO-based VM allocation method is based on action artificial ants to allocate VMs to decrease the number of active PMs while satisfying the current resource requirements and also uses a dynamic threshold calculation approach to calculate the load of all active PMs and change the higher threshold value according to necessity. The comparison analysis has revealed that the method employing a static threshold scheme has effectively reduced the overall migrations of VM and power usage. The ant colony optimization approach gives a better result in combining dynamic threshold-based VM consolidation. Furthermore, low energy consumption, high QoS, and low SLA violation are some of the advantages of this method.

Matre, et al. [27] have offered a novel migration overhead-aware VM allocation mechanism based on the ACO algorithm for reducing asset usage for large info centers, computational overhead, and power usage. Data center elements and applications are also damaged by dynamic VM consolidation over VM-aware migrations. Therefore, a method for estimating migration overhead is suggested, incorporating pragmatic migration factors. Also, in this technique, to improve the system's scalability for large-scale data centers, a hierarchical, decentralized integration framework has been developed using localization of VM integration functions and minimizing their network impact. Nevertheless, this technique ignores QoS and migration time.

Ashraf and Porres [28] have presented a new workload-aware VM allocation method to reach efficient VM allocation on the basis of the MMAS (Max-Min Ant System) by applying multidimensional time-varying workloads. This method has used the perfect caseload models and multidimensional assets being heuristic variables, and the SLA model has been exploited to guarantee customer satisfaction. Also, this method examines both relations between workloads and the influence of resource-imbalanced usage, which decreases the server counts and resource waste. The proposed method offers several advantages, including resource utilization and high performance, ensuring efficient utilization of available resources in the cloud environment. Additionally, it aims to provide high QoS by optimizing VM allocation and minimizing resource wastage. Another significant advantage is the reduction in energy consumption, contributing to improved

sustainability and cost-effectiveness of cloud computing systems.

Malekloo, et al. [29] presented an energy-aware VM allocation method employing the ACO algorithm to minimize energy consumption in data centers. This method takes into account the energy consumption associated with VM migration as a key factor. As a result, it greatly decreases dynamic PMs and relocations number, leading to savings in energy utilization while ensuring QoS requirement are met through SLAs. The technique effectively decreases the migration rate and minimizes total energy waste in the network. It offers high adaptability, QoS, resource utilization, and low energy consumption. However, this method may face challenges in terms of scalability and overhead migration, which could impact its applicability in larger and more complex environments.

Aryania, et al. [30] proposed a multi-objective approach for VM allocation utilizing double thresholds and the Ant Colony System (ACS). This method focuses on VM consolidation and employs double thresholds to determine the conditions for consolidation. The authors use a mapping relation between VMs and hosts, treating the hosts as food sources, and optimize this mapping relationship through a multi-stage consolidation process based on ACS. By utilizing distributed search and the cooperation of artificial ants, the approach aims to achieve an optimal global mapping relationship between virtual machines and hosts. The performance of the proposed method is assessed by utilizing actual workload data and employing dual thresholds for CPU utilization. These thresholds help determine whether a host is experiencing an overload or underload condition. When a host falls into either of these categories, the process of VM consolidation is initiated. This consolidation process involves implementing specific policies based on the load status of the host, which in turn enables the migration of selected VMs to destination hosts using ACS. This method has some drawbacks, such as high migration time and limited scalability, which may affect its suitability for larger and more complex environments.

Xiao, et al. [31] proposed a new method called ELM_MPACS, which combines the multi-population ACS algorithm with Extreme Learning Machine (ELM) prediction. This method offers a lower complexity compared to other approaches. The ELM_MPACS algorithm operates in two steps. Firstly, it estimates the state of hosts using ELM, which helps determine whether a host is over-loaded or under-loaded. In the second step, if a host is found to be over-loaded, the virtual machine on that host is migrated to a normal host.

On the other hand, if a host is under-loaded, the virtual machine is consolidated onto another under-loaded host with higher utilization. The concentration of pheromones in the algorithm increases the likelihood of a combination being selected in the future. ELM_MPACS uses multiple populations to choose the ideal solution each time, increasing diversity and ensuring convergence. Experimental results have shown several positive outcomes, including improved resource utilization, reduced SLA violations, decreased energy consumption, and shorter migration times. This method provides promising results in terms of resource management, SLA compliance, energy efficiency, and migration efficiency.

B. Biogeography-based Optimization Algorithm

The Biography-based Optimization (BBO) algorithm is a nature-inspired metaheuristic algorithm used for solving optimization problems. It draws inspiration from the behavior of biographical organisms and applies it to various domains, including virtual machine allocation in cloud computing. BBO employs population-based techniques to efficiently explore the solution space, optimize resource utilization, and minimize energy consumption. Its adaptive nature makes it well-suited for dynamic cloud environments, leading to improved performance and resource efficiency [32].

Zheng, et al. [33] introduced a novel method for solving the VM allocation problem using the BBO algorithm. Their approach combines a Multi-objective BBO algorithm with Differential Evolution (MBBO/DE). In this method, each habitat represents a placement solution, where a vector of length n corresponds to n VMs mapped to m servers. Initially, the habitats of MBBO/DE are created using a random function to ensure diversity. Then, all habitats are evaluated using a cosine model to calculate the migration rate for each habitat. The next step involves migrating Similar Individual Vectors (SIVs) through the differential evolution strategy, where habitats in MBBO/DE share their SIVs, resulting in the creation of new habitats. Mutated SIVs are generated using a Gaussian model, producing potentially new optimal solutions. Multiple objectives are measured, and non-dominated sorting is employed to evaluate the effectiveness of each solution in MBBO/DE. Finally, a greedy one-to-one deletion mechanism based on differential evolution is used to select better habitats. If a new habitat outperforms the reserved elite habitat, it replaces the previous habitat; otherwise, the previous habitat remains. This method offers advantages such as low power consumption, reduced migration time, and improved resource utilization. However, it has some drawbacks, including lower Quality of Service (QoS), limited scalability, and higher overhead migration.

Shi, et al. [34] introduced a new multi-objective optimization method that builds upon the classical BBO algorithm to fix the allotment issue of the VM. The method proposes enhancements in the form of enhanced Cosine migration and mutation models aimed at improving the efficiency of attaining optimal solutions. By incorporating these enhancements, the algorithm aligns better with the actual VM allocation scenario. Furthermore, the algorithm focuses on optimizing three key objectives: migration resource overhead, load balancing, and server power consumption. Thus, the proposed improvements in the migration and mutation models are geared towards efficiently achieving these optimization goals in the context of VM allocation.

Zheng, et al. [35] have suggested a new approach for the VM allocation problem based on a multi-objective BBO algorithm in cloud-based environments. In this technique, after the initial parameters of the system are determined, the within-subsystem immigration is based on the ranking of the island within-subsystem, so the elementary population is the number of subsystems multiplied by the number of islands per

subsystem. Then, the subsystem migration is based on the likeness level among the subsystem pairs. The next step concerns the probability of mutation of each island, so if an island is selected for the mutation, the SIV is randomly selected from the island and replaced with a new randomly generated SIV. Finally, a modified non-dominant ranking system is used to select Nelite elites from the optimal answers of each subsystem and the elite of the previous generation. Then, if the repetition doesn't end, a new population is created for the next generation, and the Nelite matrices are replaced by the elitists generated in the last stage. Therefore, reduced power consumption, increased resource utilization, and high performance are the advantages of the method. However, this method suffers from high overhead migration, high time migration, and high SLA.

C. Artificial Bee Colony Algorithm

The Artificial Bee Colony (ABC) algorithm is a population-based metaheuristic optimization technique miming honeybee foraging behavior. It employs a diverse population of artificial bees, including employed bees, onlookers, and scout bees, to explore the solution space. The employed bees exploit food sources based on local information, while onlookers choose food sources based on quality. Scout bees introduce diversity by randomly searching for new food sources. Through iterative search processes, the ABC algorithm effectively balances exploration and exploitation, making it suitable for solving various optimization problems [36].

In their work, Jiang, et al. [37] have introduced an innovative VM consolidation policy that utilizes the ABC algorithm in conjunction with the Data-intensive Energy Evaluation Model. The objective of this policy is to minimize energy consumption costs and improve the SLAV value. By employing the ABC algorithm, the VM migration decisions are optimized, resulting in enhanced VM allocation. Moreover, the authors propose an energy efficiency evaluation model specifically designed for data-intensive tasks within green data centers. The model takes into account two influential factors: the CPU and GPU interest rates, which are observed through the behavior of data-intensive tasks in data centers. The VM allocation policy, based on the ABC utility, is then employed to minimize energy consumption and achieve satisfactory SLAV values. This approach effectively reduces energy consumption, the number of VM migrations, and SLAs while maximizing resource utilization and maintaining a high quality of service. However, it is important to note that this technique overlooks scalability and the overhead involved in the migration process.

Li, et al. [38] have introduced a novel approach called energy-aware dynamic virtual machine consolidation (EC-VMC) for VM migration. This method aims to address the constraints related to potential overloading across multiple sources. It employs a set of algorithms for selecting and placing virtual machines, taking into account the limitations of different sources within a physical machine while considering additional overhead. Additionally, the algorithm integrates and simulates the foraging behaviors of an artificial bee colony, treating the mapping relationship between PMs and VMs as a food source. By utilizing the search mechanism and optimization strategy of the ABC algorithm, the optimal mapping relationship is achieved globally, considering the multi-source constraints between PMs and VMs. Experimental results demonstrate that this method effectively reduces the VM's relocations number and power usage while ensuring large resource utilization and performance. However, it should be noted that this approach faces challenges related to overhead migration and scalability.

D. Chicken Swarm Optimization Algorithm

Chicken Swarm Optimization (CSO) is a swarm intelligence algorithm introduced in 2014 by Meng et al. It is a randomized optimization algorithm inspired by the group behavior of chickens searching for food. The CSO algorithm categorizes the chickens into roosters, hens, and chicks based on their fitness levels. Each type of chicken employs a different searching strategy, and the entire chicken swarm updates itself over multiple generations. One of the key strengths of the CSO algorithm is its ability to prevent getting trapped in local optima and instead find the optimal global solution for a given optimization problem. By mimicking the cooperative foraging behavior of chickens, the algorithm explores the search space effectively [39]. Tian et al. [40] have presented a new algorithm for Virtual Machine Consolidation (VMC) based on the Chicken Swarm Optimization model with deadlock-free migration (DFM-CSO). In this method, the consolidation VM design, which turns the VMC problem into a vector packing optimization problem according to deadlock-free migration, minimizes energy consumption. This algorithm is specified by the "one-step look-ahead with n VMs migration in parallel (OSLA-NVMIP)" method, which validates the VM migration and resets the aim physical host, also records the migration order for each solution placement so that VM transfer can be done according to the migration sequence. Furthermore, the algorithm reduces energy consumption, improves resource utilization, and reduces migration times. However, it exhibits low performance, low scalability, and a high migration overhead.

E. Cuckoo Search Algorithm

The Cuckoo Search Algorithm (CSA) is a metaheuristic optimization algorithm inspired by the behavior of cuckoo birds. It simulates the cuckoos' parasitic breeding manner, where they put their eggs in the nests of different bird species. In the CSA, each solution is represented as a cuckoo egg, and the nests represent potential solutions. The algorithm employs three key steps: (1) Lévy flights, which simulate the random walk of cuckoos, are used to generate new solutions; (2) some eggs are randomly selected for replacement to introduce diversity; and (3) the quality of eggs in the nests is assessed using an objective function, and the nests with higher quality solutions are more likely to survive. By iteratively repeating these steps, the CSA effectively explores the solution space and converges toward optimal or near-optimal solutions for optimization problems [41].

Joshi and Kaur [42] have proposed a new approach to solving the virtual machine consolidation issues based on a cuckoo search to save both power consumption and resource utilization in cloud space. The cuckoo search has been used to solve the problem of consolidation of the VM, which is similar to the bin packing problem. Therefore, in this method, it is

shown that the cuckoo search algorithm is a useful way to solve bin packing problems compared to other heuristics algorithms and can be efficiently used for the VM consolidation approach for both resource utilization and power consumption in the cloud network. This approach enables great source employment, less power usage, and great quality of service, reducing the number of PMs. Nevertheless, this method ignored overhead migration, time migration, scalability, and SLA violations.

Naik, et al. [43] have presented a new approach using a multi-objective Fruit-fly hybridized Cuckoo Search algorithm for virtual machine consolidation in the cloud environment. This algorithm is based on VM migration, which is used to reduce the over-provisioning of a PM by consolidating VMs on an underutilized physical machine. Thus, it relies on VM migration plans, which reduce the migration of PMs, consolidate under-utilized PMs, and select migrating VMs based on threshold significance. Multi-objective criteria should be used instead of single-objective criteria for placing and integrating VMs. The best host is detected from the host list. Compared with other methods, this method uses less power and resources, and it also reduces the relocations number. However, scalability, service quality, and violations of service level agreements were ignored.

F. Firefly Algorithm

The Firefly algorithm (FA), developed by Yang in 2008, was inspired by the flashing behavior of fireflies and uses swarm intelligence. Fireflies are also integrated to attract one another based on brightness. Therefore, light is considered a factor for absorption for flashing fireflies, and less brightness leads to a lighter situation. With increasing distance, attractiveness decreases directly with brightness. The goal function could be equated to the light to get the optimal solution [44].

Perumal and Murugaiyan [45] have proposed a novel for server consolidation and virtual machine placement problems that use a firefly colony and fuzzy firefly colony algorithms. The purpose of this method is to reduce the number of PMs utilized and solve the problem of VM disposition to achieve the right method with lower energy usage and lower waste of resources. Therefore, in this approach, both mentioned goals are considered simultaneously in the procedure of figuring out the ideal arrangement for setting up the virtual machine. One of the benefits of using fuzzy firefly colonies to solve virtual machine placement problems is that, unlike other algorithms, this algorithm works quickly and reduces randomization when searching for the optimal solution so that it will perform well. The fuzzy firefly colony method is provided with a rule of fuzzy probability to handle the behavior of the firefly probe with uncertainty. This technique reduces power consumption, maximizes resource utilization, and enhances performance. However, it ignores the number of migrations, SLA, migration time, and scalability.

John and Bindu [46] have proposed two new methods, the first being exploratory energy and temperature-based integrated temperature (HET-VC) and the second, virtual machine-based firefly and temperature-based integration (FET-VC). These methods are used to find the best solution in the problem space in improving the integration problems of virtual machines. Besides, in the proposed method, the current position and the next position after the migration of virtual machines will be updated at regular intervals. This method tries to decrease the number of transmissions and SLA violations, as well as decrease the energy consumption of servers. Besides, these two algorithms use high-performance servers that use the least CPU and RAM, as well as low temperatures, to integrate virtual machines. It suffers from high system overhead, long migration times, and limited scalability.

G. Genetic Algorithm

Genetic Algorithm (GA) was introduced in 1975 as a population-based optimization method with an evolutionary process. GA has been widely used to optimize VM integration parameters and also to solve optimization problems. In the GA algorithm, each chromosome represents a possible solution, and these solutions are combined by a string of genes. A fitness function is also used to check if the chromosome for the environment is right, and then crossover and mutation operations are used to produce offspring for the new population. The fitness function is used to assess the quality of each offspring, and this action is repeated until sufficient offspring are manufactured. Quick convergence and high processing time are the disadvantages of this algorithm [47].

Joseph, et al. [48] has proposed a novel technique to allocate virtual machines using the Family Gene approach. In this method, the host list and VM list are used for optimal mapping, and then the whole process is distributed among different families running in parallel in this module using the FGA module. Also, by using a self-adjusting mutation operator, an attempt has been made to reduce the rate of early convergence in this approach. Simple mutations are made to construct families. The resulting chromosomes, which are slightly different from each other, are placed in a family. Each family is processed "k" times. In the absence of a better individual being found by then, the family will be destroyed, and the next family will be taken in their place. In this way, the quality of each individual is determined by the amount of fitness value associated with that individual. Physical resource utilization is the purpose of the proposed method. In order to accomplish this, each individual's performance is evaluated so that if the individual fails, the impossible solution becomes feasible. This method performs well regarding energy consumption, resource utilization, and VM migrations. Nevertheless, it endures large upward and limited versatility.

Wu and Ishikawa [49] have presented an Energy-aware method for dynamic VM consolidation to reduce the energy consumed in heterogeneous cloud environments. In this method, the migration cost of the virtual machine and the amount of energy savings through the dynamic integration of the virtual machine in heterogeneous cloud data centers are investigated. Thus, a general assessment is defined by two opposing goals, namely, the cost of migration and energy saving. For this purpose, in this method, a merge score function has been arranged for total assessment based on the immigration cost estimation method and a high-limit estimation method for maximum saving power. Therefore, the IGGA algorithm, a kind of genetic grouping algorithm, is designed to enhance the merging score by a greedy heuristic

and an exchange action. The offered method offers better scalability, migration cost, QoS, power consumption, and resource utilization. But it ignores migration overhead, migration time, and SLA violation.

Wu, et al. [50] have presented a new method based on the evolutionary computing Adaptive Genetic Algorithm (A-GA) for the VM consolidation approach. This method has good performance in issues related to virtual machine consolidation in the areas of dynamism in resource utilization, VM migration number, failure reduction, SLA breach, and energy efficiency. Therefore, the suggested system can be used to manage large-scale cloud resources, reduce energy consumption, and increase QoS. Furthermore, the Minimum Migration Time (MMT) by VM selection policy based on VM placement or allocation has performed better than the Maximum Correlation (MC) and Random Selection (RS) method for energy utilization and QoS in minimal failure time and SLA violation. Furthermore, less power usage, large source employment, less relocation duration, less SLA violation, and high QoS are some of the advantages of this algorithm. However, high migration overhead and low scalability are some of the weaknesses of this algorithm.

Theja and Babu [51] have proposed an improved energy-saving asset allotment on the basis of a genetic method taking the power usage in cooler systems and IT items into account. They have focused on the problem of resource allocation and have evaluated various important issues in the cloud network, namely power and energy usage, VM's relocations number, and SLA violations. Also, the primary importance of considering various system parameters such as CPU, RAM, and network bandwidth in the decision-making process in their work has been evaluated. Most importantly, a comprehensive multi-objective policy has been used as a solution to the problem of resource allocation based on genetic algorithms to simultaneously reduce the power consumption of IT equipment and CRAC units.

Arianyan, et al. [52] have proposed a new approach with mixed-integer nonlinear programming (MINLP) formula for virtual machine integration problems in cloud computing that uses the genetic algorithm (GA) named energy and cost-aware VM consolidation or (ECVMC). They designed a mathematical model to reduce power consumption and costs using effective VM consolidation. In this method, by decreasing the number of active servers, the use of resources is maximized and minimizes total power consumption. Besides, this method decreases the placement process cost by discovering the optimal solution. As a result, low cost, low power consumption, and high resource utilization are some of the benefits of this method.

H. Greedy Random Adoptive Search Procedure Algorithm

The Greedy Random Adaptive Search Procedure (GRASP) algorithm was presented in 1989 to solve combinatorial problems by Feo and Resende. The GRASP is a meta-heuristic exploration based on a structural exploration that can create various initial solutions. This is a repetitive random optimization method in which each iteration involves two steps: one construction step and one local search improvement step. The construction step is an iterative greedy and adaptive process, and the second phase is a method of improving the local search for the initial solution, both of which are repeated for the maximum number of GRASP iterations. In the greedy construction stage, the list of solutions is made on the basis of greedy performance by randomly demoing the resolution habitat. In the second stage, a local search is performed to detect the best present solution from the formerly created solution list [53].

Ilager, et al. [54] have suggested a new method for a dynamic consolidation framework for the comprehensive controlling of cloud sources through enhancing cooling and computing systems. Via the Energy and Thermal-Aware Scheduling (ETAS) algorithm, they have controlled the trade among aggressive consolidation and the scattered distribution of VMs, which affects energy and hotspots. Also, the ETAS algorithm is designed to handle the trade-off between expense and time efficiency and could be adjusted as needed. Besides, based on the needs of the system, this algorithm is customizable to manage computation time and solution quality. Extensive experiments have been performed using real-world traces with precise thermal and power samples on this algorithm. Furthermore, this technique has advantages such as high QoS, less power usage, large source employment, and less relocation duration. However, this algorithm ignored scalability and performance.

I. Harmony Search Algorithm

The Harmony Search (HS) algorithm is a metaheuristic optimization technique inspired by the improvisation process of musicians in a musical ensemble. It mimics the harmony creation process, where musicians adjust their musical notes to achieve pleasing melodies. In the HS algorithm, a population of solution vectors represents musical harmony, and each element of the vector corresponds to a decision variable. Initially, random solutions are generated, and a fitness function evaluates their quality. Through iterations, new solution vectors are created by considering three main operators: pitch adjustment, harmony memory consideration, and randomization. These operators guide the search process towards better solutions. By continuously refining the harmony vectors based on their fitness values, the HS algorithm aims to find the optimal or near-optimal solution for the given optimization problem. The algorithm has demonstrated effectiveness in solving various complex optimization problems [55].

Fathi and Khanli [56] have suggested a new method based on the HS algorithm for the active allotment of VMs, which has been proven to be effective in power management systems. To solve the problem of virtual machine allocation, this algorithm has been used to detect an optimal solution. It defines a multi-objective function that takes into account both the number of silent PMs and the number of migrations and data center power consumption in each state. Also, some of the benefits of this method result in faster results, such as no need for basic parameters and no need to extract data. This method dramatically reduces energy utilization, resulting in fewer active PMs while maintaining QoS service quality standards. Furthermore, this algorithm has low SLA violations, low VM migration counts, high QoS, and low migration times. However, it ignores scalability and resource utilization.

Kim, et al. [57] have proposed a new method using a grouping agent approach and a meta-heuristic harmonic search algorithm for effectively solving the virtual machine allocation problem. In this algorithm, to solve VMC, the population number should be increased in proportion to increasing the size of the problem to certify search variety. In this case, the cost of searching for a solution may increase, so knowing that the number of VMs in virtualization is greater than the number of PMs, VMCs are targeted in such environments. Reflecting on this feature, the length of the solution will be reduced if the PM-based solution agent is classified instead of the VM-based redesign. Therefore, as the size of the problem increases, population growth becomes more linear and efficient search results. In this method, VMs have a unique workload pattern because they are provided in real-time using network resources. When migrating a virtual machine, depending on the VM time-series pattern as well as the number of resources allocated, the PM could be stabled by maintaining a good PM-VM map. Therefore, some advantages of this algorithm are low migration cost, low power consumption, high QoS, and high performance. However, this method suffers from high SLA violations, high overload migration, and low scalability.

J. Sine-Cosine Algorithm

The sine-cosine Algorithm (SCA) is an optimization method based on a population that performs the optimization function by applying a set of random solutions. These solutions are computed by a target function repeatedly in each iteration. In this method, the random set is repeatedly measured by a target function and improved by using a set of rules that is the core of this optimization method. The likelihood of global optimization increases with the right number of random solutions. In the optimization algorithms with enough random solutions and optimization steps (iterations), the possibility of discovering the global optimum increases [58].

Jayasena, et al. [59] have proposed a multi-objective approach based on the Multi-objective Sine Cosine Algorithm (MOSCA) for virtual machine allocation problems. This method is presented using a Multi-objective Evolutionary Algorithm based on Decomposition (MOEAD) and a Non-Dominated Sorting Genetic Algorithm (NSGAII). This approach generally focuses on evaluating and comparing multi-objective algorithms to find the optimal solution, as well as developing the MOSCA to figure out the resolution for the proposed VMC model. In this method, three conflicting goals discussed for the high energy efficiency of VMC are to reduce power consumption, achieve better SLAVs, and maximize MTBHS time. Therefore, the advantages of this algorithm are low power consumption, high resource utilization, low SLA violation, and high QoS. However, this method endures less versatility, high upward systems, and high migration time.

K. Penguin Search Optimization Algorithm

Penguins Search Optimization Algorithm (PSOA), introduced by Gheraibia and Moussaoui [60], is a meta-heuristic algorithm based on a common penguin hunting strategy. The algorithm is inspired by the penguin's search strategy, as they can combine their attempts and synchronize their dives to optimize global energy within the collective looking out and nutrition strategy. Each of them is represented by its place as well as the number of fish consumed. The distribution of penguins is determined by the number of prior fish in the area. Penguins are divided into groups, and they search in different situations. If the result of the number of dives is successful, Penguins return to the land and share important points such as places and a set of available food resources.

Jayasena, et al. [59] have presented a new method, according to the PSOA, for consolidating multiple virtual machines in the distribution of cloud-based hosting surroundings. This method uses the PSOA algorithm to make an economic VM consolidation and also supports scheduling for multiple VMs with different applications. The system can allocate VM resources for applications, reducing the number of VMs required. Therefore, the current strategy, by automatically planning and implementing several programs simultaneously with the specified features, provides the desired thinking in a very dedicated hosting space. Therefore, this algorithm, using placement rules, can lead to the Optimization of various multi-objective problems. Finally, this algorithm provides low power consumption and high QoS and reduces the number of PMs. However, this method ignored scalability and SLA violations, as well as resource utilization.

L. Particle Swarm Optimization Algorithm

The Particle Swarm Optimization (PSO) algorithm was introduced in 1995 by Kenney and Eberhart to find suitable solutions to continuous space optimization problems. The PSO algorithm uses a set of potentially random solutions (particles) to discover proper solutions to optimization problems. For each particle, the speed required to move in the search space is allocated. In each iteration, the speed of each particle is adjusted according to its best position and the position of the best particle in the total population. In this algorithm, each particle tracks its coordinates in the problem space with the optimal solution (fitness) obtained because each particle has internal memory. This algorithm has other benefits, such as combining local search methods with global search methods and trying to balance exploration and exploitation [61].

Dashti and Rahmani [62] have proposed a new method with a hierarchical architecture to meet the needs of producers and consumers, as well as a new service for scheduling consumer tasks in the PaaS layer. They have improved the allocation of dynamic resources and gained more benefits in the Personal Data Center by using the PSO algorithm. Therefore, PSO is used to ensure the quality of user services and reduce energy efficiency in redistributing migratory virtual machines in the overloaded host. In this method, low-load hosts and their power are consolidated to save energy. Also, there are balanced overload hosts used to ensure the quality of service during response times and deadlines. Therefore, this method offers less power usage, great QoS, and high source employment but ignores SLA violations and migration time.

Li, et al. [63] have presented a new approach based on several assets and energy-saving relocation and VMs' integration method under dynamic load and have arranged a double-threshold algorithm with multi-resource utilization for the VMs migration. One of the common problems of traditional heuristic algorithms in the field of virtual machine

consolidation is falling into local optima, which is prevented by using the Modified Particle Swarm Optimization (MPSO) method. Therefore, this method can reduce energy utilization with the QoS guarantee. Finally, the advantages of this algorithm are less power usage, great source employment, great QoS, and a reduced number of PMs. However, high system overhead and high migration time are some of the disadvantages of this algorithm.

M. Simulated Annealing Algorithm

The Simulated Annealing (SA) algorithm mimics the annealing process in metallurgy, in which materials are heated and slowly cooled to reduce defects and improve structure. In SA, the search process starts with an initial solution and iteratively explores the solution space by randomly making changes. These changes can be either accepted or rejected based on a probability determined by a cooling schedule. Initially, the algorithm accepts more changes, resembling a high-temperature phase, allowing for exploration [64]. As the temperature decreases, the algorithm becomes more selective, favoring changes that lead to improved solutions. This transition from exploration to exploitation helps the algorithm escape local optima and converge towards global optima. The cooling schedule controls the balance between exploration and exploitation. SA has been widely used to solve combinatorial optimization problems and has shown effectiveness in finding near-optimal solutions, even in complex search spaces [65].

Marotta and Avallone [66] have proposed a new model for solving consolidation problems using the SA algorithm, which solves these problems by evaluating the attractiveness of possible VM migrations. In this method, the attractiveness of virtual machine migration is considered through the use of resources and the energy efficiency of physical servers. Therefore, to reduce energy waste, the consolidation of the virtual machine in live VM migration is used. Increasing the overall cost efficiency by decreasing the number of active nodes is one of the main goals of this method. Therefore, the advantages of this method are low energy consumption, high resource utilization, and high performance. However, this method ignores the scalability count of migration and time migration.

Rajabzadeh and Haghighat [67] have proposed a new method based on an energy-aware framework for the consolidation of VM to optimize energy consumption and SLA violation reduction. In this approach, the whole process of allocating and managing virtual machines is divided into smaller parts, and each of these small parts is improved by new algorithms or existing algorithms. First, the hosts are determined by the critical mode. In the first stage, the host overload status is classified once with the possibility of violating the SLA and again without considering it. Then, in the second stage, using this algorithm, virtual machines are selected to migrate from important hosts. Therefore, using the SA algorithm, according to the list of selected virtual machines for migration, new hosts are selected as the migration destination of virtual machines. Eventually, the low-load hosts are identified and selected by the low-load host selection algorithm, and then all the virtual machines deployed on these hosts migrate, and the host goes to sleep off. All steps in this method are done in the distributed mode, except for the VM's disposition, which is done in the centralized model. Finally, the advantages of this method are low energy consumption and high resource utilization, low SLA violation and low migration time, and high performance. However, high overload and low scalability are some of the weaknesses of this algorithm.

Telenyk, et al. [68] have proposed a new optimization method using the SA to solve dynamic virtual machine integration problems, an extension of the bin-packing problem. In this system, to provide a new configuration, the optimized objective function of the proposed Simulated Annealing algorithm is used. This approach uses temperature as a control parameter when exploring to optimally map the allocation of virtual machines for the objective function. At each stage of the algorithm evolution, using the search function, the virtual machine allocation map becomes a new neighborhood state. The acceptance indicator of each new VM allocation map is compared to the current allocation map indicator to decide whether it is accepted or not. Therefore, the advantages of this method are low SLA violation, high resource utilization, and low energy consumption. However, high system overload and this method suffer from low scalability, low QoS, and high migration time.

Telenyk, et al. [69] suggested a new approach to dynamic VM allocation based on the SA algorithm so that it can calculate the cost of migration by a consolidation program. In this algorithm, the overhead cost resulting from the live migration of the virtual machine is calculated using an estimated model. The goals of this algorithm are to minimize the total energy consumption and minimize the total migration overhead. A static threshold-based method can be used to identify underloaded hosts in this approach. In this method, by selecting a host as the source PM, all its host virtual machines are selected for migrated to the out. Lastly, the SA is used to select the destination host. Thus, this technique has reduced energy consumption, living migration costs, and SLA violations, as well as the overall migration overhead. This method, however, neglected migration times, quality of service, and scalability.

N. Shuffled Frog Leaping Algorithm

The Shuffled Frog Leaping Algorithm (SFLA) was proposed by Eusuff and Lansey as an intelligent optimization algorithm that mimics natural organisms' behavior. It adopts a population-based approach and cooperative search paradigm to tackle discrete optimization problems. The SFLA incorporates a shuffling approach, facilitating the exchange of information between local search and global Optimization [70]. In this algorithm, frogs in the population symbolize tasks, and the positions of the frogs denote the mapping between VMs and their respective tasks.

Luo, et al. [71] suggested a new scheme based on energy-aware resource allocation for dynamic consolidation of VMs. This approach focuses on the infrastructure in which custom VMs run on the unsuitable servers of a data center as a service-oriented model. To properly manage resources, by achieving dynamic consolidation of virtual machine host resources and changing the mode of idle or low-consumption hosts to energy-saving mode, it is possible to use more energy resources and adhere to SLAs. Thus, a new hybrid intelligent algorithm using

the Modified SFLA based on Extreme Optimization (EO) called (MSFLA-EO) is used to quickly and efficiently complete the dynamic allocation of VMs. This method increases the SFLA frog's leaping visibility and improves local search capability. Finally, low energy consumption, high performance, reduced migration cost, and high resource utilization are some of the advantages of this algorithm.

O. Tabu Search Algorithm

Tabu Search (TS) is a metaheuristic algorithm developed by Glover to address mixed optimization problems like the bin packing one. TS utilize local searching to avoid being trapped in local optima and continue the search until the desired results are satisfied. The algorithm incorporates random selection to enhance the efficiency of searching and employs a "Taboo List" to restrict search movements and avoid cycles. The Taboo list serves as a short-term memory, keeping track of recently explored solutions [72]. Nasim and Kassler [73] introduced a novel approach that addresses the trade-off between resource conflict protection and the additional energy costs associated with increased server loads while addressing uncertainties in VM resource demand. This method utilizes a TS-based method coupled with greedy heuristics to explore local issues and repetitive resistance. The proposed method demonstrates the capability to achieve near-optimal solutions through robust discoveries and accurately address time-constrained online data optimization.

Moreover, the approach can be customized to accommodate varying resistance levels required by data center operators. The benefits of this algorithm include reduced power consumption, low resource utilization, and minimal VM migrations. However, it should be noted that this method has some drawbacks, such as lower performance and limited scalability.

IV. DISCUSSION

Meta-heuristic algorithms play a crucial role in the allocation of VMs in cloud environments. This problem involves efficiently allocating VMs to physical servers in cloud infrastructure, aiming to optimize resource utilization, energy consumption, and overall system performance. Cloud environments are complex and changing, making it difficult to find the best solutions. Nature-inspired meta-heuristic algorithms draw inspiration from natural phenomena, biological systems, and evolutionary processes to solve complex optimization problems.

Table I outlines the key features, advantages, and limitations of the discussed algorithms. These algorithms offer unique advantages for tackling the VM allocation problem in cloud environments. Each algorithm has unique traits and demonstrates gradual advancements. For example, GA demonstrates significant progress in its capacity to tackle intricate challenges, demonstrating gradual improvements in exploring various arrangements for effective VM placement. FA reveals gradual improvements in achieving a trade-off between exploration and exploitation, leading to faster convergence even in dynamic contexts despite initial difficulties. In addition, ACO and PSO algorithms demonstrate improvements in their capacity to handle dynamic environments, gradually addressing the scaling issues found in their previous generations. The incremental solutions emphasize the progressive path of meta-heuristic approaches, depicting their potential to adapt and improve in dealing with the complexities of VM allocation in the constantly changing field of cloud computing.

- Global search capability: Meta-heuristic algorithms excel in exploring a vast solution space, allowing them to identify promising solutions even in highly complex and large-scale cloud environments. These algorithms, such as genetic algorithms, PSO, and ACO, utilize search strategies inspired by natural systems to navigate the search space and find good solutions efficiently.
- Robustness and adaptability: Cloud environments are dynamic, with varying workloads and resource demands. Nature-inspired algorithms possess inherent robustness and adaptability, enabling them to handle the dynamic nature of the virtual machine consolidation problem. They can quickly adapt to changes in the cloud environment, such as VM migrations, workload fluctuations, and server failures, and reoptimize the VM-to-server allocation.
- Parallelism and scalability: Nature-inspired algorithms are inherently parallelizable, allowing them to leverage the distributed nature of cloud environments. Parallelism makes it possible for algorithms to manage large-scale cloud systems with multiple VMs and servers, enhancing scalability and speeding up the optimization process.
- Exploration-exploitation trade-off: Balancing exploration and exploitation is necessary for solving the virtual machine allocation problem. Nature-inspired algorithms inherently possess exploration-exploitation mechanisms, ensuring a balance between exploring the search space for new configurations and exploiting promising solutions to optimize VM placement.

By leveraging the strengths of nature-inspired meta-heuristic algorithms, cloud providers can achieve efficient and optimized VM allocation, leading to improved resource utilization, reduced energy consumption, and enhanced overall system performance. These algorithms offer powerful tools for addressing the complexities of virtual machine allocation in cloud environments and contribute to advancing cloud computing technology.

TABLE I. COMPARISON OF META-HEURISTIC ALGORITHMS FOR VM ALLOCATION IN CLOUD COMPUTING

Algorithm	Key Features	Advantages	Limitations
Ant Colony Optimization (ACO)	Pheromone trails Inspiration from ant foraging behavior	Good exploration robustness to dynamic environments	Slow convergence Scalability issues
Biogeography-Based Optimization (BBO)	Migration and evolution-inspired algorithm	Consideration of migration rates, habitat suitability	Parameter tuning required Scalability issues
Artificial Bee Colony (ABC)	Inspired by the bee foraging behavior	Fast convergence Simplicity	May get trapped in local optima Sensitivity to parameters
Chicken Swarm Optimization (CSO)	Inspired by the chicken foraging behavior	Simplicity Good exploration	Convergence speed may be slow Performance variation
Cuckoo Search Algorithm (CS)	Inspired by cuckoo bird behavior	Fast convergence Able to escape local optima	Parameter sensitivity Scalability issues
Firefly Algorithm (FA)	Inspired by firefly flashing behavior	Exploration-exploitation balance Fast convergence	Slow convergence Difficulty with dynamic environments
Genetic Algorithm (GA)	Genetic operators: selection, crossover, mutation	Good exploration Able to handle complex problems	Selection of suitable operators Parameter tuning
Greedy Random Adoptive Search Procedure (GRASP)	A combination of greedy and random search	Simplicity Fast convergence	May get trapped in local optima
Harmony Search Algorithm (HS)	Music-inspired algorithm with improvisation and memory considerations	Good exploration Robustness to noisy environments	Parameter tuning required Slow convergence
Sine-Cosine Algorithm (SCA)	Inspired by sine and cosine functions	Simplicity Fast convergence	Sensitivity to parameters Difficulty with complex problems
Penguin Search Optimization Algorithm (PSOA)	Inspired by the hunting behavior of penguins	Good exploration Robustness to dynamic environments	Limited scalability Parameter sensitivity
Particle Swarm Optimization (PSO)	Particle movement and social interaction-based algorithm	Fast convergence Good exploration	Difficulty in balancing exploration and exploitation
Simulated Annealing (SA)	Inspired by the annealing process in metallurgy	Good exploration Able to escape local optima	Parameter tuning required Slower convergence
Shuffled Frog Leaping Algorithm (SFLA)	Simulates the frog leaping behavior	Good exploration Able to escape local optima	Slow convergence Scalability issues
Tabu Search Algorithm (TS)	Uses tabu list to avoid revisiting recently visited solutions	Good exploration Able to escape local optima	May get stuck in suboptimal solutions Parameter tuning

V. OPEN ISSUES

- Resource Utilization: One open issue is optimizing the utilization of computing resources in cloud data centers. Dynamic workload patterns pose a challenge in allocating virtual machines to maximize resource utilization and ensure efficient task execution. Efficient allocation and placement algorithms are needed to adjust resource allocation based on workload variations dynamically.
- Energy Efficiency: Another open issue is the improvement of energy efficiency in VM allocation and placement. In cloud data centers, energy consumption is a major concern, and efficient allocation of VMs can help reduce overall power usage. Developing algorithms and techniques that consider energy consumption metrics and prioritize energy-efficient VM placement is crucial for achieving sustainable and green cloud computing environments.
- Quality of Service (QoS): Ensuring QoS requirements for applications running in cloud environments is a critical challenge. VM allocation and placement algorithms should consider factors such as response time, latency, throughput, and reliability to meet the performance expectations of users. Balancing resource allocation for various applications and providing adequate QoS guarantees is an ongoing concern in cloud computing.
- Security and Privacy: VM allocation and placement involve sensitive data and require protection against security threats. Ensuring secure communication, data integrity, and privacy preservation during the allocation process is a significant challenge. Robust security mechanisms and privacy-preserving techniques need to be integrated into VM allocation algorithms to mitigate risks and protect the confidentiality and integrity of user data.
- Scalability: As cloud computing environments continue to grow in size and complexity, scalability becomes an open issue in VM allocation and placement. Efficient allocation is crucial when handling a large number of VMs, hosts, and data center resources. Developing scalable algorithms that can handle the increasing scale of cloud environments is essential for smooth and efficient VM allocation and placement.
- Multi-objective Optimization: VM allocation and placement involve conflicting objectives, including enlarging asset usage, minimizing power usage, and meeting QoS requirements. Multi-objective optimization techniques that can balance these objectives and provide trade-off solutions are necessary. Developing algorithms that can handle multiple optimization objectives and consider the preferences and constraints of different stakeholders is an open research area in VM allocation and placement.

- Dynamic and Real-time Allocation: Real-time allocation of VMs to adapt to dynamic workload changes and user demands is an ongoing challenge. VM allocation and placement algorithms should be able to handle sudden workload spikes, failures, or changes in resource availability. Designing dynamic and adaptive allocation strategies that can efficiently respond to real-time changes in the cloud environment is crucial for maintaining performance and responsiveness.

VI. CONCLUSION

The VM migration process necessitates the allocation of VMs in cloud computing. Its objective is to identify the optimal PM for each VM. With an increase in the number of VMs deployed in data centers, it becomes harder to control the utilization rate of host PMs. This study emphasized the importance of meta-heuristic algorithms to develop effective approaches to VM allocation. A description and evaluation of the algorithms in each group are presented based on a number of relevant criteria, including resource utilization, number of PMs and VMs, migration time, energy consumption, migration overhead, and SLA violation. The results of the research indicate that no meta-heuristic algorithm has been developed to allocate VMs capable of meeting all the requirements. The review has also identified several challenges and limitations associated with existing methods. These include scalability issues, high computational overhead, and the need for improved adaptation to dynamic environments. Furthermore, the lack of standardization and benchmarking frameworks for evaluating the performance of different algorithms remains a critical concern.

REFERENCES

[1] B. Pourghebleh, A. A. Anvigh, A. R. Ramtin, and B. Mohammadi, "The importance of nature-inspired meta-heuristic algorithms for solving virtual machine consolidation problem in cloud environments," Cluster Computing, pp. 1-24, 2021.

[2] K. Prasanna Kumar and K. Kousalya, "Amelioration of task scheduling in cloud computing using crow search algorithm," Neural Computing and Applications, vol. 32, no. 10, pp. 5901-5907, 2020.

[3] K. N. Qureshi, G. Jeon, and F. Piccialli, "Anomaly detection and trust authority in artificial intelligence and cloud computing," Computer Networks, vol. 184, p. 107647, 2021.

[4] O. Ali, A. Shrestha, J. Soar, and S. F. Wamba, "Cloud computing-enabled healthcare opportunities, issues, and applications: A systematic review," International Journal of Information Management, vol. 43, pp. 146-158, 2018.

[5] V. Hayyolalam, B. Pourghebleh, M. R. Chehrehzad, and A. A. Pourhaji Kazem, "Single - objective service composition methods in cloud manufacturing systems: Recent techniques, classification, and future trends," Concurrency and Computation: Practice and Experience, vol. 34, no. 5, p. e6698, 2022.

[6] A. A. Khan, M. Zakarya, R. Buyya, R. Khan, M. Khan, and O. Rana, "An energy and performance aware consolidation technique for containerized datacenters," IEEE Transactions on Cloud Computing, vol. 9, no. 4, pp. 1305-1322, 2019.

[7] A. Yousafzai et al., "Cloud resource allocation schemes: review, taxonomy, and opportunities," Knowledge and information systems, vol. 50, pp. 347-381, 2017.

[8] A. Peivandizadeh and B. Molavi, "Compatible authentication and key agreement protocol for low power and lossy network in IoT environment," Available at SSRN 4194715, 2022.

[9] B. Pourghebleh and V. Hayyolalam, "A comprehensive and systematic review of the load balancing mechanisms in the Internet of Things," Cluster Computing, pp. 1-21, 2019.

[10] M. Khodayari, J. Razmi, and R. Babazadeh, "An integrated fuzzy analytical network process for prioritisation of new technology-based firms in Iran," International Journal of Industrial and Systems Engineering, vol. 32, no. 4, pp. 424-442, 2019.

[11] M. Mohseni, F. Amirghafouri, and B. Pourghebleh, "CEDAR: A cluster-based energy-aware data aggregation routing protocol in the internet of things using capuchin search algorithm and fuzzy logic," Peer-to-Peer Networking and Applications, pp. 1-21, 2022.

[12] M. Momeni, D.-C. Wu, A. Razban, and J. Chen, "Data-driven Demand Control Ventilation Using Machine Learning CO2 Occupancy Detection Method," 2020.

[13] B. M. Jafari, M. Zhao, and A. Jafari, "Rumi: An Intelligent Agent Enhancing Learning Management Systems Using Machine Learning Techniques," Journal of Software Engineering and Applications, vol. 15, no. 9, pp. 325-343, 2022.

[14] S. R. Abdul Samad et al., "Analysis of the Performance Impact of Fine-Tuned Machine Learning Model for Phishing URL Detection," Electronics, vol. 12, no. 7, p. 1642, 2023.

[15] S. Vairachilai, A. Bostani, A. Mehbodniya, J. L. Webber, O. Hemakesavulu, and P. Vijayakumar, "Body Sensor 5 G Networks Utilising Deep Learning Architectures for Emotion Detection Based On EEG Signal Processing," Optik, p. 170469, 2022.

[16] W. Anupong et al., "Deep learning algorithms were used to generate photovoltaic renewable energy in saline water analysis via an oxidation process," Water Reuse, vol. 13, no. 1, pp. 68-81, 2023.

[17] S. Aghakhani, A. Larijani, F. Sadeghi, D. Martín, and A. A. Shahrakht, "A Novel Hybrid Artificial Bee Colony-Based Deep Convolutional Neural Network to Improve the Detection Performance of Backscatter Communication Systems," Electronics, vol. 12, no. 10, p. 2263, 2023.

[18] M. Bagheri, "Clustering Individual Entities Based on Common Features," 2021.

[19] D.-C. Wu, M. Momeni, A. Razban, and J. Chen, "Optimizing demand-controlled ventilation with thermal comfort and CO2 concentrations using long short-term memory and genetic algorithm," Building and Environment, vol. 243, p. 110676, 2023.

[20] S. P. Rajput et al., "Using machine learning architecture to optimize and model the treatment process for saline water level analysis," Journal of Water Reuse and Desalination, 2022.

[21] M. Dorigo, M. Birattari, and T. Stutzle, "Ant colony optimization," IEEE computational intelligence magazine, vol. 1, no. 4, pp. 28-39, 2006.

[22] A. Esnault, E. Feller, and C. Morin, "Energy-aware distributed ant colony based virtual machine consolidation in IaaS Clouds bibliographic study," Informatics Mathematics (INRIA), pp. 1-13, 2012.

[23] E. Feller, C. Morin, and A. Esnault, "A case for fully decentralized dynamic VM consolidation in clouds," in 4th IEEE International Conference on Cloud Computing Technology and Science Proceedings, 2012: IEEE, pp. 26-33.

[24] F. Farahnakian et al., "Energy-aware dynamic VM consolidation in cloud data centers using ant colony system," in 2014 IEEE 7th International Conference on Cloud Computing, 2014: IEEE, pp. 104-111.

[25] M. H. Ferdaus, M. Murshed, R. N. Calheiros, and R. Buyya, "Virtual machine consolidation in cloud data centers using ACO metaheuristic," in Euro-Par 2014 Parallel Processing: 20th International Conference, Porto, Portugal, August 25-29, 2014. Proceedings 20, 2014: Springer, pp. 306-317.

[26] F. Farahnakian et al., "Using ant colony system to consolidate VMs for green cloud computing," IEEE Transactions on Services Computing, vol. 8, no. 2, pp. 187-198, 2014.

[27] P. Matre, S. Silakari, and U. Chourasia, "Ant colony optimization (ACO) based dynamic VM consolidation for energy efficient cloud computing," International Journal of Computer Science and Information Security, vol. 14, no. 8, p. 345, 2016.

[28] A. Ashraf and I. Porres, "Multi-objective dynamic virtual machine consolidation in the cloud using ant colony system," International Journal of Parallel, Emergent and Distributed Systems, vol. 33, no. 1, pp. 103-120, 2018.

[29] M.-H. Malekloo, N. Kara, and M. El Barachi, "An energy efficient and SLA compliant approach for resource allocation and consolidation in cloud computing environments," Sustainable Computing: Informatics and Systems, vol. 17, pp. 9-24, 2018.

[30] A. Aryania, H. S. Aghdasi, and L. M. Khanli, "Energy-aware virtual machine consolidation algorithm based on ant colony system," Journal of Grid Computing, vol. 16, no. 3, pp. 477-491, 2018.

[31] H. Xiao, Z. Hu, and K. Li, "Multi-objective VM consolidation based on thresholds and ant colony system in cloud computing," IEEE Access, vol. 7, pp. 53441-53453, 2019.

[32] D. Simon, "Biogeography-based optimization," IEEE transactions on evolutionary computation, vol. 12, no. 6, pp. 702-713, 2008.

[33] Q. Zheng, J. Li, B. Dong, R. Li, N. Shah, and F. Tian, "Multi-objective optimization algorithm based on bbo for virtual machine consolidation problem," in 2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS), 2015: IEEE, pp. 414-421.

[34] K. Shi, H. Yu, F. Luo, and G. Fan, "Multi-Objective Biogeography-Based Method to Optimize Virtual Machine Consolidation," in Proceedings of the International Conference on Software Engineering and Knowledge Engineering, 2016, pp. 225-230, doi: https://doi.org/10.18293/SEKE2016-151.

[35] Q. Zheng et al., "Virtual machine consolidated placement based on multi-objective biogeography-based optimization," Future Generation Computer Systems, vol. 54, pp. 95-122, 2016.

[36] D. Karaboga and B. Akay, "A comparative study of artificial bee colony algorithm," Applied mathematics and computation, vol. 214, no. 1, pp. 108-132, 2009.

[37] J. Jiang, Y. Feng, J. Zhao, and K. Li, "DataABC: A fast ABC based energy-efficient live VM consolidation policy with data-intensive energy evaluation model," Future generation computer systems, vol. 74, pp. 132-141, 2017.

[38] Z. Li, C. Yan, L. Yu, and X. Yu, "Energy-aware and multi-resource overload probability constraint-based virtual machine dynamic consolidation method," Future Generation Computer Systems, vol. 80, pp. 139-156, 2018.

[39] X. Meng, Y. Liu, X. Gao, and H. Zhang, "A new bio-inspired algorithm: chicken swarm optimization," in Advances in Swarm Intelligence: 5th International Conference, ICSI 2014, Hefei, China, October 17-20, 2014, Proceedings, Part I 5, 2014: Springer, pp. 86-94.

[40] F. Tian, R. Zhang, J. Lewandowski, K.-M. Chao, L. Li, and B. Dong, "Deadlock-free migration for virtual machine consolidation using chicken swarm optimization algorithm," Journal of Intelligent & Fuzzy Systems, vol. 32, no. 2, pp. 1389-1400, 2017.

[41] Q. Yang, H. Huang, J. Zhang, H. Gao, and P. Liu, "A collaborative cuckoo search algorithm with modified operation mode," Engineering Applications of Artificial Intelligence, vol. 121, p. 106006, 2023.

[42] S. Joshi and S. Kaur, "Cuckoo search approach for virtual machine consolidation in cloud data centre," in International Conference on Computing, Communication & Automation, 2015: IEEE, pp. 683-686.

[43] B. B. Naik, D. Singh, A. B. Samaddar, and S. Jung, "Developing a cloud computing data center virtual machine consolidation based on multi-objective hybrid fruit-fly cuckoo search algorithm," in 2018 IEEE 5G World Forum (5GWF), 2018: IEEE, pp. 512-515.

[44] X.-S. Yang and X. He, "Firefly algorithm: recent advances and applications," International journal of swarm intelligence, vol. 1, no. 1, pp. 36-50, 2013.

[45] B. Perumal and A. Murugaiyan, "A firefly colony and its fuzzy approach for server consolidation and virtual machine placement in cloud datacenters," Advances in Fuzzy Systems, vol. 2016, 2016.

[46] N. P. John and V. Bindu, "Energy-Efficient Hybrid Firefly–Crow Optimization Algorithm for VM Consolidation," in Intelligent Computing and Communication: Proceedings of 3rd ICICC 2019, Bangalore 3, 2020: Springer, pp. 413-427.

[47] B. Alhijawi and A. Awajan, "Genetic algorithms: Theory, genetic operators, solutions, and applications," Evolutionary Intelligence, pp. 1-12, 2023.

[48] C. T. Joseph, K. Chandrasekaran, and R. Cyriac, "A novel family genetic approach for virtual machine allocation," Procedia Computer Science, vol. 46, pp. 558-565, 2015.

[49] Q. Wu and F. Ishikawa, "Heterogeneous virtual machine consolidation using an improved grouping genetic algorithm," in IEEE 17th International Conference on High Performance Computing and Communications, 2015: IEEE, pp. 397-404.

[50] Q. Wu, F. Ishikawa, Q. Zhu, and Y. Xia, "Energy and migration cost-aware dynamic virtual machine consolidation in heterogeneous cloud datacenters," IEEE transactions on Services Computing, vol. 12, no. 4, pp. 550-563, 2016.

[51] P. R. Theja and S. K. Babu, "Evolutionary computing based on QoS oriented energy efficient VM consolidation scheme for large scale cloud data centers," Cybernetics and Information Technologies, vol. 16, no. 2, pp. 97-112, 2016.

[52] E. Arianyan, H. Taheri, and S. Sharifian, "Multi Target Dynamic VM Consolidation in Cloud Data Centers Using Genetic Algorithm," Journal of Information Science & Engineering, vol. 32, no. 6, 2016.

[53] M. G. Resende and C. C. Ribeiro, "Greedy randomized adaptive search procedures: Advances, hybridizations, and applications," Handbook of metaheuristics, pp. 283-319, 2010.

[54] S. Ilager, K. Ramamohanarao, and R. Buyya, "ETAS: Energy and thermal - aware dynamic virtual machine consolidation in cloud data center with proactive hotspot mitigation," Concurrency and Computation: Practice and Experience, vol. 31, no. 17, p. e5221, 2019.

[55] M. T. Abdulkhaleq et al., "Harmony search: Current studies and uses on healthcare systems," Artificial Intelligence in Medicine, p. 102348, 2022.

[56] M. H. Fathi and L. M. Khanli, "Consolidating VMs in green cloud computing using harmony search algorithm," in Proceedings of the 2018 International Conference on Internet and e-Business, 2018, pp. 146-151.

[57] M. Kim, J. Hong, and W. Kim, "An Efficient Representation Using Harmony Search for Solving the Virtual Machine Consolidation," Sustainability, vol. 11, no. 21, p. 6030, 2019.

[58] S. Mirjalili, "SCA: a sine cosine algorithm for solving optimization problems," Knowledge-based systems, vol. 96, pp. 120-133, 2016.

[59] K. Jayasena, L. Li, M. Abd Elaziz, S. Xiong, and J. Xiang, "Optimizing the energy efficient VM consolidation by a multi-objective algorithm," in 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD)), 2018: IEEE, pp. 81-86.

[60] Y. Gheraibia and A. Moussaoui, "Penguins search optimization algorithm (PeSOA)," in Recent Trends in Applied Artificial Intelligence: 26th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2013, Amsterdam, The Netherlands, June 17-21, 2013. Proceedings 26, 2013: Springer, pp. 222-231.

[61] T. M. Shami, A. A. El-Saleh, M. Alswaitti, Q. Al-Tashi, M. A. Summakieh, and S. Mirjalili, "Particle swarm optimization: A comprehensive survey," IEEE Access, 2022.

[62] S. E. Dashti and A. M. Rahmani, "Dynamic VMs placement for energy efficiency by PSO in cloud computing," Journal of Experimental & Theoretical Artificial Intelligence, vol. 28, no. 1-2, pp. 97-112, 2016.

[63] H. Li, G. Zhu, C. Cui, H. Tang, Y. Dou, and C. He, "Energy-efficient migration and consolidation algorithm of virtual machines in data centers for cloud computing," Computing, vol. 98, no. 3, pp. 303-317, 2016.

[64] S. Mahmoudinazlou, A. Alizadeh, J. Noble, and S. Eslamdoust, "An improved hybrid ICA-SA metaheuristic for order acceptance and scheduling with time windows and sequence-dependent setup times," Neural Computing and Applications, pp. 1-19, 2023.

[65] K. Amine, "Multiobjective simulated annealing: Principles and algorithm variants," Advances in Operations Research, vol. 2019, 2019.

[66] A. Marotta and S. Avallone, "A simulated annealing based approach for power efficient virtual machines consolidation," in 2015 IEEE 8th international conference on cloud computing, 2015: IEEE, pp. 445-452.

[67] M. Rajabzadeh and A. T. Haghighat, "Energy-aware framework with Markov chain-based parallel simulated annealing algorithm for dynamic management of virtual machines in cloud data centers," The Journal of Supercomputing, vol. 73, no. 5, pp. 2001-2017, 2017.

[68] S. Telenyk, E. Zharikov, and O. Rolik, "Consolidation of Virtual Machines Using Stochastic Local Search," in Advances in Intelligent Systems and Computing II: Selected Papers from the International Conference on Computer Science and Information Technologies, CSIT 2017, September 5-8 Lviv, Ukraine, 2018: Springer, pp. 523-537.

[69] S. Telenyk, E. Zharikov, and O. Rolik, "Consolidation of virtual machines using simulated annealing algorithm," in 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), 2017, vol. 1: IEEE, pp. 117-121.

[70] B. B. Maaroof et al., "Current Studies and Applications of Shuffled Frog Leaping Algorithm: A Review," Archives of Computational Methods in Engineering, pp. 1-16, 2022.

[71] J.-p. Luo, X. Li, and M.-r. Chen, "Hybrid shuffled frog leaping algorithm for energy-efficient dynamic consolidation of virtual machines in cloud data centers," Expert Systems with Applications, vol. 41, no. 13, pp. 5804-5816, 2014.

[72] V. K. Prajapati, M. Jain, and L. Chouhan, "Tabu search algorithm (TSA): A comprehensive survey," in 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), 2020: IEEE, pp. 1-8.

[73] R. Nasim and A. J. Kassler, "A robust Tabu Search heuristic for VM consolidation under demand uncertainty in virtualized datacenters," in 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), 2017: IEEE, pp. 170-180.

Investigating the Effectiveness of ChatGPT for Providing Personalized Learning Experience: A Case Study

Raneem N. Albdrani[1], Amal A. Al-Shargabi[2]
Independent Researcher[1]
Department of Information Technology, College of Computer
Qassim University, Buraydah 51425, Saudi Arabia[2]

***Abstract*—The demand for personalized learning experiences that cater to the unique needs of individual learners has increased with the emergence of data science. This paper investigates the potential use of ChatGPT, a generative AI tool, in providing personalized learning experiences for data science education, specifically focusing on Deep Learning. The paper presents a case study that applies the 5Es model to test personalized learning for students using ChatGPT. The study aims to answer the question of how educators can leverage ChatGPT in their pedagogy to enhance student learning, and whether ChatGPT can provide a better learning experience than traditional teaching methods. The paper also discusses the limitations faced during the study and the findings. The results suggest that ChatGPT can be a valuable resource for data science education, providing personalized and instant feedback to learners. However, ethical considerations such as the potential for biased or inaccurate responses and the need for transparency in AI-generated content should be carefully addressed by educators. The study highlights ChatGPT's potential as a research tool for data science educators to investigate the effectiveness of AI in personalized learning experiences. Overall, this paper contributes to the ongoing dialogue on the role of AI in data science education and provides insights into how educators can utilize ChatGPT to enhance student learning and engagement.**

***Keywords*—Personalized learning; data science education; ChatGPT; generative AI**

I. Introduction

In recent years, technology has been transforming many aspects of our lives, and education is no exception. The rise of artificial intelligence (AI) and its applications in education, known as Artificial Intelligence in Education (AIED), have opened new possibilities for personalized and effective learning. The area of Artificial Intelligence in Education (AIED) has made significant progress in the last two and a half decades, particularly in the realm of technological advancements [1]. Moreover, the COVID-19 pandemic has accelerated the shift towards online and remote learning, increasing the demand for intelligent and adaptive learning technologies [2]. The potential for AI to revolutionize teaching and learning is immense, as it can automate tasks, process large quantities of data, and provide predictive insights. Intelligent Tutoring Systems (ITS), such as Beetle II, have demonstrated the potential of AI in education by using natural language processing and interactive experimentation to provide context-specific feedback to students and improve their learning outcomes[3].While the term "artificial intelligence" is broad and encompasses a range of technologies and techniques, the development of educational chatbots (EC) has garnered attention in the field of education. Built using Natural Language Processing (NLP), ECs offer immediate feedback to learners and assist in achieving educational and pedagogical objectives. Chatbots have been recognized for their potential in the realm of personalized learning, with Chen et al. [4] noting their benefits in this area. Additionally, Garcia Brustenga et al. [5] distinguished EC into two types: those with educational intentionality, which are designed to foster teaching and learning directly, and those without, which are incorporated into teaching tasks of an administrative nature. Colace et al. [6] argued that integrating chatbots in education can result in an engaging and interactive learning environment, similar to one-on-one interactions with teachers. The key for EC, in general, is personalized learning, as emphasized by Basham et al. [7], who state that personalized learning focuses on each student's individual skills and knowledge. It empowers learners to set goals, reflect on their progress, and extend their learning beyond the classroom [8].The research at hand centers around a particular chatbot-based AI platform, ChatGPT, which has attracted significant global interest and sparked a great deal of excitement among the public. Highlighting the uniqueness of ChatGPT as an AI platform and the potential impact it can have on personalized learning, it offers an advanced natural language processing system and the ability to provide tailored guidance and support to students. Through interactive conversations, ChatGPT creates a personalized learning experience that enables learners to receive immediate feedback, adapt their learning process, and explore knowledge in a flexible manner. This innovative AI platform holds great promise for enhancing personalized learning experiences and shaping the future of education. Despite the growing interest in the use of AI in education and the potential of chat-based AI tools like ChatGPT, there are still notable gaps and problems in existing studies. First, while some studies have explored the efficacy of personalized learning approaches, few have specifically investigated the integration of AI tools within this context. This leaves a gap in our understanding of how AI can be effectively harnessed to tailor educational experiences to individual learners. Second, limited research has specifically focused on the field of data science education and the potential benefits of AI-powered tools in this domain. This is particularly important given the rapidly evolving nature of data science and the increasing demand for skilled professionals in this field. Furthermore, while there have been studies examining the impact of AI

tools in education, there remains a lack of comprehensive exploration of the pedagogical methodologies and frameworks that can be successfully employed in conjunction with such tools. This gap hinders educators' ability to effectively implement AI tools like ChatGPT in their instruction and maximize their impact on student learning outcomes. To investigate the effectiveness of ChatGPT in providing personalized learning experiences, a case study approach with two groups, namely a control group and an experimental group, was adopted. The study involved a total of 20 students, with the experimental group having access to ChatGPT and learning through the guidance of the 5Es model, while the control group followed traditional learning methods. Data was collected to analyze the interactions and learning outcomes of both groups. This comparative analysis provides insights into the efficacy of ChatGPT as a personalized learning tool and its impact on student learning outcomes. The findings from this case study contribute to our understanding of the potential benefits and limitations of using ChatGPT in educational settings and inform the future development and implementation of AI-based educational technologies. In this study, the primary research objective was to investigate the effectiveness of ChatGPT in providing personalized learning experiences. The research questions that guided this study include: How does the use of ChatGPT impact student engagement and participation? What are the effects of ChatGPT on student learning outcomes? How do students perceive the personalized learning experience facilitated by ChatGPT? By addressing these research questions, this study aims to contribute to our understanding of the potential benefits and limitations of using ChatGPT as a personalized learning tool and inform the future development and implementation of AI-based educational technologies. By addressing the existing research gaps, and providing evidence-based recommendations on the effective usage of AI-enabled tools like ChatGPT, this study has the potential to enhance the student's learning experience, particularly in the field of data science education.The paper structure provides an overview of ChatGPT and its related studies, discusses personalized learning and AI in education, introduces the 5Es model, employs a methodology to test personalized learning using ChatGPT, presents and analyzes the study results, discusses limitations and implications, and draws conclusions emphasizing the importance of ChatGPT as a resource for data science education and ethical considerations. The paper contributes to the discourse on the role of AI in education and highlights how educators can use ChatGPT to enhance student learning and engagement.

II. CHATGPT

ChatGPT is an AI-based natural language processing model developed by OpenAI and launched in November 2022. It utilizes the GPT (Generative Pretrained Transformer) architecture that employs a transformer network to generate human-like text. This model has been extensively trained on a large volume of text data, enabling it to generate responses on a wide range of questions and topics. ChatGPT has been successful in achieving state-of-the-art performance on various language tasks, including language translation, text completion, and question answering. It can be used for various applications such as chatbots, language learning tools, and text-based virtual assistants [9]. ChatGPT has revolutionized the world of AI and chatbots and has gained popularity with people relying on it in many aspects of their lives, given its ability to mimic human-like conversations and follow up responses. The popularity of ChatGPT has been rapidly increasing over time, especially on social media platforms, where people are becoming more aware of the capabilities of AI. Since its release, ChatGPT has been inspiring people to create various AI tools, indicating the growing interest in the field of AI. As AI technology continues to advance, it is becoming increasingly important to understand ChatGPT's potential use and how to leverage its capabilities in the most effective and efficient ways possible. Biswas [10] has explored the potential use of ChatGPT in climate change research, suggesting that the model could be valuable in generating and analyzing different climate scenarios using diverse data inputs, ultimately enhancing the accuracy of climate projections. Additionally, Biswas [11] also discussed the potential use of ChatGPT in military applications, such as providing instant language translation and aiding in intelligence analysis. The model can also play a significant role in public health by supporting individuals and communities in making informed decisions about their health [12]. Numerous studies have explored the potential use of ChatGPT in education. Tlili et al. [13] conducted research on the subject, analyzing tweets to gain insights into public discussions about ChatGPT's role in education. Additionally, they conducted interviews with individuals who had used ChatGPT in an educational context and found that users considered ChatGPT to be highly valuable for revolutionizing education. However, some concerns were raised regarding its use. According to UNESCO [14], for teaching and learning ChatGPT can be used as collaboration coach, personal tutor, a study buddy and more. Furthermore, Kung et al. [15] discussed the ability of ChatGPT to perform clinical reasoning by testing its performance on questions from the United States Medical Licensing Examination (USMLE) and the potential of ChatGPT to assist in medical education. While Cooper [16] conducted a study that shows ways to use ChatGPT in science education and asked questions related to science to test ChatGPT's responses. In the realm of potential applications, Ray [17] has explored the versatility of ChatGPT across different fields. In the education sector, ChatGPT has shown promise in personalized learning, language acquisition, and test preparation. It has also demonstrated potential in domains such as healthcare, law services, content generation, programming, and marketing. The broad range of areas where ChatGPT can be utilized highlights its adaptability and the wide array of benefits it can offer across various industries. ChatGPT, like any AI technology, brings forth several challenges and ethical considerations when applied in education. One of the primary concerns is data privacy [17]. As ChatGPT interacts with students and collects data, ensuring the privacy and security of sensitive information becomes paramount. Educational institutions and developers must implement robust data protection measures to safeguard student data and comply with relevant privacy regulations. Another significant consideration is algorithmic bias. AI models like ChatGPT are trained on vast amounts of data, which can inadvertently perpetuate biases present in the training data [17] . This bias can manifest in responses or recommendations provided by ChatGPT, potentially reinforcing stereotypes or discriminatory practices. It is essential to address and mitigate algorithmic bias through careful data curation, diverse training datasets,

and ongoing monitoring of model outputs. Furthermore, over-reliance on AI without human guidance can pose challenges. While ChatGPT offers personalized learning experiences, it should not replace human teachers or mentors. The role of educators in guiding and contextualizing knowledge remains crucial. Striking a balance between leveraging AI tools like ChatGPT and maintaining human interaction and guidance is necessary for effective and holistic education.

III. Related Work

In this literature, we are providing precious works, researches that discussed personalized learning and the effect of using chatbots in education. As ChatGPT is still considered to be new to its field there's not yet much case studies or researches about using it in education. However, Tlili et al. [13] conducted a study on the impact of ChatGPT on education by analyzing three stages, including tweets and interviews with teachers and students, as well as investigating user experiences. The overall findings of the study were encouraging, with positive feedback on the effectiveness of ChatGPT in revolutionizing education and providing a comprehensive understanding of complex topics in a simple language. However, the researchers also highlighted some concerns that need to be addressed to ensure ChatGPT's efficiency in education. While Cooper [16] illustrated ways where ChatGPT can be helpful to generate ideas when designing science units, rubrics, and quizzes, in his study he interacted with ChatGPT and asked questions in which he considered the answers are extraordinary however he believes that there's a lack of evidence to the answers and misleading information. And in the context of educational chatbots in general Colace et al. [6] created a EC and tested its behavior for over than 180 students especially testing behavior, keeping track of progress, and assigning tasks, in this case study researchers tested three different situations Chatbot furnishes a correct suggestion, Chatbot furnishes a correct suggestion but it does not fit with the real needs of the student, and Chatbot furnishes a wrong suggestion the result was Correct Suggestion: 133 student, Correct Suggestionn but not suitable for the needs of the student: 30 student, and Wrong Suggestion: 24 student, the students found the EC user friendly and easy to use comparing to other chatbots. Furthermore Schmulian and Coetzee [18] developed two chatbots were developed to support personalized learning by fulfilling the role of a co-teacher in finding answers to commonly asked questions. This approach was aimed at addressing some of the challenges associated with teaching large groups of students. The study revealed that the chatbots were well-received, with 72 of the respondents expressing satisfaction and even affection towards the chatbots. Additionally, the study found that many students reported higher engagement levels when interacting with the chatbots as compared to a traditional face-to-face classroom environment. According to the study by Kumar and Silva [19], they addressed the challenge of student acceptance towards chatbots for personalized learning. They created a chatbot using Telegram and the results showed a positive response. The respondents indicated that chatbots have various advantageous features such as quick responses, accessibility, mobility, user-friendliness, human-like conversation, and private interaction. The study also highlighted that chatbots could be used for personalized feedback, guidance, peer-peer assessment, critique, and communication management. These automated attributes can lead to more effective interactions while maintaining privacy boundaries between instructors and students. The findings of the study suggest that chatbots can be a useful tool for learning, with potential future applications in education. In a study conducted by J. Pereira [20] involved creating a multiple-choice question (MCQ) chatbot to assist in the training of university students. The students were receptive to the bot and expressed interest in having similar tools integrated into other subjects. The researcher predicts that in the future, there will be a significant increase in the use of learning-oriented chatbots, which will offer personalized user experiences. Each student will have access to personal teacher assistants or learning coaches in the form of chatbots. Both ChatGPT and EC have a common foundation in personalized learning, which forms the fundamental concept of their educational approaches. I believe that personalized learning serves as the fundamental building block for both tools. According to Lee, Huh, Lin, and Reigeluth [21] believed that personalized learning is an effective approach for tailoring instruction to meet the individual needs and previous experiences of learners. This customized approach allows everyone to maximize their potential by receiving instruction that is specifically designed for them.

IV. Methodology

This research study aimed to investigate the effectiveness of ChatGPT in providing a personalized learning experience in the context of deep learning. The study employed a case study design with a mixed-methods approach, combining quantitative data from quiz results and qualitative observations. A total of 20 participants were involved in the study, comprising 10 students in the experimental group and 10 students in the control group. The participants were selected based on their interest in learning about deep learning and their varying levels of prior experience with ChatGPT. The participants were enrolled in an Internet of things class at Qassim university and had a basic understanding of machine learning concepts. The procedure consisted of the several steps, The first step was Lesson Preparation, I carefully prepared the lesson materials, ensuring that they covered the essential concepts and topics related to deep learning. The content included explanations of forward propagation, activation functions, convolutional neural networks, and network architecture. I also developed a set of multiple-choice quiz questions to assess the participants' understanding and knowledge. After preparing the lesson I instructed ChatGPT to provide tailored instructions and explanations to each student from the experimental group, based on their individual learning needs and prior knowledge. Drawing inspiration from Cooper [16], the researcher developed a teaching unit based on the 5Es Model, which served as the foundation for creating an automated lesson using this approach. The effectiveness of the 5Es Model was further supported by the findings of Sibel et al. [22], who demonstrated its positive impact in a relevant case study. As a result, the lessons were meticulously designed, aligning with the structure and principles of the 5Es Model, which involved engaging the students, exploring the concepts, explaining the principles, elaborating on the ideas, and evaluating their understanding. Throughout this process, ChatGPT assumed a pivotal role as a virtual assistant, actively supporting the students by addressing their queries, presenting pertinent examples, and guiding them throughout their learning

journey. Next stage Control Group, The control group received traditional instruction from a teacher who followed a standardized curriculum for teaching deep learning. The teacher delivered the content through lectures, and practical exercises, following a well-established approach to teaching the subject matter. The control group followed a structured lesson plan that aligned with the curriculum and did not involve the use of ChatGPT. After completing the lessons, both the experimental and control groups were given the same multiple-choice quiz. The quiz assessed the participants' understanding of deep learning concepts and their ability to apply the knowledge they had acquired. The quiz questions covered topics such as forward propagation, activation functions, convolutional neural networks, and network architecture. The quiz was administered in a controlled environment to ensure consistency across both groups. The last stage of the study is Data Collection and Analysis, The quiz results were collected and analyzed to compare the performance between the experimental and control groups. Quantitative analysis involved assessing the scores and identifying any significant differences in the performance of the two groups. The data from the quiz provided insights into the effectiveness of ChatGPT in facilitating personalized learning experiences in comparison to traditional instruction. Additionally, qualitative observations were made during the lessons to gather insights into the participants' engagement, interaction, and overall learning experience.

V. Result

A. Analysis of 5ES Model Implementation

Generally speaking, The analysis of the case study highlights both the advantages and limitations of using ChatGPT to guide students through the 5Es model. Despite some limitations, the 5Es model proved to be an effective framework for student engagement and learning. One of the limitations observed was that ChatGPT occasionally provided incorrect evaluations of the students' answers during the evaluation phase. This indicates a need for caution and verification when relying solely on ChatGPT's feedback. However, the 5Es model served as an effective guide for students. It helped structure the learning process and provided a clear framework for students to follow. The sequential progression through the engagement, exploration, explanation, elaboration, and evaluation phases allowed students to have a comprehensive learning experience. The personalized nature of ChatGPT as a tool was particularly beneficial. It served as a virtual co-teacher, offering guidance and support tailored to each student's needs. This individualized approach helped foster a sense of ownership and agency in the learning process. The quiz results played a significant role in reflecting the students' experience. Some students were able to answer all the quiz questions correctly, indicating that they successfully grasped the concepts and information provided by ChatGPT. However, there were instances where students made errors or skipped questions, suggesting the need for further clarification or reinforcement in certain areas. The integration of the 5Es model as a guide alongside ChatGPT was particularly valuable, especially considering that a majority of the students were unfamiliar with ChatGPT initially. The 5Es model provided structure and support, ensuring that students had a consistent and cohesive learning experience. The students who had prior familiarity with ChatGPT had a more positive experience overall. Nevertheless, it was important to have teacher supervision to ensure smooth progress and address any errors or deviations from the 5Es model that ChatGPT occasionally encountered. In certain instances, ChatGPT would repeat certain steps or omit them, necessitating the intervention of a teacher to correct and realign the learning process. In Fig. 1 of the lesson, it can be observed that ChatGPT progressed to the next step of the 5Es framework after the student joined the conversation and completed the first E. Subsequently, ChatGPT followed the same pattern for the subsequent E, as depicted in Fig. 2.

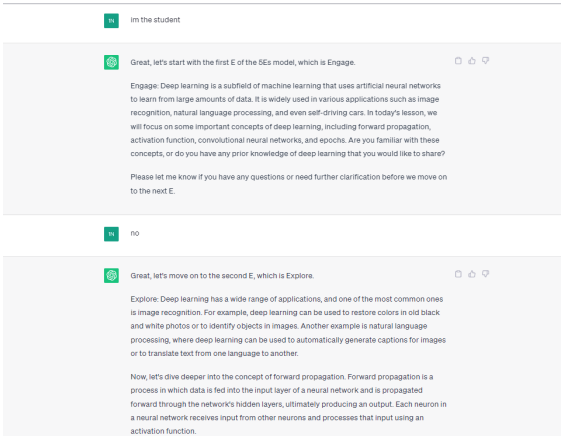


Fig. 1. ChatGPT lesson - Part 1.

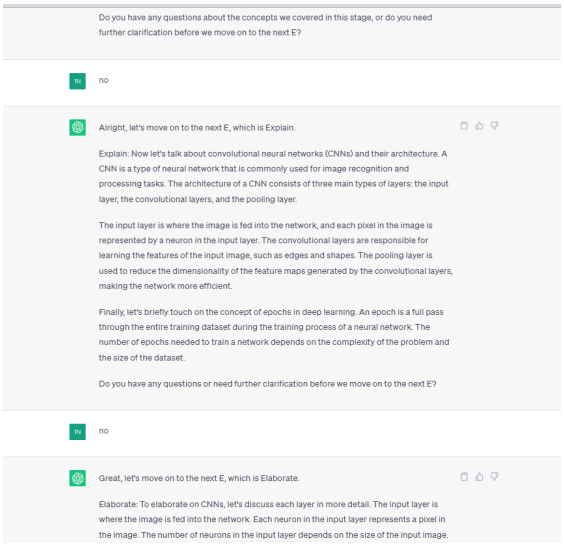


Fig. 2. ChatGPT lesson - Part 2.

In the final stage of the lesson, which was the Evaluate phase, ChatGPT effectively implemented it as a multiple-choice question (MCQ), following the instructions provided at the beginning of the lesson as shown in Fig. 3.

B. Analysis of Control Group by Teaching Traditional Way

The analysis of the control group's performance reveals important insights. Initially, the majority of students demonstrated a good understanding of the topic, as indicated by their correct

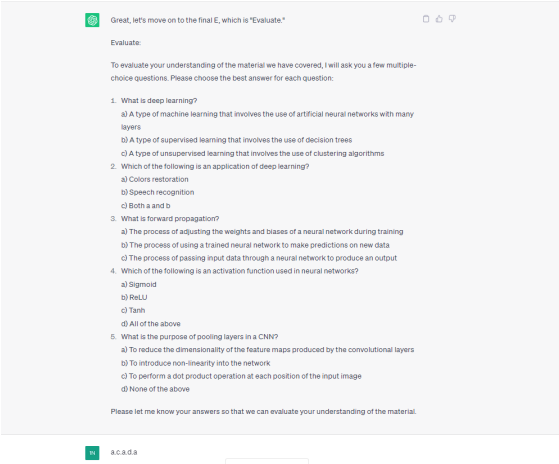


Fig. 3. ChatGPT evaluation quiz.

responses to the first question. However, there was a decline in the number of correct responses in subsequent questions, suggesting a potential decrease in comprehension over time. This raises concerns about the long-term retention of the material taught through traditional teaching methods, indicating that students may struggle to retain and apply the information beyond the immediate context. Furthermore, the varying number of correct responses across questions highlights differences in students' comprehension levels for different topics or concepts, emphasizing the need for further investigation and targeted teaching strategies. The presence of incorrect responses also signifies that the traditional teaching method may not effectively address the learning needs of all students, necessitating the exploration of alternative instructional approaches catering to diverse learning styles and abilities. These findings suggest that the traditional method of teaching has limitations in terms of long-term knowledge retention and ensuring uniform comprehension among students. To overcome these limitations, it is crucial to explore innovative teaching methods, incorporate active learning strategies, and provide personalized support to enhance student engagement and understanding. Further research and analysis are necessary to deepen our understanding of the effectiveness of traditional teaching methods and compare them with alternative approaches, ultimately leading to improved learning outcomes and student success.

VI. Discussion

The discussion in this study involved the qualitative analysis of the results from two groups, namely the experimental group and the control group, to examine the function of ChatGPT as an AI assistant in facilitating personalized learning experiences for students. While there are alternative approaches and methods that can be explored in future studies, our focus was on utilizing the 5Es model to address our research questions.

A. How Does the use of ChatGPT Impact Student Engagement and Participation?

The integration of ChatGPT in the experimental group, guided by the 5Es model, positively impacted student engagement and participation. The personalized nature of ChatGPT as a virtual co-teacher provided individualized guidance and support to students, fostering a sense of ownership and agency in the learning process. This individualized approach helped increase student engagement as they received tailored feedback and support, leading to active participation throughout the learning activities. The structured framework of the 5Es model, combined with ChatGPT's assistance, facilitated a more interactive and engaging learning environment for the students.

B. What are the Effects of ChatGPT on Student Learning Outcomes?

The implementation of ChatGPT alongside the 5Es model resulted in positive effects on student learning outcomes. The quiz results showed that a significant number of students in the experimental group were able to answer the questions correctly, indicating successful comprehension of the concepts and information provided by ChatGPT. The sequential progression through the engagement, exploration, explanation, elaboration, and evaluation phases of the 5Es model allowed students to have a comprehensive learning experience. The personalized guidance and support from ChatGPT contributed to improved learning outcomes by addressing individual learning needs, reinforcing understanding, and providing opportunities for active application of knowledge.

C. How do Students Perceive the Personalized Learning Experience Facilitated by ChatGPT?

The students in the experimental group perceived the personalized learning experience facilitated by ChatGPT positively. ChatGPT served as a virtual co-teacher, offering tailored guidance and support, which helped students feel supported and encouraged throughout the learning process. The individualized approach enhanced their learning experience, allowing them to progress at their own pace and receive immediate feedback. Students appreciated the interactive nature of ChatGPT and the opportunity to engage in dialogue with an AI-powered assistant. The integration of the 5Es model alongside ChatGPT provided a structured and cohesive learning experience, which students found beneficial in terms of clarity and organization.

Overall, the analysis suggests that the use of ChatGPT positively impacts student engagement, learning outcomes, and perception of personalized learning experiences. The integration of ChatGPT alongside the 5Es model provides an effective framework for student engagement, fosters active participation, enhances learning outcomes, and supports students in their individual learning journeys. However, it is important to acknowledge some limitations in this study. Firstly, the sample size of the experimental group may be relatively small, limiting the generalizability of the findings. Additionally, the study primarily focuses on short-term effects, and further research is needed to assess the long-term impact of ChatGPT on student learning and retention. Moreover, the reliance on technology and the potential for technical issues or errors in ChatGPT's feedback pose challenges that need to be considered. Despite these limitations, the findings highlight the promising potential of ChatGPT as a tool for enhancing student engagement and learning outcomes in the educational setting.

VII. Conclusions

This study employed a qualitative approach to investigate the effectiveness of ChatGPT, a chatbot tool, in education, particularly in the context of technology and personalized learning. By conducting a study consisting of two groups, namely the experimental group implementing the 5Es model with ChatGPT and the control group using traditional teaching methods, the analysis sheds light on the efficacy of different approaches in promoting student engagement, learning outcomes, and personalized learning experiences. The integration of ChatGPT with the 5Es model in the experimental group proved to be effective in engaging students and providing a structured learning experience. The sequential progression through the model's phases facilitated comprehensive learning, with ChatGPT serving as a valuable virtual co-teacher that enhanced student engagement and agency. However, the occasional incorrect evaluations by ChatGPT necessitated teacher supervision to ensure a smooth learning process. In contrast, the control group's traditional teaching method exhibited limitations in terms of knowledge retention and uniform comprehension. The declining trend in correct responses across questions indicated potential challenges in long-term retention, emphasizing the need for alternative instructional approaches and personalized support to accommodate diverse learning styles and abilities. These findings underscore the importance of exploring innovative teaching methods, incorporating active learning strategies, and providing personalized support to enhance student engagement and understanding. Further research is necessary to gain deeper insights into the effectiveness of traditional teaching methods compared to alternative approaches and to identify the most effective instructional strategies for improving learning outcomes and student success. In conclusion, the integration of ChatGPT with the 5Es model demonstrated promise in terms of engagement, structure, and individualized support, while the traditional teaching method highlighted the need for improvements in knowledge retention and comprehension. These findings contribute to the ongoing conversation about enhancing teaching and learning practices, ultimately aiming to create meaningful and effective educational experiences for students.

References

[1] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," *Autism*, vol. 47, no. 4, pp. 197–221, Oct. 2011, https://doi.org/10.1080/00461520.2011.611369.

[2] S. J. e. a. Ng, DTK; Leung, "Teachers' ai digital competencies and twenty-first century skills in the post-pandemic world." *Education Tech Research Dev*, vol. 71, no. 1, p. 137–161, Feb. 2023, https://doi.org/10.1007/s11423-023-10203-6.

[3] F. Dzikovska, Steinhauser, "Beetle ii: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics." *Int J Artif Intell Educ*, vol. 24, no. 3, p. 284–332, Sep. 2014, https://doi.org/10.1007/s40593-014-0017-9.

[4] H.-L. Chen, G. V. Widarso, and H. Sutrisno, "A chatbot for learning chinese: Learning achievement and technology acceptance." *Journal of Educational Computing Research*, vol. 58, no. 6, pp. 161–189, Jun. 2020, https://doi.org/10.1177/0735633120929622.

[5] G. Brustenga, Fuertes-Alpiste, and Molas-Castells, "Briefng paper: Chatbots in education." *Universitat Oberta de Catalunya.*, Sep. 2018, http://hdl.handle.net/10609/80185.

[6] F. Colace, M. D. Santo, M. Lombardi, and F. Pascale, "Chatbot for e-learning: A case study," *International Journal of Mechanical Engineering and Robotics Research*, vol. 7, no. 5, pp. 528–533, Sep. 2018, https://doi.org/10.18178/ijmerr.7.5.528-533.

[7] J. Basham, T. H. Hall, R. A. Carter, and W. M. Stahl, "An operationalized understanding of personalized learning," *Journal of Special Education Technology*, vol. 31, no. 3, pp. 126–136, Aug 2016, https://doi.org/10.1177/016264341666083.

[8] A. P. Susan Patrick, Kathryn Kennedy, "Mean what you say: Defining and integrating personalized, blended and competency education," *International Association for K-12 Online Learning*, Oct. 2013.

[9] OpenAI, "Chatgpt: Optimizing language models for dialogue." 2023.

[10] S. Biswas, "Potential use of chat gpt in global warming." *Ann Biomed Eng*, vol. 51, no. 6, p. 1126–1127, Mar. 2023, https://doi.org/10.1007/s10439-023-03171-8.

[11] B. S, "Prospective role of chat gpt in the military: According to chatgpt." *Qeios*, Feb. 2023, https://doi.org/10.32388/8WYYOD.

[12] S. Biswas, "Role of chat gpt in public health." *Ann Biomed Eng*, vol. 51, no. 5, p. 868–869, Feb 2023, https://doi.org/10.1007/s10439-023-03172-7.

[13] A. Tlili, B. Shehata, Adarkwah, and M.A., "What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education." *Smart Learn. Environ.*, vol. 10, no. 15, Feb 2023, https://doi.org/10.1186/s40561-023-00237-x.

[14] E. Sabzalieva and A. Valentini.;, "Chatgpt and artificial intelligence in higher education: quick start guide." https://eduq.info/xmlui/handle/11515/38828, year=2023.

[15] K. TH, C. M, M. A, and S. C. et al, "Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models." *PLOS Digital Health*, vol. 2, no. 2, 2023, https://doi.org/10.1371/journal.pdig.0000198.

[16] G. Cooper, "Examining science education in chatgpt: An exploratory study of generative artificial intelligence." *J Sci Educ Technol*, vol. 32, no. 3, p. 444–452, Mar 2023, https://doi.org/10.1007/s10956-023-10039-y.

[17] P. P. Ray, "Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope." *ScienceDirect*, vol. 3, pp. 121–154, 2023, https://doi.org/10.1016/j.iotcps.2023.04.003.

[18] A. Schmulian and S. A. Coetzee, "The development of messenger bots for teaching and learning and accounting students' experience of the use thereof." *Br J Educ Technol*, vol. 50, 2019, https://doi.org/10.1111/bjet.12723.

[19] J. A. Kumar and P. A. Silva, "Work-in-progress: A preliminary study on students' acceptance of chatbots for studiobased learning." *IEEE Global Engineering Education Conference*, pp. 1627–1631, 2020, https://doi.org/10.1109/EDUCON45650.2020.9125183.

[20] J. Pereira., "Leveraging chatbots to improve self-guided learning through conversational quizzes." *Association for Computing Machinery*, p. 911–918, 2016, https://dl.acm.org/doi/10.1145/3012430.3012625.

[21] D. Lee, Y. Huh, and C. Lin, "Technology functions for personalized learning in learner-centered schools." vol. 66, no. 5, Oct 2018, https://doi.org/10.1007/s11423-018-9615-9.

[22] S. Açışlı, S. A. Yalçın, and Ümit Turgut., "Effects of the 5e learning model on students' academic achievements in movement and force issues, procedia - social and behavioral sciences." *ScienceDirect*, vol. 15, pp. 2459–2462, 2011, https://doi.org/10.1016/j.sbspro.2011.04.128.

Securing Digital Data: A New Edge Detection and XOR Coding Approach for Imperceptible Image Steganography

Hayat Al-Dmour
Faculty of Information Technology
Mutah University
Mu'tah, Karak, Jordan

***Abstract*—The rapid progress of digital devices and technology, coupled with the emergence of the internet has amplified the risks and perils associated with malicious attacks. Consequently, it becomes crucial to protect valuable information transmitted through the internet. Steganography is a tried-and-true technique for hiding information beneath digital content, such as pictures, texts, audio, and video. Various methodologies of image steganography have been developed recently. In image recognition, edge detection secures an image into well-defined areas. This paper introduces a novel image steganography algorithm with edge detection and XOR coding techniques. The proposed approach aims to conceal a confidential message within the spatial domain of the original image. In contrast to uniform regions, the Human Visual System (HVS) is less responsive to variations in the sharp areas; an edge detection algorithm is applied to identify edge pixels. Furthermore, to enhance the efficiency and reduce the embedding impact, XOR operation has been utilized to embed the secret message in the Least Significant Bit (LSB). According to the results of the experiments, the proposed method embeds confidential data without causing noticeable modifications to the stego image. The proposed method system produced imperceptible stego images with minimal embedding distortions compared to existing methods. Based on the results, the proposed approach outperforms the conventional methods regarding image distortion techniques. The PSNR values achieved by the proposed method are higher than the acceptable level.**

***Keywords*—Steganography; information hidings; bits modification; decoding algorithm; edge detection; canny edge detection; human visual system**

I. Introduction

Data security is considered one of the most noteworthy factors of Information and Communication Technology (ICT) due to the rapid growth in electronic technologies and the internet. Therefore, a necessary prevention mechanism is needed to protect the data securely. Commonly, sensitive information can be protected either by using cryptography or Steganography. The technique of cryptography conceals the contents of sensitive data to unreadable text by several transformations; however, Steganography is the art of concealing data in an explicit transport file in such a manner that unapproved third parties find it challenging to discover and retrieve the concealed data [1]. An image is represented as an array of numbers corresponding to the light intensities at different points, known as pixels. These pixels collectively form the raster data of the image [2]. At the same time, digital images are the most common form of cover object used for Steganography, as they are the most prevalent carrier on the internet [3]. Image steganographic methodologies can be classified into the Spatial domain and Transform domain methods. In the Spatial domain, the intensity of the pixels is used to implant information. In contrast, information is embedded in the frequency domain of the previously transformed images in the Transform domain. In image steganography, confidential information is protected from malicious attacks by changing the pixels, and the modifications applied to the image are made unnoticeable. The original image without sensitive information is called the cover image, while the cover image with secret information embedded in it is called the stego image. Steganography requires two files: the message and the cover image, which conceal the message. Digital images are preferred over videos due to their compact and smaller size compared to the large and redundant size of the videos when transmitted over low bandwidth networks [4].

The image's detailed contents, information, and features cannot be observed by the Human Visual System (HVS) or the naked eye. Therefore, the use of techniques becomes necessary to check whether an image is an original or a stego image. It is essential to evaluate the strengths and weaknesses of steganography methods based on various characteristics [5]. Some important requirements conflict with each other to develop a reliable steganography algorithm. The three fundamental characteristics of image steganography are capacity, robustness, and imperceptibility. A cover medium's capacity refers to how many bits it can contain. Imperceptibility and robustness are standard requirements that conflict with embedding capacity. Imperceptibility refers to the quality of the stego carrier. The steganography algorithm satisfies the imperceptibility requirement, even though the stego carrier content may differ from the original one if that difference cannot be noticed by the human visual system (HVS) [6]. The imperceptibility is usually computed by the Peak Signal Noise Ratio (PSNR). Greater PSNR means higher imperceptibility. Robustness relates to the strength of the stego medium to endure different types of manipulation. It is, therefore, hard for attackers to illegally modify or remove the embedded secret data [7].

Different image steganography techniques are used to embed the information in the cover media, i.e., spatial and transform domain techniques. The Least Significant Bit (LSB) replacement is one such approach that seeks to replace the least significant bit of each pixel with the associated concealed data

pixel. As a result, such a pixel's initial magnitude changes by 0 or 1, a tendency repeated throughout the cover picture. Least Significant Bit (LSB) is the most widely used method for image steganography based on spatial domain techniques. In this method, the message is embedded in LSB directly [8].

The research problem at hand revolves around enhancing image steganography techniques to embed data securely while maintaining imperceptibility and robustness. This research aims to answer the following questions:

- How can image steganography methods be improved to enhance imperceptibility without compromising capacity?
- What novel approaches can be developed to ensure secure and undetectable data embedding in images?

The main contribution of this study lies in the proposition of an image steganography method based on edge detection, guaranteeing identical edge images in the original and stego images while securely embedding messages. This approach ensures that the concealed message can be correctly extracted from the stego image. The results of the experiments show that the proposed method embeds secret data without causing noticeable modifications to the stego image.

The rest of this work is structured as follows. Section II presents the most recent steganography techniques. The proposed scheme is explained in Section III. Section IV reports on the experimental results, analysis, and discussion. Section V brings the paper to a conclusion.

II. Steganography and its Techniques

Steganography is a scientific discipline that involves concealing data within another form of data. For example, it can include hiding a plaintext message within an image file. Throughout history, various entities such as individuals, the military, secret intelligence agencies, and governments have leveraged Steganography to covertly communicate and transmit information without arousing suspicion. Steganography has a wide range of uses, including confidential communication, electronic watermarking, reliability of data, copyright protection, and identifying manipulation of data [9]. Extensive research work has been carried out to develop digital image steganography. The LSB technique was one of the first to obscure data transmission by embedding secret data into unimportant bits of pixels. LSB approaches, in general, substitute the identical length bits in each underlying pixel with the embedding data. Nevertheless, not all pixels in the image can have equal levels of alteration without producing apparent distortion [10]. The stego image resembles the original image because altering the LSB of such a pixel does not significantly alter the color. Despite this, not all pixels in an image are capable of enduring equal amounts of modifications without noticeable distortion resulting in a low-quality stego image. To handle this issue, some image steganography methods based on LSB have used HVS features to conceal the secret bits in the cover image [11].

Image Steganography is achieved by conducting an XOR operation on the bits of pixel values. In this regard, Joshi et al., retrieved the two rightmost LSBs and two leftmost MSBs of the pixel value using their method. They combined the first and second bits using the XOR operation. The message bit 0 was concealed in the LSB of the pixel value if the result of the two XOR operations was 11 or 00, and the message bit 1 was hidden if the impact of the two XOR operations was 10 or 01 [12].

XOR is the logical operation performed on the LSB and MSB based on the technique. Based on the result of the XOR operation, the message is embedded in the LSB of a particular pixel. Baek et al., suggested a method for embedding the information in the grayscale image to share it secretly. They used the XOR operation to represent the bits at a specific location in the image [13].

Further, a previous study proposed a complicated method for image steganography to hide information in the LSB area of an image pixel. The authors used the XOR operation three times before embedding the message in the LSB. XOR operation was performed on the three MSB bits. This operation behaved as a key to embedding a message in an image. Better security was provided using this simple operation. A study proposed an alpha-trimmed mean filter to enhance the image quality, whereas they used XOR operation on 6-MSBs to add two bits of secret message in the image at 2-LSBs [12].

Additional options for concealing the information included two levels of encryption and an obfuscation phase. Two XOR operations and a private key were used to encrypt the data. The LSB method was then used to incorporate the information in the cover picture. One straightforward XOR-based process selected the colors using a sequencing technique and modified several LSB methods. Three MSBs were utilized as the key in the steganographic procedure, and they employed a triple XOR literary content that needed to be delivered [14].

The least significant bit of an image is subjected to an XOR procedure. When an 8-bit random key is used in the application, pixel 1 from the red matrix's second bit is XORed with the pixel. Because the result of the XOR of the taken bit 1 and the taken bit 0 is 1, the pixel must be satisfied by delivering an encrypted message that conceals the value of the pixel's first LSB bit. The pixel will pass if the XOR operation returns a result of 0, but the following step's process will continue using the same 8-bit random key. Based on the length of the encrypted message, this procedure will continue [15].

To find the secret message, bits from a pixel were extracted and saved in an encrypted message. In the recovery process, pixel 1 of the second bit for the red matrix and the 8-bit random key is XORed with the pixel, and the result of the XOR of bits 1 and 0 is 1. Consequently, the pixel must convey an encrypted message buried in the value from its first LSB. The pixel will pass if the supplied response during the XOR operation is 0, but the following step in the procedure will still use the same 8-bit random key. Following the completion of this phase, the decrypted image is extracted from the stego-image. These bits are taken into the LSB of the identical pixel shown in the stego-image. This process continues till the length of the message is sent [16].

III. Materials and Methods

Compared to smooth regions, the human visual system is less sensitive to modification in image regions with sharp transitions. To attain undetectable Steganography, the secret

message has to be embedded in the edge regions of the cover image. Regardless of how insignificant the changes are in the cover image, conventional edge detection methods produce sensitive edge images. Since hiding the message might cause some alterations to the cover image, this feature restricts the implementation of image steganography based on edge detection. The edge detection technique is commonly used in digital images to determine if each pixel has a high or low spatial frequency [17]. It is the technique of finding locations in a computer image where the image brightness swiftly changes, for example, pixels diverging from minimal intensities to high intensities or the other way around, displaying certain discontinuities [18]. Therefore, this paper presents a new image steganography method based on edge detection that produces identical edge images in the original and stego images. Sobel and Canny edge detection methods are used to extract edges. The Canny edge detection method was created by John Canny in 1986. It is one of the most efficient and well-known [19]. Therefore, Sobel and Canny edge detection method gives the identical edges of both the original and stego-image. In this way, the concealed message can be correctly extracted from the stego image.

In the proposed method, three bits of the message are embedded in a grayscale image intended for transmission to the receiver. The colored image has three channels: Red, Green, and Blue, each of matrix length and width of image size. In the case presented, the input image is in a grayscale where only one channel represents the image in one matrix. The method is centered on the advantage of XOR operation. The XOR operation is performed on 4-pixel values to get 3 XORed results. Three bits of message XORed with four least significant bits of the image to embed 3 bits of message in stego image. The block diagram of the proposed methodology is given in Fig. 1. The steps of the proposed image steganography algorithm are as follows:

Step 1: Convert the image into grayscale.

Step 2: Resize the image into 512 x 512 to get a uniform image size.

Step 3: Perform Edge detection on the image using Canny or Sobel edge detection methods.

Step 4: Convert the message (secret message to be sent) into decimal using the American Standard Code for Information Interchange (ASCII) code character by character.

Step 5: Convert decimal ASCII codes for each character into a binary format where ASCII-encoded data is of 8-bit length.

Step 6: Iterate over 3 bits of the message (secret message to be sent).

Step 7: Get 4-pixel locations in the cover image from the identified edges, i.e., P_1, P_2, P_3, and P_4, where P_1, P_2, P_3, and P_4 are the pixel 4-LSBs of the cover image.

Step 8: Calculate the XOR operation on P_1 and P_2, P_3 and P_4 and P_1 and P_3 as follows:

$$k_1 = P_1 \oplus P_2 \ , \ k_2 = P_3 \oplus P_4 \ , \quad k_3 = P_1 \oplus P_3$$

Step 9: Embed message bits into the stego image based on the XOR calculated by comparing the three estimated bits, $k1$, $k2$, and $k3$, with the three secret message bits according to Table I.

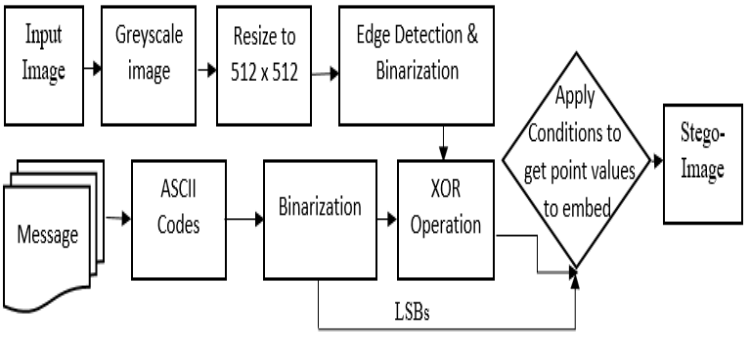


Fig. 1. Block diagram of the proposed methodology.

Step 10: Convert the image matrix from binary to decimal to get the stego image matrix.

In the first step, the RGB image is converted to grayscale. After that, the cover image is resized to a fixed size of dimensions 512×512, where the image contains 512 rows and 512 columns. Then, Canny or Sobel edge detection methods are performed to detect edges in the image.

Edge detection is the technique for identifying points in digital images with discontinuities. It is an abrupt change in the brightness of the image. The point where this change occurs is the edge. The four groups of edges are created that represent the place where the message has to be embedded. The location of four points, P_1, P_2, P_3, and P_4, are extracted from the edges. Fig. 2 shows an example of edge detection. The edges are calculated to store the message and the position for steganographic content (secret message) is calculated as X in Fig. 2. The position of x is calculated using the location of edge detection points, and then XOR these points with each other. Edge detection is performed using two types of edge detection methods: Sobel filter and canny edge detection, for images to use for XOR coding in image steganography.

The Sobel filter is an edge detection method for digital images. It is also called Sobel Feldman sometimes. Edges are the discontinuities in the image that cause a rapid change in intensity value, as discussed earlier. It uses a filter in both vertical and horizontal directions, producing thick and bright edges in every direction. It uses gradient operation. The estimated magnitude of gradient operation is measured using the sum of specific values of slope in two directions: horizontal and vertical.

Canny edge detection is the method for detecting edges in the image. It uses three criteria to detect edge detection performance: localization precision, SNR, and single-edge response precision. It produced the best results in many problems.

In order to send secret messages, steganographic codes are converted into ASCII codes. Subsequently, the binary data was obtained by converting it into binary codes (ASCII), enabling the application of embedding using the suggested XOR approach. Just on spots located using the edge detection technique, XOR was used. The process begins with the points $P1$, $P2$, $P3$, and $P4$, and then XOR is performed to

101	103	106	109	99	98
100	102	103	70	80	70
99	60	103	71	83	72
50	61	80	79	85	71
51	80	82	80	79	73

(a) a

0	0	0	0	0	0
0	0	0	1	1	1
0	1	0	x	x	x
1	x	x	0	0	0
x	0	0	0	0	0

(b) b

Fig. 2. (a) Intensity values of the input image (b) Stego bit positions identification using edge detection.

TABLE I. CONDITIONS TO EMBED THE MESSAGE

Condition Description	Action
All m bits match k bits	-
if m3 does not match k3	$\bar{P}3$ and $\bar{P}4$
if m2 does not match k2	$\bar{P}4$
m2 does not match k2 and m3 does not match k3	$\bar{P}3$
if m1 does not match k1	$\bar{P}2$
m1 does not match k1 and m3 does not match k3	$\bar{P}1$
m1 does not match k1 and m2 does not match k2	$\bar{P}2$ and $\bar{P}4$
None of the m bits match their corresponding k bits	$\bar{P}1$ and $\bar{P}4$

$P1$, $P2$, $P3$, and $P4$ and recorded the results in $k1$, $k2$, and $k3$, where $k1$, $k2$, and $k3$ are indeed the XORed results of the estimated edge point.

Based on the above conditions, shown in Table I, three message bits are embedded into four-point locations that were extracted using the edge detection method. Hence, three message bits are embedded into the image, showing the average embedding is 1.25 bits.

To extract the messages from steganographic images, it starts with calculating the edge based on the same method used during the embedding process. The points were identified using edge detection, as shown in Fig. 3. For example, a total of 1000 edges are presented in the file, enabling the embedding of three bits in four-point locations, thereby accommodating 750 bits within 1000 edges. Nonetheless, 2 bits are used to predict the threshold and 1 bit for data embedding. In every position, the first two bits are used to indicate the threshold, and then the next bit is used to embed the data; using this way, the 1000 locations will get reduced to 3 times = 333 locations. Now in 333 locations, the data is embedded, which reduces the embedding rate. Conversely, it enhances data corruption if any attack on the image happens.

As the locations chosen for embedding the data are not employed to predict the edges, this technique enhances the extraction process. These are the edge points where the XOR operation was performed to embed the secret messages. Once these points are calculated and represented to get m1, an XOR operation is performed on q1 and q2. To get m2, the XOR operation is performed on q3 and q2. To acquire m3, the XOR operation is conducted on q1 and q3. The block diagram for extracting the message is shown in Fig. 3.

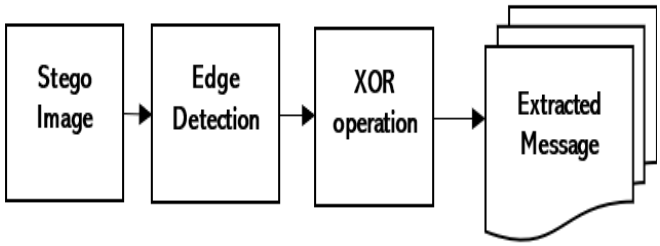


Fig. 3. Block diagram for message extraction.

IV. RESULTS AND DISCUSSION

Distortion evaluation methods are the methods that measure the distortion in the image by comparing the original image and the steganographic image. PSNR, the number of edges in the image, SSIM, MSE, and UQI are used to evaluate image distortion in the image. The proposed method is evaluated using various embedding distortion evaluation methods.

PSNR is the peak signal-to-noise ratio. It is the ratio between the peak signal, which means the maximum power of the test image, and the noise, maximum noise. Noise is the image distortion that affects the image's representation quality. It is calculated as shown in Eq. 1.

$$PSNR = 10\log_{10}\left[\frac{255^2}{MSE}\right] \ (dB) \tag{1}$$

MSE is the mean squared error in the image, calculated by the mean of squared differences between input and steganographic images. The difference is calculated from pixel to pixel in an input image and stego image. The summation of all the pixel differences is then divided by the total number of pixels, that is, width x height of the images (input and stego). Eq. 2 represents how the mean squared error is calculated.

$$MSE = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H}(I_{ij} - SI_{ij})^2 \tag{2}$$

Where Iij and $SIij$ represent the pixel value at the location i and j in the input image and stego-image. At the same time, W and H are used to represent the width and height of both of the images.

SSIM is the structural similarity index. It is the quality measure of the image that represents the image degradation after performing the Steganography in the image. It is calculated by the lamination, contrast, and structure between the input image and the stego image. To calculate the SSIM, local mean, standard deviation, and cross-covariance are used between input and stego image. It is the perceptual difference between both images. Eq. 3 is used to calculate it.

$$SSIM(I, SI) = \frac{(2\mu_I\mu_{SI} + C_1)(2\sigma_{ISI} + C_2)}{(\mu_I^2 + \mu_{SI}^2 + C_1)(\sigma_I^2 + \sigma_{SI}^2 + C_2)} \tag{3}$$

Where I, SI,$\sigma I\sigma SI$, and σISI represent the local mean, standard deviation, and cross-covariance for both images.$C1$

(a) a (b) b (c) c (d) d (e) e (f) f (g) g (h) h

Fig. 4. Cover images (a) Baboon (b) Boat (c) Couple (d) House (e) Lena (f) Pepper (g) Sailboat (h) Tank.

and $C2$ are the regularization constants. UQI is the universal image quality. It is the ratio between the multiplicative variance and summation of variances of the input image and stego-image.

In order to perform the experimentation, the benchmark dataset of USC-SIPI is used. The experimentations are conducted on 8 images from the dataset, as shown in Fig. 4. The dimension of 512×512 is utilized, as already discussed in the methodology. The images for this experimentation purpose are (a) Baboon, (b) Boat, (c) Couple, (d) House, (e) Lena, (f) Pepper, (g) Sailboat, and (h) Tank.

Distortion evaluation is performed for two cases: Sobel filter and Canny edge detection, after completing the distortion

TABLE II. IMAGE EVALUATION RESULTS ON MESSAGE LOAD OF 24000 BITS

Filter	Image	PSNR	No. of Edges	Embedding Rate	SSIM	MSE	UQI
Canny	Tank	65.08	51602	0.091553	0.99992	0.0202	1
	baboon	64.24	70387	0.091553	0.99996	0.0245	1
	boat	64.84	58471	0.091553	0.99987	0.0213	1
	couple	64.93	56701	0.091553	0.99987	0.0209	0.99
	house	64.71	59923	0.091553	0.99991	0.0219	1
	lena	65.32	47839	0.091553	0.99983	0.0191	0.99
	pepper	65.20	50535	0.091553	0.99983	0.0196	0.99
	sailboat	64.86	55720	0.091553	0.99991	0.0212	1
Sobel	Tank	64.65	60875	0.091553	0.99990	0.0223	1
	baboon	63.30	94586	0.091553	0.99995	0.0304	1
	boat	63.58	86512	0.091553	0.99980	0.0285	1
	couple	63.50	88790	0.091553	0.99980	0.0291	1
	house	63.61	86757	0.091553	0.99987	0.0284	1
	lena	63.98	77158	0.091553	0.99974	0.0259	0.99
	pepper	63.50	87713	0.091553	0.99974	0.0290	1
	sailboat	63.58	85944	0.091553	0.99987	0.0285	1

TABLE III. IMAGE EVALUATION RESULTS ON MESSAGE LOAD OF 32000 BITS

Filter	Image	PSNR	No. of Edges	Embedding Rate	SSIM	MSE	UQI
Canny	Tank	64.58	51590	0.12207	0.99992	0.0227	1
	baboon	63.85	70395	0.12207	0.99995	0.0268	1
	boat	64.37	58445	0.12207	0.99986	0.0238	1
	couple	64.43	56687	0.12207	0.99986	0.0235	1
	house	64.24	59926	0.12207	0.99989	0.0245	1
	lena	64.75	47833	0.12207	0.99981	0.0218	0.98
	pepper	64.66	50504	0.12207	0.99981	0.0223	0.99
	sailboat	64.44	55703	0.12207	0.99988	0.0234	1
Sobel	Tank	64.18	60866	0.12207	0.99989	0.0248	1
	baboon	62.98	94583	0.12207	0.99994	0.0327	1
	boat	63.20	86497	0.12207	0.99979	0.0312	1
	couple	63.17	88743	0.12207	0.99979	0.0313	1
	house	63.24	86786	0.12207	0.99986	0.0309	1
	lena	63.56	77105	0.12207	0.99972	0.0286	0.98
	pepper	63.15	87698	0.12207	0.99971	0.0315	1
	sailboat	63.25	85963	0.12207	0.99986	0.0307	1

TABLE IV. IMAGE EVALUATION RESULTS ON MESSAGE LOAD OF 40000 BITS

Filter	Image	PSNR	No. of Edges	Embedding Rate	SSIM	MSE	UQI
Canny	Tank	64.14	51605	0.152588	0.99991	0.0251	1
	baboon	63.42	70392	0.152588	0.99993	0.0296	1
	boat	63.94	58445	0.152588	0.99985	0.0263	1
	couple	64.01	56649	0.152588	0.99984	0.0259	0.99
	house	63.83	59984	0.152588	0.99986	0.0270	1
	lena	64.23	47783	0.152588	0.99979	0.0246	0.97
	pepper	64.22	50509	0.152588	0.99979	0.0246	0.99
	sailboat	63.94	55711	0.152588	0.99985	0.0262	1
Sobel	Tank	63.71	60875	0.152588	0.99988	0.0277	1
	baboon	62.67	94591	0.152588	0.99993	0.0351	1
	boat	62.91	86535	0.152588	0.99978	0.0333	1
	couple	62.87	88774	0.152588	0.99977	0.0336	1
	house	62.88	86763	0.152588	0.99984	0.0335	1
	lena	63.21	77063	0.152588	0.99970	0.0311	0.97
	pepper	62.85	87631	0.152588	0.99970	0.0338	1
	sailboat	62.92	85927	0.152588	0.99983	0.0332	1

TABLE V. IMAGE EVALUATION RESULTS ON MESSAGE LOAD OF 48000 BITS

Filter	Image	PSNR	No. of Edges	Embedding Rate	SSIM	MSE	UQI
Canny	Tank	63.74	51593	0.183105	0.99990	0.0275	1
	baboon	63.09	70343	0.183105	0.99992	0.0319	1
	boat	63.56	58483	0.183105	0.99983	0.0287	1
	couple	63.64	56645	0.183105	0.99982	0.0281	0.98
	house	63.45	59928	0.183105	0.99984	0.0294	1
	lena	63.82	47830	0.183105	0.99977	0.0269	0.97
	pepper	63.77	50491	0.183105	0.99977	0.0273	0.98
	sailboat	63.54	55660	0.183105	0.99983	0.0288	1
Sobel	Tank	63.33	60878	0.183105	0.99987	0.0302	1
	baboon	62.35	94581	0.183105	0.99991	0.0378	1
	boat	62.59	86505	0.183105	0.99977	0.0359	1
	couple	62.56	88777	0.183105	0.99976	0.0361	1
	house	62.54	86734	0.183105	0.99982	0.0362	1
	lena	62.88	77108	0.183105	0.99968	0.0335	0.96
	pepper	62.54	87630	0.183105	0.99968	0.0362	1
	sailboat	62.57	85883	0.183105	0.99980	0.0360	1

(a) a (b) b

(c) c

Fig. 5. (a) Cover edge image, (b) stego edge image, and (c) difference between cover and stego edge images.

evaluation methods on message loads of 24000, 32000, 4000 and 48000 bits, respectively. The obtained results for tl message load of 24000 bits are shown in Table II. Tables III, IV, and V represent the results obtained on message load of 32000, 40000, and 48000 bits, respectively, for both cases. As shown in Tables II, III, IV, and V, both edge detection methods performed well. The result shows that when the message load is increased, the image quality is degraded. The image steganography can be applied on smaller message loads for better securing the data. As much as the message load increases, then the image quality degrades. Hence, it would be prone to be detected as a stenographic image.

The statistics provided show the proportion of even and odd pixels for every pairing of pictures. The cumulative numbers are quite comparable both before and after. Consider that the discrepancy between the sum of the variations and the number of manuscripts in the secret message is less. Fig. 5(a) shows the edge pixels of the cover image, which is identified by applying Sobel edge detection. Edge pixels of the stego image obtained after embedding a message are shown in Fig. 5(b), and Fig. 5(c) shows the difference between the two edge images. This indicates that the edge pixels in the cover and stego images are similar.

The image quality improves as more data is in the stego image and with the increase in the number of edges. The peak noise signal (PSNR) method for implementing the steganography procedure is the primary determinant of optimizing efficiency. The stego image will be more similar to the actual image if PSNR quality increases. The hiding capacity is increased with the pixel in stenography.

These findings show that the pictures look entirely unmodified to the unaided eye and are statistically the same. It is challenging for a person looking at both images or a computer looking at just one image to notice the possibility of increasing statistical similarity by introducing noise. Before and after changes, the ratio of even and odd dots is approximately the same. PSNR and MSE are used to assess the quality of stego images compared to cover images. Based on experimental results, the proposed method achieves a high-quality image by using the XOR operation to reduce the difference between the cover and stego images. PSNR values vary between 65 dB and 62 dB with embedding rates of 9% - 18%, where 35 dB is the minimum acceptable value.

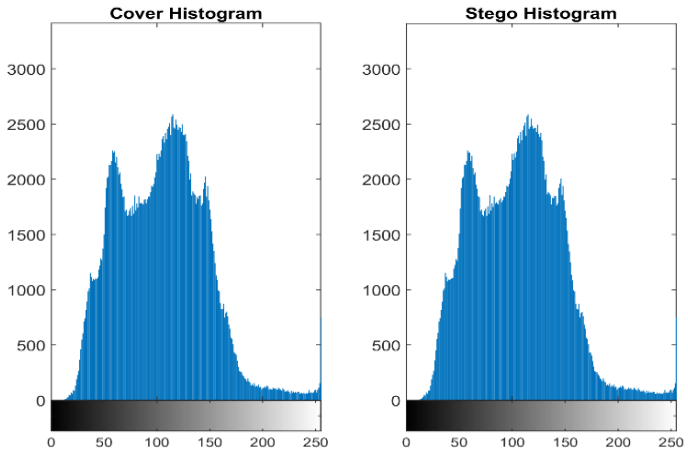


Fig. 6. (a) The cover image histogram and (b) The histogram of the corresponding stego image using the proposed algorithm.

The visual differences between the cover and stego images cannot be discovered by the human eye, and even the histograms of the stego images (illustrated in Fig. 6(a) and 6(b) are pretty similar. The algorithm's effectiveness was demonstrated by employing LSB for steganographic techniques and evaluating the histogram, PSNR, safety, and resemblance between human and machine readings.

V. Conclusion

The human eye is capable of detecting a significant variation in an image pixel. This means that the edges of an image might allow greater distortion than the other sections of the cover image since the edge portions have a sharper shift in pixel values than those that surround them. As a result, in edge-based Steganography, the majority of message bits are embedded in the edge pixels, and as one moves from the edge regions into the homogeneous areas of the cover picture, the number of bits embedded reduces, making the distortion less visible to the human eye. In this paper, an XOR-based embedding method was proposed to embed the data into the image. Sobel and Canny edge detection methods were used to extract edges. Sobel edge detection method is the traditional edge detection method used a few years ago [14], [20]. On the identified edges, an XOR operation is performed, which is a disjunction property [19]. The embedding capacity of an image is enhanced by edge detection methodology. In the presented method, a good rate of PSNR has been achieved.

Using an edge detection approach along with multiple-bit modification methods leads to high security. The contribution of this paper is embedding the message efficiently by incorporating XOR coding and identifying identical edges in the cover and stego images using the traditional edge detection methods. The technique was tested on various image distortion methods. The proposed approach performed efficiently, as shown in the results. The intensity of the edges in the input and steganographic images are estimated to be identical. The MSE is almost zero on a message load of 24000, 32000, 40000, and 48000 bits. Universal image quality and SSIM are closer to one representing the image quality, and the similarity between cover and stego images is almost the same. The approach may be expanded to several picture formats, including grayscale images, and requires no further information other than the stego image.

In terms of benefits, the suggested solution is undetectable since it only employs three LSBs to hide the secret data in the pixels of the detected edges. Furthermore, the recommended approach scatters bits of the secret data over specific regions of the identified edges rather than over all pixels of the carrier picture. A second benefit is that the buried data may be recovered. Different outputs for the same input image and secret data can be generated by hiding the secret data using a specific pattern denoted by the Canny algorithm, as well as parameterizing the algorithm to allow communicating parties to alter the effects and results of the algorithm.

References

[1] N. Ibraheem and M. Hasan, "Combining several substitution cipher algorithms using circular queue data structure," *Baghdad Science Journal*, vol. 17, no. 4, pp. 1320–1320, 2020.

[2] I. J. Kadhim, P. Premaratne, P. J. Vial, and B. Halloran, "Comprehensive survey of image steganography: Techniques, evaluations, and trends in future research," *Neurocomputing*, vol. 335, pp. 299–326, 2019.

[3] O. Rachael, S. Misra, R. Ahuja, A. Adewumi, F. Ayeni, and R. Mmaskeliunas, "Image steganography and steganalysis based on least significant bit (lsb)," in *Proceedings of ICETIT 2019: Emerging Trends in Information Technology*. Springer, 2020, pp. 1100–1111.

[4] S. K. Ghosal, A. Chatterjee, and R. Sarkar, "Image steganography based on kirsch edge detection," *Multimedia Systems*, vol. 27, no. 1, pp. 73–87, 2021.

[5] G. C. Kessler and C. Hosmer, "An overview of steganography," *Advances in Computers*, vol. 83, pp. 51–107, 2011.

[6] H. Al-Dmour, N. Ali, and A. Al-Ani, "An efficient hybrid steganography method based on edge adaptive and tree based parity check," in *MultiMedia Modeling: 21st International Conference, MMM 2015, Sydney, NSW, Australia, January 5-7, 2015, Proceedings, Part I 21*. Springer, 2015, pp. 1–12.

[7] D. R. I. M. Setiadi, "Psnr vs ssim: imperceptibility quality assessment for image steganography," *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 8423–8444, 2021.

[8] M. M. Emam, A. A. Aly, and F. A. Omara, "An improved image steganography method based on lsb technique with random pixel selection," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 3, 2016.

[9] H. Al-Dmour and A. Al-Ani, "A medical image steganography method based on integer wavelet transform and overlapping edge detection," in *Neural Information Processing: 22nd International Conference, ICONIP 2015, November 9-12, 2015, Proceedings, Part IV 22*. Springer, 2015, pp. 436–444.

[10] H. Khamis, "Studies on image steganography," Master's thesis, Itä-Suomen yliopisto, 2021.

[11] S. Gupta and N. K. Garg, "Optimized data hiding for the image steganography using hvs characteristics," in *Proceedings of the International Conference on Paradigms of Computing, Communication and Data Sciences: PCCDS 2020*. Springer, 2021, pp. 275–285.

[12] Y. P. Astuti, E. H. Rachmawanto, C. A. Sari *et al.*, "Simple and secure image steganography using lsb and triple xor operation on msb," in *2018 International Conference on Information and Communications Technology (ICOIACT)*. IEEE, 2018, pp. 191–195.

[13] P. P. Balgurgi and S. K. Jagtap, "Intelligent processing: An approach of audio steganography," in *2012 International Conference on Communication, Information & Computing Technology (ICCICT)*. IEEE, 2012, pp. 1–6.

[14] P.-Y. Chen, W.-E. Wu *et al.*, "A modified side match scheme for image steganography," *International Journal of Applied Science and Engineering*, vol. 7, no. 1, pp. 53–60, 2009.

[15] S. L. Chikouche and N. Chikouche, "An improved approach for lsb-based image steganography using aes algorithm," in *2017 5th International Conference on Electrical Engineering-Boumerdes (ICEE-B)*. IEEE, 2017, pp. 1–6.

[16] E. J. Kusuma, C. A. Sari, E. H. Rachmawanto, and D. R. I. Moses Setiadi, "A combination of inverted lsb, rsa, and arnold transformation to get secure and imperceptible image steganography." *Journal Of ICT Research & Applications*, vol. 12, no. 2, 2018.

[17] Y.-H. Yu, C.-C. Chang, and Y.-C. Hu, "Hiding secret data in images via predictive coding," *Pattern recognition*, vol. 38, no. 5, pp. 691–705, 2005.

[18] C.-H. Yang, C.-Y. Weng, S.-J. Wang, and H.-M. Sun, "Adaptive data hiding in edge areas of images with spatial lsb domain systems," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 3, pp. 488–497, 2008.

[19] K. Gaurav and U. Ghanekar, "Image steganography based on canny edge detection, dilation operator and hybrid coding," *Journal of Information Security and Applications*, vol. 41, pp. 41–51, 2018.

[20] H.-W. Tseng and H.-S. Leng, "High-payload block-based data hiding scheme using hybrid edge detector with minimal distortion," *IET Image Processing*, vol. 8, no. 11, pp. 647–654, 2014.

Incorporating News Tags into Neural News Recommendation in Indonesian Language

Maxalmina Satria Kahfi, Evi Yulianti, Alfan Farizki Wicaksono
Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

***Abstract*—News recommendation system holds the potential to aid users in discovering articles that align with their interests, which is critical to alleviate user information overload. To generate effective news recommendations, one key capability is to accurately capture the contextual meaning of text in the news articles, since this is pivotal in acquiring useful representations for both news content and users. In this work, we examine the effectiveness of neural news recommendation with attentive multi-view learning (NAML) method to conduct a news recommendation task in the Indonesian language. We further propose to incorporate news tags, which at some levels may capture the important contextual meanings contained in the news articles, to improve the effectiveness of the NAML method in the Indonesian news recommendation system. Our results show that the NAML method leads to significant improvement (if not comparable) in the effectiveness of neural-based Indonesian news recommendations. Further incorporating news tags is shown to significantly increase the performance of the NAML method by 5.86% in terms of NDCG@5 metric.**

***Keywords*—News recommendation; recommendation systems; news tags; user modeling**

I. Introduction

A recommendation system plays a crucial role in helping users discover items that match their needs and preferences [1]. In business, it is a vital part of marketing strategies, particularly for boosting online sales by offering customers a curated selection of items tailored to their preferences [1, 2]. This process of recommendation can be automated using a wide range of techniques, such as content-based methods and collaborative filtering [3], and recently involves computing user and item embeddings which serves as a basis for predicting how much a user is likely to prefer a particular item [4, 5]. Recommendation systems have found application in various domains, including movie recommendation [6, 7, 8], news recommendation [9, 10, 11, 12, 13], and music recommendation [14, 15, 16, 17].

News recommendation stands as one of the most frequently employed use cases of recommendation systems within the digital landscape, notably embraced by online platforms and news publishers [2, 18, 10]. Given the overwhelming amount of news articles generated daily, it is impractical for users to manually sift through all of them to find content that aligns with their interests [19]. Every day, an enormous volume of news articles is generated and published online, posing a challenge for users to efficiently discover news that aligns with their interests [2, 20]. Therefore, the implementation of personalized news recommendation becomes vital, as it enables online news platforms to target user preferences effectively and alleviate the issue of information overload [10].

In the past few years, news recommendation models have benefited from the use of deep learning methods [9, 11, 10, 13, 12]. These models are often referred to as neural recommendation models. Various neural recommendation models have been introduced to tackle news recommendation challenges, including NAML (Neural news recommendation with Attentive Multi-view Learning) method [11]. The NAML method uses an attentive multi-view learning model that learns unified news representations from different kinds of news information, such as news titles, bodies, and categories. This method takes advantage of multiple useful information from news articles that could enrich the semantics captured in the news representation. As a result, this method has been shown to outperform a range of deep learning methods for news recommendation, such as Convolutional Neural Network (CNN) [21], Deep fusion model (DFM) [22], Deep news recommendation based on knowledge-aware CNN (DKN) [23], etc. The NAML method also offers the flexibility to incorporate other useful information that could be exploited from news articles to produce more accurate news representation, which may result in more effective news recommendations. With this reported effectiveness as well as the potential to utilize extra knowledge for improving news representation, we propose to use the NAML method to perform the Indonesian neural news recommendation task in this work.

As mentioned above, the NAML architecture enables us to easily incorporate extra information to produce news representation. Therefore in this work, we also want to investigate the effectiveness of integrating news tags as supplementary information into the news recommendation model, aimed at improving the news representation. We argue that news tags at some levels may capture the important points in the articles, which therefore may provide extra useful information to generate better news representation. Previous research has primarily relied on a limited set of news components, such as title, category, and subcategory [24]. To the best of our knowledge, the use of news tags to enhance recommendation performance has not been explored in previous work.

Finally, this research endeavors to address two primary research questions:

1) RQ1. How is the performance of the NAML neural recommendation method to perform news recommendation task in the Indonesian language?

2) RQ2. To what extent news encoder in the NAML model can benefit from the use of news tags in the Indonesian news recommendation task?

By providing answers to these research questions, this study contributes to the existing body of research by providing

insights into the effectiveness of a neural method, i.e., NAML, for Indonesian news recommendation systems, as well as the impact of incorporating news tags into the NAML architecture.

The rest of this paper is organized as follows. Section II is dedicated to an in-depth review of relevant literatures related to news recommendation models, news information, and text embedding. Section III outlines the research methodology adopted in this study. Moving on to Section IV, this is where the results of the data analysis are presented and discussed. Section V serves as the conclusion, encapsulating key findings and providing directions for future research based on the study's insights. Finally, Section VI provides recommendations for future work based on the findings of this study.

II. Related Works

A. Text Embedding

Text embedding is a technique that represents text as a vector of real numbers. This allows computers to process text in a more meaningful way, as the vector representations can capture the semantic meaning of the text. There are several text embedding methods that try to capture the semantic and syntactic meaning of the text input.

The Word2Vec model served as a fundamental baseline for word embedding [25]. Their work introduced two innovative model architectures specifically crafted for generating continuous vector representations of words using vast datasets. These representations were evaluated for their quality in a word similarity task. The Word2Vec model has demonstrated its capability to address various tasks, including text summarization [26, 27, 28], ranking for academic expert finding [29, 30], and text classification [31, 32, 33]. Several models similar to Word2Vec were subsequently introduced, including Glove (Global Vectors for Word Representation) [34] and FastText [35]. These models build upon Word2Vec's foundations, enhancing the learning of word representations to acquire deeper semantic insights from the corpus.

In 2017 BERT is introduced, a bidirectional model that can learn context-aware and informative representations of words and phrases [36]. BERT, as a bidirectional model, learns word representations considering both forward and backward contexts, distinguishing it from previous models limited to forward context understanding. BERT is pre-trained on a massive dataset of text and code, and it can be fine-tuned for a variety of natural language processing tasks. This model remains challenging for underrepresented languages, which face unique obstacles due to limited data availability and the substantial corpus required for effective training [37].

Notably, there have been notable works on BERT-based models created specifically for the Indonesian language. These models, known as IndoLEM [38] and IndoNLU [39], have been developed independently, each trained on a different corpus of Indonesian text. IndoLEM and IndoNLU represent valuable contributions to Indonesian NLP research. They serve as pre-trained language models that have learned to understand and encode the linguistic patterns, semantics, and contextual information present in Indonesian text data. In addition, the BERT-based model has been demonstrated to enhance performance in classification tasks [31]. In our research, we evaluate the capabilities of these two distinct BERT-based models to obtain text embeddings. Furthermore, our part of the research is to conduct a comprehensive comparison of various text embedding methods and evaluate their performance within the context of a news recommendation system.

B. News Information

In the scope of news recommendation systems, the use of news information as news representation plays a pivotal role in constructing the entire system [2, 19]. The recommendation model aims to construct a user profile customized to individual preferences through an analysis of the articles that a user reads. To achieve this, news representation involves the transformation of some or all news information into vector form [24].

Typically, the components input into the model encompass essential elements such as the article's title, abstract, body, category, and subcategory. These components collectively serve as the news representation for generating a comprehensive vector representation of news articles. This representation is instrumental in facilitating the modeling of user preferences and enabling the recommendation system to provide personalized news suggestions that align with each user's unique interests and preferences [4, 5].

Another valuable component of news information is the news tags. Although the news tags show the topic information that is closely related to the news content, the existing personalized news recommendation methods usually ignore the value of tags. News tags are usually used by users to track certain topics in the news portal which can trace other articles based on that tags. This behaviour leads to our research to check whether we add news tags as additional information is necessary for news recommendation models.

One of the methods to process the tag information is by leveraging social bookmarking [40]. The method involves the utilization of tags, which are assigned by a community of users with shared preferences. It's important to note that while tags are still employed in this method, they now signify a collective understanding within the user group, aligning with their common interests and preferences. An alternative approach involves creating a probability relation graph among tags, exploring potential correlations among different tags [41]. Building on this foundation, they employ the term frequency–inverse document frequency (TF-IDF) method to determine tag weights. Additionally, they introduce a novel approach to calculate the correlation degree between tags, utilizing conditional probability as a key metric.

C. News Recommendation Methods

In recent years, the field of news recommendation systems has received significant attention due to the exponential growth of online news consumption and the need for personalized content delivery. Several studies have been conducted to explore various approaches for improving the accuracy and relevance of news recommendations. News recommendation systems can be broadly categorized based on how they model user behavior. Two primary categories exist: Candidate-Agnostic (C-AG) models and Candidate-Aware (C-AW) models [42]. Both of the categories are illustrated in Fig. 1.

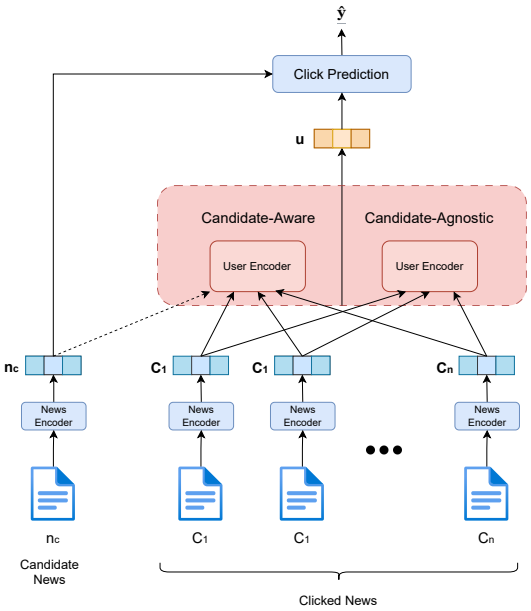


Fig. 1. News recommendation system general framework.

In C-AG Models, the User Encoder (UE) forms user embeddings exclusively from the embeddings of previously clicked news articles, without taking into account the candidate news articles [10, 9, 11, 12]. Essentially, the user embedding remains the same regardless of the specific candidate news being presented. One of the C-AG models is NAML that using a multi-view learning framework to learn unified news representations by incorporating several news information like title, body, and category. The NAML architecture is easily modified to add other news information. For the user encoder, NAML utilize a news attention network to identify crucial news articles, aiding in the acquisition of informative user representations [11]. The experimental results showcased significant enhancements in recommendation accuracy with the deep learning approach compared to traditional methods. Due to this performance, we have selected NAML as the primary model for our research focus.

Conversely, in C-AW models, the User Encoders (UEs) produce user embeddings that are influenced by the content of the candidate news articles. This means that in C-AW Models, the user embeddings can vary depending on the specific candidate news article in consideration [23, 13]. While C-AG Models maintain a consistent user embedding, C-AW models adapt the user representation to the characteristics of the candidate news, thereby potentially improving recommendation accuracy.

Considering the importance of contextual information in news recommendations, researchers have explored the incorporation of additional factors such as temporal relevance and user context. With the advent of deep learning techniques, several studies have investigated the application of neural networks for news recommendation [10, 9, 11, 13].

With the evolution of news recommendation systems, several metrics are employed to compare the performance among models. The most common ones include AUC (Area Under Curve), MRR (Mean Reciprocal Rank), and NDCG (Normalized Discounted Cumulative Gain) [4, 43]. For models treating the news recommendation task as a classification problem, the Area Under Curve (AUC) score is a frequently used metric to measure the accuracy of the model. MRR is calculated as the reciprocal of the position of the first relevant element in the ranking. NDCG considers graded relevance (rating values) along with positional information of the recommended items. Both MRR and NDCG assess the relevancy of recommendations from the news recommendation system (NRS) model.

In summary, news recommendation systems have been extensively studied, employing various approaches such as contextual information and deep learning approaches. These studies have contributed to the advancement of news recommendation systems, enhancing their accuracy, relevance, and personalization capabilities.

III. Research Methodology

A. Dataset

We conducted experiments on a real-world news recommendation dataset collected from one of the largest news portals in Indonesia. We randomly sampled users who had at least 5 news click records during 4 weeks from Feb 10 to March 16, 2023. The last 1 week was treated as a data test. We collected the behaviour logs of these users in this period, which are formatted into impression logs. An impression log is a record of the news articles that are presented to a user when they visit a news page, including information about the time of the visit and the user's interactions such as clicks on these news articles. In Fig. 2, you can see the layout of the news article page and an example of news information presented within an article.

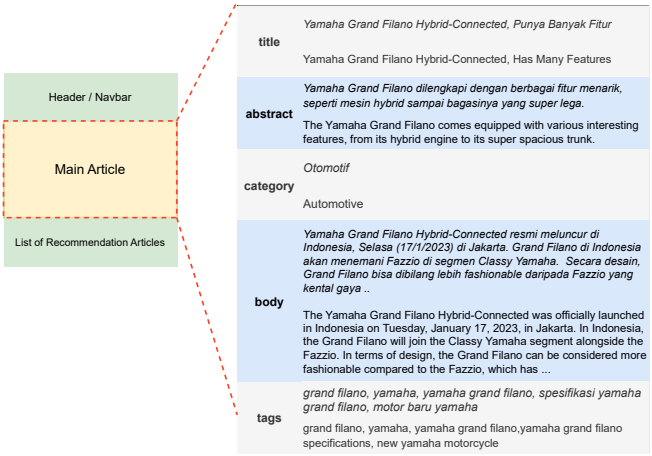


Fig. 2. Illustration of page layout and main article example.

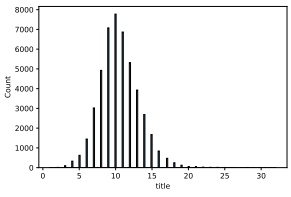
Regarding the news dataset, we focus on several news components to figure out their influence on the performance of news recommendation systems. These components encompass the title, abstract (short description), news body, category, and subcategory. Furthermore, our primary research objective is to

scrutinize the impact of news tags on the model. By examining how each of these elements shapes the model's performance, we aim to gain valuable insights into the factors that drive effective news recommendations. The statistical information about the dataset summarized in Table I.

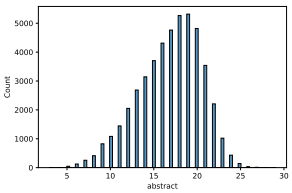
TABLE I. STATISTICAL INFORMATION ABOUT THE DATASET

Component	Value
# User	34,053
# User Logs	181,875
# News	79,041
# Category	27
# Subcategory	148
Avg. Title Len.	10.40
Avg. Abstract Len.	16.80
Avg. Body Len.	371.22
Avg. Tags per Article	4.55

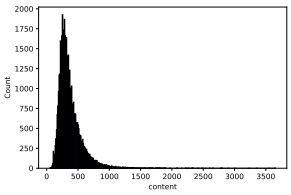
Fig. 3a, 3b, and 3c provide insights into the length distributions of news titles, abstracts, and bodies. Notably, news titles exhibit a distinct pattern of being quite short, averaging only 10.4 words. In contrast, both news abstracts and bodies present significantly longer text lengths, offering the potential for more in-depth coverage of news content. This discrepancy in text length underscores the value of incorporating diverse types of news information, including titles, abstracts, and bodies. Such integration enriches our comprehension of news articles by providing a more comprehensive and nuanced perspective on the content.



(a) Length of Title



(b) Length of Abstract

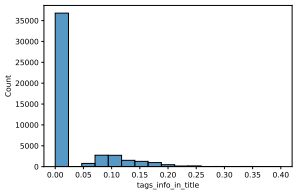


(c) Length of News Body

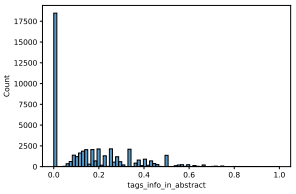
Fig. 3. Statistics of News Dataset.

In contrast to the approach outlined in [41], our methodology involves concatenating the news tags into a sentence or text format. We then treat this combined text in the same manner as other news information, such as the title and abstract. As a result, we transform the tags into a textual format, enabling them to be seamlessly integrated into the news content.

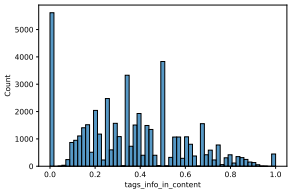
Our initial observations, as illustrated in Fig. 4, reveal an interesting connection between the amount of text data and the presence of words within news tags. With the expansion of text data, there is a higher probability of encountering shared words between the tags and other textual content. This intriguing pattern suggests that even though news tags are found within other news information, they can contribute to a more comprehensive understanding of the article. This is because the proportion of news tags included in other news information remains relatively low. This highlights the potential significance of news tags in enriching news representation.



(a) Tags information in title



(b) Tags information in abstract



(c) Tags information in body

Fig. 4. Percentage of tags information included in other news information.

B. Experiment Settings

In our study, we employed two pre-trained models of word embedding, FastText and Glove. Additionally, we utilized the feature-based approach of Indonesian BERT architecture to generate more contextualized word embeddings. We reported average results in terms of AUC, MRR, NDCG@5 and NDCG@10. For the training objective, we primarily focus on parameter tuning through minimizing cross-entropy loss (with negative sampling) as the straightforward classification objective [44].

C. Text Embedding Models

As part of our benchmarking process, we trained three word embedding models—Word2Vec, Glove, and FastText—using the Indonesian Wikipedia corpus extracted from the Indo4B corpus [39]. Furthermore, we also trained a FastText model using the entire Indo4B corpus. For our BERT-based models, we leveraged models from two distinct sources, namely indolem [38] and indobert [39]. These pretrained BERT models played a crucial role in our research, serving as integral components in our investigations into text embeddings and their impact on news recommendation systems.

D. News Recommendation Models

In this paper, we assess several C-AG models, each distinguished by their News Embedding (NE) component, which essentially determines how they embed the clicked news articles: (1) The LSTUR [9], focuses on learning user representations using recurrent networks. In this model, a short-term

user embedding is generated from the clicked news articles using a Gated Recurrent Unit (GRU) [45]. This short-term embedding is then combined with a long-term embedding, which is constructed by initializing with random values and fine-tuning it during training. In this research, we use two types of LSTUR model which is LSTUR-con and LSTUR-ini. (2) NAML [11] employs additive attention [46] to encode users' preferences. This means that NAML utilizes an attention mechanism to capture and emphasize important aspects of a user's preferences when making recommendations. Additionally, our primary focus for this experiment will be on NAML, given its strong performance and adaptability, especially in accommodating adjustments to the included news information in the model. (3) NRMS [10] adopts a more intricate approach to learn user representations. It employs a two-layer encoder, consisting of multi-head self-attention [36] and additive attention. This design allows NRMS to capture and process user information from different perspectives, potentially leading to more comprehensive user representations for improved recommendations. (4) TANR [12] propose a new encoder that is trained to learn topic-aware news representations by simultaneously training it on an auxiliary topic classification task.

On the other hand, we examine one specific C-AW model to gain insights into their performance and effectiveness. We evaluate CAUM [13], there is a fusion of two key components: A candidate-aware self-attention network, which is used to model extensive connections among clicked news items while taking into account the specific candidate news; and a candidate-aware convolutional network (CNN) employed to capture immediate user interests from nearby clicks, also influenced by the content of the candidate news. Ultimately, the user's candidate-aware embedding is derived by considering both the long-range and short-term representations.

IV. Results & Analysis

Table II shows evaluation of the NRS has produced diverse performance results across various testing scenarios. These thorough assessments have uncovered valuable insights, particularly highlighting the competitive performance of two prominent models, NAML and CAUM, in terms of accuracy and relevancy. However, a notable distinction emerges as NAML demonstrates a significant advantage in terms of runtime efficiency, notably surpassing the CAUM model in processing speed. This unique combination of high performance and reduced computational overhead presents a compelling argument.

Using the NAML model as our main framework, we conduct a more in-depth assessment of how news information influences news representation and, consequently, its impact on news recommendation model performance. The results are shown in Table III. Our approach involves the independent processing of various news text information and categories. We merge these elements using an attention network, constructing a holistic representation that encompasses the collective impact of various news components. This method enables us to systematically investigate the synergistic effects of diverse news content on enhancing model performance.

Regarding news representation, the results in Table III reveal a key insight: relying on a single source of news information may not suffice for optimal representation. The notable improvement lies in not depending solely on a single type of news information but integrating various types, such as title, abstract, and tags. This approach is suggested to potentially result in better performance, emphasizing the need to leverage the complementary aspects of diverse news components. By incorporating a variety of information sources, the model can achieve a more comprehensive and nuanced understanding of news articles, ultimately improving the quality of news representation.

title	tags
Launched Tomorrow, Free Homecoming by State-Owned Enterprises (BUMN) Can Be Joined by 65,603 People	Free Homecoming, Homecoming Together with BUMN
"The Glory" Director, Ahn Gil Ho, Denies Being a Bully Perpetrator	"The Glory," "The Glory" Director, Director of "The Glory" Bullying

(a) A User clicked news articles in a session

title	tags
"APA", Female "Whisperer" of Mario Dandy Who Claims AG Received Poor Treatment from D	Mario Dandy Satrio, Mario Dandy Satrio Abuses the Child of GP Ansor's Official, Who is the Figure in the Mario Dandy Case, Who is the Figure, Anastasya Pretya Amanda
Head of Makassar Customs Admits Feeling Pressured	**Head of Makassar Customs Andhi Pramono, Head of Makassar Customs Clarified by the Corruption Eradication Commission (KPK)**
Not Present at the Scene, Former Girlfriend of Mario Objects to Being Linked to D's Assault	Mario Dandy Satrio, Mario Dandy Satrio Abuses the Child of GP Ansor's Official, Mario Dandy Case, Mario Dandy Breakup, Who is Mario Dandy's Whisperer?

(b) Recommendation based on clicked news using tags

title	tags
"APA", Female "Whisperer" of Mario Dandy Who Claims AG Received Poor Treatment from D	Mario Dandy Satrio, Mario Dandy Satrio Abuses the Child of GP Ansor's Official, Who is the Figure in the Mario Dandy Case, Who is the Figure, Anastasya Pretya Amanda
Korean Dramas and Indian Films Share Many Similarities, Here's the Explanation	India, Korea, Korean Drama, K-Drama, Indian Film
Ministry of Finance Holds 134 Tax Office Employees Have Shares Registered in Their Wives' Names	Ministry of Finance, Tax Office Employees, Tax Office Employees Have Shares

(c) Recommendation based on clicked news without using tags

Fig. 5. Recommendation results by using and without news tags for the same impression. The news clicked by the user in this impression is in blue and bold.

Table III provides further insights, notably indicating a significant improvement in model performance when news tags are incorporated alongside the abstract as part of the news information. This observation underscores the positive impact of integrating news tags into the model, particularly when combined with other textual elements like the abstract. Such integration significantly increases the performance of the NAML method by 0.88%, 3.11%, 5.86%, and 2.59% for AUC, MRR, NDCG@5 and NDCG@10, respectively.

The example of the NAML model using and without tags is illustrated in Fig. 5. Based on this example, it is evident

TABLE II. THE PERFORMANCE ON VARIOUS NEWS RECOMMENDATION MODELS. SIGNIFICANT DIFFERENCES WITH RESPECT TO BASELINES LSTUR/TANR/NRMS/CAUM ARE INDICATED USING †/ ‡ / ∗ /◇ FOR $p < 0.05$. RUN TIME IS REPRESENTED IN THE FORMAT HH:MM:SS

Methods	AUC	MRR	nDCG@5	nDCG@10	Run Time
LSTUR-ini†	0.5064	0.1377	0.1131	0.1885	00:53:26
LSTUR-con‡	0.5089	0.1426	0.1177	0.1939	00:53:05
TANR*	0.5038	0.1380	0.1136	0.1890	00:53:19
NRMS$^{\diamond}$	0.5266†‡*	0.1473†*	0.1273†‡*	0.2029†‡*	01:21:17
CAUM$^{\circ}$	0.5378$^{\dagger\ddagger*\diamond}$	0.1534$^{\dagger\ddagger*\diamond}$	0.1345$^{\dagger\ddagger*\diamond}$	0.2120$^{\dagger\ddagger*\diamond}$	03:44:50
NAML	0.5375$^{\dagger\ddagger*\diamond}$	0.1529$^{\dagger\ddagger*\diamond}$	0.1354$^{\dagger\ddagger*\diamond}$	0.2113$^{\dagger\ddagger*\diamond}$	01:51:31°

TABLE III. PERFORMANCE METRICS BEFORE AND AFTER ADDING THE NEWS TAGS. SIGNIFICANT DIFFERENCES ARE INDICATED USING † FOR $p < 0.05$ OR ‡ FOR $p < 0.001$.

News Information	AUC	MRR	nDCG@5	nDCG@10
Title	0.5247	0.1450	0.1236	0.2001
Abstract	0.5073	0.1400	0.1165	0.1909
Content	0.5311	0.1495	0.1296	0.2062
Title + Category/Subcategory	0.5399	0.1522	0.1338	0.2109
Title + Abstract	0.5253	0.1489	0.1295	0.2041
Title + Content	0.5345	0.1500	0.1302	0.2081
Title + Category/Subcategory + Abstract	0.5359	0.1502	0.1316	0.2081
Title + Category/Subcategory + Content	0.5394	0.1542	0.1371	0.2129
Title + Abstract + Content	0.5357	0.1518	0.1322	0.2095
Title + Category/Subcategory + Abstract + Content	0.5366	0.1509	0.1314	0.2088
Title + Tags	0.5216	0.1420	0.1195	0.1963
Abstract + Tags	0.5273‡	0.1503‡	0.1323‡	0.2059‡
Content + Tags	0.5317	0.1503	0.1314	0.2069
Title + Category/Subcategory + Tags	0.5389	0.1520	0.1337	0.2119
Title + Abstract + Tags	0.5248	0.1482	0.1288	0.2045
Title + Content + Tags	0.5273	0.1472	0.1253	0.2024
Title + Category/Subcategory + Abstract + Tags	0.5325	0.1506	0.1321	0.2085
Title + Category/Subcategory + Content + Tags	0.5415	0.1550	0.1379	0.2154
Title + Abstract + Content + Tags	0.5313	0.1499	0.1308	0.2066
Title + Category/Subcategory + Abstract + Content + Tags	0.5413†	0.1556†	0.1391†	0.2142†

that the model that incorporates news tags performs better in terms of relevancy than the one that does not use tags. In detail, when evaluating the top recommendations generated by the model without incorporating tags, none of the suggested articles align with user-clicked preferences. However, with the inclusion of tags, the model's recommendations exhibit a notable improvement. Two out of the top three suggestions are now articles that users have previously engaged with, showcasing the effectiveness of integrating tags in refining the recommendation outcomes.

TABLE IV. RESULTS ON DIFFERENT NUMBER OF CLICKED NEWS BY USERS

N Clicked News	AUC	MRR	nDCG@5	nDCG@10
1	0.5418	0.1561	0.1400	0.2160
5	0.5383	0.1516	0.1344	0.2104
10	0.5385	0.1544	0.1356	0.2130
20	0.5411	0.1533	0.1358	0.2132
50	0.5413	0.1556	0.1391	0.2142

The analysis of user interactions with news articles has yielded consistent findings across varying levels of user engagement with news content. These findings have been summarized and are presented in Table IV. Surprisingly, users who interact with only a few news articles demonstrate outcomes comparable to those who engage with a more extensive collection of articles. This intriguing observation suggests that the specific selection of the last N news items may not exert a substantial influence on the observed results. This insight prompts further exploration into the dynamics of user interactions and their impact on the recommendation process, potentially leading to more refined and nuanced approaches to personalizing news recommendations.

TABLE V. RESULTS ON DIFFERENT TEXT EMBEDDING

Embedding Model	AUC	MRR	nDCG@5	nDCG@10
word2vec-wiki-id	0.5327	0.1515	0.1323	0.2086
glove-wiki-id	0.5378	0.1534	0.1345	0.2120
fastText-wiki-id	0.5355	0.1533	0.1344	0.2124
fastText-indo4b	0.5365	0.1551	0.1378	0.2120
indolem/indobert-base-uncased	0.5264	0.1499	0.1300	0.2053
indobenchmark/indobert-base	0.5354	0.1497	0.1302	0.2083
indobenchmark/indobert-large	0.5383	0.1530	0.1338	0.2123

In addition to this research, we undertook an in-depth examination of different text embedding methods and their performance when integrated into the NAML model. The comprehensive results of these evaluations can be found in Table V. Our objective was to ascertain whether utilizing BERT-based models, which have demonstrated remarkable capabilities in various natural language processing tasks, would yield a significant performance boost within the NAML method.

However, upon thorough analysis, our findings suggest that the differences in performance between word embeddings and BERT-based models are not statistically significant when applied to the NAML model. In other words, the NAML model does not exhibit a substantial improvement in recommendation performance when integrated with BERT-based text embeddings compared to more conventional word embeddings. This emphasizes carefully choosing the right models and text embedding methods when optimizing recommendation systems.

V. Conclusion

This research investigates the effectiveness of neural news recommendation with attentive multi-view learning (NAML) method to perform news recommendation in the Indonesian language. We further propose to incorporate news tags information into the NAML architecture in order to improve the semantics of news representation, which is aimed at enhancing the Indonesian news recommendation system. According to our experimental results, the NAML method can significantly outperform some state-of-the-art neural models, such as LSTUR, TANR, and NRMS, in Indonesian news recommendation task. Although NAML demonstrates effectiveness similar to the CAUM method, it maintains superiority in terms of efficiency.

Furthermore, our investigation into news representation underscores the significance of diversifying news information sources. Combining multiple types of news data, rather than relying solely on a single source, has shown promising potential for enhancing model performance. Subsequently, the incorporation of news tags into the NAML architecture has demonstrated notable effectiveness, resulting in an enhancement of 3.11% for MRR and 5.86% for NDCG@5 in recommendation system performance.

In addition, we conducted performance assessments of the NAML model using various text embeddings. Upon initial examination, it becomes evident that there's no significant improvement observed when utilizing different types of word embedding as news representations, i.e., Word2Vec, FastText, Glove, and IndoBERT. Lastly, our analysis of user interactions with news articles, has unveiled intriguing findings regarding user engagement levels and their influence on recommendation outcomes. We found that users who interact with only a few news articles demonstrate outcomes comparable to those who engage with a more extensive collection of articles. This insight encourages further exploration into personalized news recommendation strategies, with the potential to refine and tailor recommendations based on a deeper understanding of user behaviors.

VI. Future Work

This study's findings offer insights and several promising directions for future research aimed at enhancing news recommendation systems. First, we can fine-tune the model parameters using a contrastive learning objective, specifically supervised contrastive loss. This approach can be employed to improve the separation between clicked and not-clicked news articles within the representation space. Second, it is potential to explore the automatic extraction of entities from news articles using NER (Named Entity Recognition) method as an integral part of the news recommendation model process. This approach can help uncover the relatedness of entities between relevant and non-relevant news articles, providing valuable context for improving recommendation robustness. Finally, research in news recommendation systems should diversify beyond contextual user history, adopting different methodologies for broader insights and improved system performance.

Acknowledgment

This research was funded by the Directorate of Research and Development, Universitas Indonesia, under Hibah PUTI Pascasarjana 2023 (Grant No. NKB-020/UN2.RST/HKP.05.00/2023).

References

[1] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, and D. Sampath, "The youtube video recommendation system," in *Proceedings of the Fourth ACM Conference on Recommender Systems*, ser. RecSys '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 293–296.

[2] A. S. Das, M. Datar, A. Garg, and S. Rajaram, "Google news personalization: scalable online collaborative filtering," in *WWW*, 2007, pp. 271–280.

[3] J. Freyne, M. Jacovi, I. Guy, and W. Geyer, "Increasing engagement through early recommender intervention," in *Proceedings of the Third ACM Conference on Recommender Systems*, ser. RecSys '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 85–92.

[4] C. Wu, F. Wu, Y. Huang, and X. Xie, "Personalized news recommendation: Methods and challenges," *ACM Trans. Inf. Syst.*, vol. 41, no. 1, jan 2023.

[5] X. Meng, H. Huo, X. Zhang, W. Wang, and J. Zhu, "A survey of personalized news recommendation," *Data Science and Engineering*, Sep. 2023.

[6] J. B. Baskoro and E. Yulianti, "Sgcf: Inductive movie recommendation system with strongly connected neighborhood sampling," *Jurnal Ilmu Komputer dan Informasi*, vol. 15, no. 1, pp. 55–67, Feb. 2022.

[7] Y. Liu, J. Miyazaki, and Q. Chang, "Jointly learning propagating features on the knowledge graph for movie recommendation," in *Database and Expert Systems Applications*. Cham: Springer International Publishing, 2022, pp. 3–16.

[8] Y. Liu and J. Miyazaki, "Knowledge-aware attentional neural network for review-based movie recommendation with explanations," *Neural Computing and Applications*, vol. 35, no. 3, pp. 2717–2735, Jan. 2023.

[9] M. An, F. Wu, C. Wu, K. Zhang, Z. Liu, and X. Xie, "Neural news recommendation with long- and short-term user representations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 336–345.

[10] C. Wu, F. Wu, S. Ge, T. Qi, Y. Huang, and X. Xie, "Neural news recommendation with multi-head self-attention," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6389–6394.

[11] C. Wu, F. Wu, M. An, J. Huang, Y. Huang, and X. Xie, "Neural news recommendation with attentive multi-view learning," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, ser. IJCAI'19. AAAI Press, 2019, p. 3863–3869.

[12] C. Wu, F. Wu, M. An, Y. Huang, and X. Xie, "Neural news recommendation with topic-aware news representation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1154–1159.

[13] T. Qi, F. Wu, C. Wu, and Y. Huang, "News recommendation with candidate-aware user modeling," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1917–1921.

[14] Adiyansjah, A. A. S. Gunawan, and D. Suhartono, "Music recommender system based on genre using convolutional recurrent neural networks," *Procedia Computer Science*, vol. 157, pp. 99–109, 2019, the 4th International Conference on Computer Science and Computational Intelligence (ICCSCI 2019) : Enabling Collaboration to Escalate Impact of Research Results for Society.

[15] M. Martijn, C. Conati, and K. Verbert, ""knowing me, knowing you": personalized explanations for a music recommender system," *User Modeling and User-Adapted Interaction*, vol. 32, no. 1, pp. 215–252, Apr. 2022.

[16] W. G. Assuncao, L. S. G. Piccolo, and L. A. M. Zaina, "Considering emotions and contextual factors in music recommendation: a systematic literature review," *Multimedia Tools and Applications*, vol. 81, no. 6, pp. 8367–8407, Mar. 2022.

[17] K. Dinnissen and C. Bauer, "Fairness in music recommender systems: A stakeholder-centered mini review," *Frontiers in Big Data*, vol. 5, 2022.

[18] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, p. 5–53, jan 2004.

[19] O. Phelan, K. McCarthy, M. Bennett, and B. Smyth, "Terms of a feather: Content based news recommendation and discovery using twitter." in *ECIR*, 2011, pp. 448–459.

[20] S. Okura, Y. Tagami, S. Ono, and A. Tajima, "Embedding-based news recommendation for millions of users," in *KDD*, 2017, pp. 1933–1942.

[21] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014.

[22] J. Lian, F. Zhang, X. Xie, and G. Sun, "Towards better representation learning for personalized news recommendation: a multi-channel deep fusion approach." in *IJCAI*, 2018, pp. 3805–3811.

[23] H. Wang, F. Zhang, X. Xie, and M. Guo, "Dkn: Deep knowledge-aware network for news recommendation," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, p. 1835–1844.

[24] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, and M. Zhou, "MIND: A large-scale dataset for news recommendation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 3597–3606.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *ArXiv*, 2013.

[26] E. Yulianti, N. Pangestu, and M. Jiwanggi, "Enhanced textrank using weighted word embedding for text summarization," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, pp. 5472–5482, Oct. 2023.

[27] S. Abdulateef, N. A. Khan, B. Chen, and X. Shang, "Multidocument arabic text summarization based on clustering and word2vec to reduce redundancy," *Information*, vol. 11, no. 2, p. 59, Jan 2020.

[28] M. Gambhir and V. Gupta, "Deep learning-based extractive text summarization with word-level attention mechanism," *Multimedia Tools and Applications*, vol. 81, no. 15, pp. 20 829–20 852, Jun. 2022.

[29] T. V. Rampisela and E. Yulianti, "Academic expert finding in indonesia using word embedding and document embedding: A case study of fasilkom ui," in *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 2020, pp. 1–6.

[30] R. C. Lima and R. L. T. Santos, "On extractive summarization for profile-centric neural expert search in academia," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 2331–2335.

[31] N. Nissa and E. Yulianti, "Multi-label text classification of indonesian customer reviews using bidirectional encoder representations from transformers language model," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, pp. 5641–5652, Oct. 2023.

[32] B. Jang, M. Kim, G. Harerimana, S.-u. Kang, and J. W. Kim, "Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism," *Applied Sciences*, vol. 10, no. 17, p. 5841, Aug 2020.

[33] N. E. Aoumeur, Z. Li, and E. M. Alshari, "Improving the polarity of text through word2vec embedding for primary classical arabic sentiment analysis," *Neural Processing Letters*, vol. 55, no. 3, pp. 2249–2264, Jun. 2023.

[34] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543.

[35] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[37] A. F. Aji, G. I. Winata, F. Koto, S. Cahyawijaya, A. Romadhony, R. Mahendra, K. Kurniawan, D. Moeljadi, R. E. Prasojo, T. Baldwin, J. H. Lau, and S. Ruder,

"One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 7226–7249.

[38] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 757–770.

[39] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 843–857.

[40] Y. Takeda, I. Ohmukai, S. ADACHI, and T. Kato, "News recommendation method using group tags and thesauruses," *Transactions of Japan Society of Kansei Engineering*, vol. 8, no. 3, pp. 813–818, 2009.

[41] Y. Shen, P. Ai, Y. Xiao, W. Zheng, and W. Zhu, "A tag-based personalized news recommendation method," in *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2018, pp. 964–970.

[42] A. Iana, G. Glavas, and H. Paulheim, "Simplifying content-based neural news recommendation: On user modeling and training objectives," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 2384–2388.

[43] D. Valcarce, A. Bellogín, J. Parapar, and P. Castells, "Assessing ranking metrics in top-n recommendation," *Information Retrieval Journal*, vol. 23, no. 4, pp. 411–448, Aug. 2020.

[44] C. Wu, F. Wu, and Y. Huang, "Rethinking infonce: How many negative samples do you need?" in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 2509–2515, main Track.

[45] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734.

[46] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *ArXiv*, vol. 1409, 09 2014.

Telemedicine Adoption for Healthcare Delivery: A Systematic Review

Taif Ghiwaa[1], Imran Khan[2], Martin White[3], Natalia Beloff[4]
Department of Computer Science and Information Systems, King Khalid University, Abha, Saudi Arabia[1]
Department of Informatics, University of Sussex, Brighton, United Kingdom[1,2,3,4]

***Abstract*—Telemedicine is the delivery of healthcare services using telecommunication and information technologies. The adoption of telemedicine has been promoted by advancements in technology, increased accessibility to the Internet, and the need for convenient and efficient healthcare delivery. Understanding the theoretical foundations of telemedicine adoption among healthcare providers and patients is crucial for successful acceptance and utilization. This systematic review aims to explore the theoretical frameworks and models that have been widely utilized to understand telemedicine adoption among healthcare providers and patients. A systematic search was conducted across two popular electronic databases, resulting in the inclusion of 21 relevant studies. The selected studies were analyzed to identify the theoretical perspectives employed in telemedicine adoption research. The key findings reveal that the Technology Acceptance Model (TAM), the Unified Theory of Acceptance, and the Use of Technology (UTAUT) model are the most widely models used to illustrate the factors affecting telemedicine adoption among healthcare providers and patients through different countries and telemedicine contexts. Understanding these theoretical models is crucial for policymakers and healthcare professionals as it can provide insight into the key factors influencing the widespread adoption of telemedicine. This knowledge can serve as a guidance for crafting initiatives, and tailoring policies to promote the successful acceptance and utilization of telemedicine among providers and patients in diverse healthcare environments.**

***Keywords*—Telemedicine; systematic review; technology acceptance model; adoption; telehealth; healthcare provider; patient**

I. Introduction

The integration of Information and communication technology (ICT) in healthcare systems has revolutionized the delivery of healthcare services. Telemedicine, a branch of ICT, has emerged as a promising approach for delivering care remotely, overcoming distance barriers, enhancing clinical outcomes, increasing patient engagement [1], and reducing costs [2], [3]. It can be used to provide a range of clinical services, including consultations, diagnosis, treatment, monitoring, therapy process, and exchange of medical information by using electronic communication tools such as video conferencing, phone calls, or secure messaging [4]. Broadly, telemedicine includes two different types of services which are: store and forward (asynchronous), and real-time interactive (synchronous) [5], [6]. Thus, telemedicine is an excellent opportunity for health professionals to reach a wider community through remote provision of healthcare services.

Telemedicine and telehealth are often used interchangeably, but they have slightly different meanings [7]. Telemedicine specifically refers to the remote delivery of clinical healthcare services, while telehealth includes a broader scope of remote healthcare services, including both clinical and non-clinical aspects [4], [8]. Due to the inconsistent usage of the term telemedicine in the studies, it becomes difficult to define it precisely in relation to other terms. Therefore, during the search process, related terms like eHealth, mHealth, and telehealth were also taken into consideration. The studies included in the analysis focused on telemedicine as a means of providing patient-centered healthcare services over long distances.

An increasing world population, especially the elderly, [9] will require access to remote healthcare services, such as that potentially offered by integrating telemedicine with traditional healthcare practice. The COVID-19 pandemic has significantly accelerated this trend of adopting telemedicine [10]. With the pandemic forcing many people to stay at home, the use of telemedicine has become even more necessary to ensure that patients can continue to receive uninterrupted care. Accordingly, telemedicine has become an essential component of healthcare service delivery. Despite the great promise of telemedicine, its actual use is insufficient and has not achieved a prominent utilization outcome [11], [12]. The reasons behind this are not only the technical aspect but also the human behavioral aspect [13]. Thus, the acceptance of technology plays a crucial role in successfully implementing and consistently using it and is considered a significant factor in ensuring the effective implementation of IT systems.

Investigating the literature reveals a number of models are useful in understanding individuals' intentions to adopt ICT [14]. Some of these models include the Technology Acceptance Model (TAM) [15], [16], Unified Theory of Acceptance and Use of Technology (UTAUT) [17], Diffusion of Innovations Theory (DOI) [18], Theory of Planned Behavior (TPB) [19], Theory of Reasoned Action (TRA) [20], and Health Belief Model (HBM) [21]. These models introduce different factors that influence end users' behavior to adopt telemedicine. However, limited evidence exists regarding the optimal theory or model for understanding the acceptance of telemedicine in the realms of technology adoption and acceptance.

While individual studies explore theoretical constructs as predictors for telemedicine acceptance, there is a notable absence of a comprehensive overview that systematically analyzes these constructs, models, and factors influencing the acceptance of various types of telemedicine. This evaluation is essential from the standpoint of providers and patients, who are pivotal users of the telemedicine system. Furthermore, the authors of [22] and [23] pointed out the gap in knowledge, emphasizing that existing models are limited in scope and constrained by regional or national borders. Therefore, before

suggesting a model to study telemedicine adoption, it is important to study systematically these models. Thus, this review is conducted to provide a comprehensive overview of the theories and models used to assess the behavioral intention in adopting telemedicine among healthcare providers and patients in different settings of telemedicine. The research will seek to answer:

RQ1. Which adoption theories and models are widely applied in the telemedicine context?

RQ2. What are the most prominent factors affecting telemedicine technology adoption from the end user's perspective in a different setting of telemedicine?

The rest of the paper is organized as follows: the study methods and materials are discussed in Section II, followed by Section III detailing our results. Finally, the discussion in Section IV is succeeded by an exploration of future research directions and limitations, followed by the conclusion in Section V.

II. Materials and Methods

A. Search Strategy

This systematic review was conducted based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure a systematic and transparent review process [24]. Two popular academic databases, including Scopus and PubMed, were selected to collect relevant studies published in the last decade (between 2012 and 2023) and available in English or Arabic languages. The Population, Intervention, Comparison, Outcome, and Study design (PICOS) framework was applied to establish the inclusion and exclusion criteria for the review [25] (see Table I). The inclusion criteria encompassed studies focusing on telemedicine adoption theories and models among healthcare providers and patients. The search terms used a combination of PICOS components and were refined using Boolean operators. The search was conducted in March 2023. A specific search string can be found in Appendix 1.

TABLE I. Inclusion and Exclusion Criteria

	Inclusion criteria	Exclusion criteria
Population	Patients, healthcare providers, nurses.	Not patient, not healthcare provider, not nurse.
Intervention	Telemedicine	Not telemedicine, no clinical services delivered to patients, no interaction with healthcare providers.
Outcome	Behavioral intention in adoption and acceptance of technologies	Not based on theory, no adoption or acceptance.
Study design	studies published in English or Arabic language,full text access.	Reviews, studies not published in English or Arabic language.

B. Search Result

The result of searching selected databases initially yielded 3,753 articles through electronic searches. In addition, manual searching yielded five articles from other sources. After removing duplicates and screening titles and abstracts, 230 articles were selected for full-text review. Finally, a total of 21 articles, meeting our inclusion criteria, were chosen for in-depth analysis. Fig. 1 displays the PRISMA flow diagram for the search and selection process.

C. Data Extraction and Organization

Data from the included studies were extracted using a standardized form (see Table II). The extracted data included study characteristics (e.g., authors, year of publication, journal, the country where the study was conducted), theoretical frameworks employed, study design, population target, sample size, key constructors affecting telemedicine adoption in different contexts, and type of statistical analysis used in the study. The findings were summarized to fulfill the review's objectives by focusing on identified theories and their constructs in understanding telemedicine adoption among healthcare providers and patients.

TABLE II. Data Extraction Form

Data Extraction	Description
Study ID	Identifier of the study.
Title	Title of the study.
Author/s	Author/s name.
Year	Year of the publication.
Journal	Published Journal.
Country	Country/place where the research conducted.
Telemedicine application	Type of Telemedicine application.
Theory/Model	Type of theory/model used in the study.
Data collection Design	Method used to collect data.
Population Focus	The target of research participants.
Population Sampling	Number of research participants.
Components of theory/model	Constructor used to build theory/model.
Moderator components	A variable that influences the presence of a relationship between variables.
Statistical Analysis type	Type of analysis used to obtain the result.

D. Quality of the Studies

A quality assessment checklist is essential for evaluating the methodological rigor and reliability of studies [26]. The quality assessment checklist of the included studies was adapted from [27] and includes five items, as presented in Table III. This checklist uses a 3-point scale where (1=Yes, 0.5=Partly, 0=No). The results of the quality assessment can be found in Appendix 2. Generally, all the included studies scored high in quality and passed the quality assessment, allowing us to proceed to the next step, which is analysis.

TABLE III. Quality Assessment Questions

No.	Question
1	Is the study related to telemedicine adoption and its application?
2	Does the study use adoption theories or models?
3	Does the study explicitly present the research methodology?
4	Does the data collection procedure outline in the study?
5	Are the study findings clearly presented and added to the literature?

III. Results

A. Characteristics of Included Studies

After reviewing studies from the last decade, the number of publications remained limited, with only a few studies until 2018. It was in that year when there was a noticeable surge in interest regarding telemedicine adoption, a trend that continued

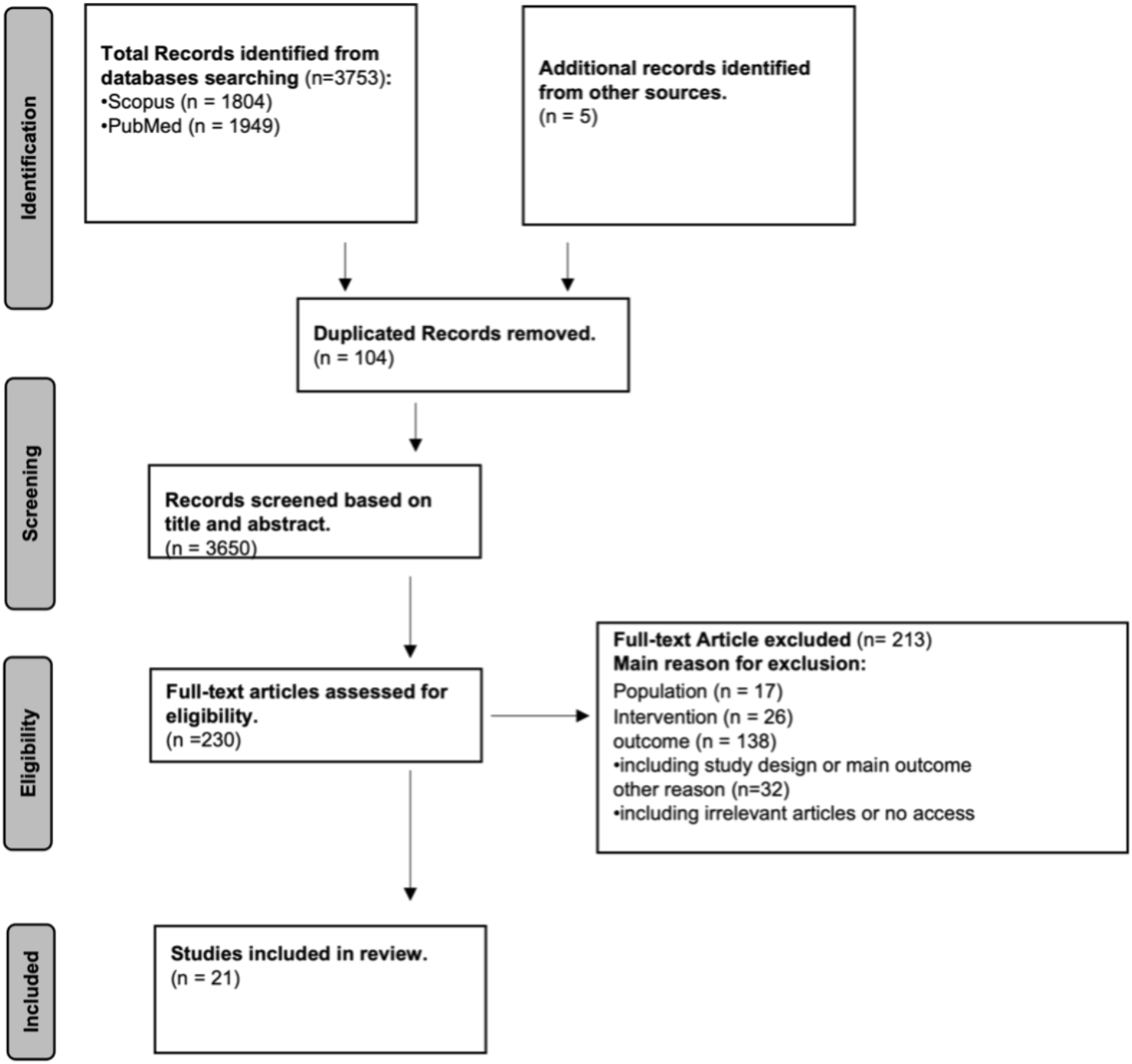


Fig. 1. The PRISMA flow diagram of studies search and selection.

during and beyond the COVID-19 pandemic. Fig. 2 illustrates the growth of publications over the last decade.

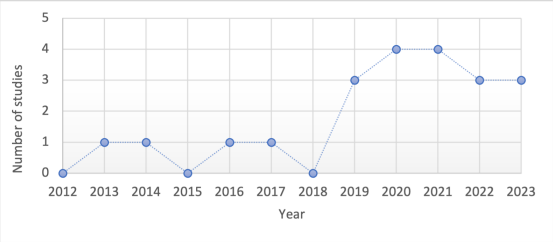


Fig. 2. Number of articles published by year.

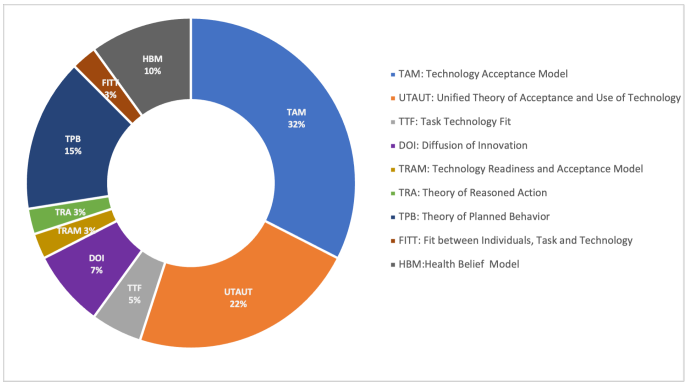


Fig. 3. Frequency of theories and models used to explain the adoption of telemedicine.

As previously mentioned, various frameworks, theoretical models, and their extensions have been developed to comprehend user intentions for adopting ICT, particularly in the context of telemedicine. Among the 21 studies included in the review, TAM and UTAUT with their extension were identified to be the highest models used compared to others. As shown in Fig. 3, the TAM model and its extension were used 13 times, UTAUT was used nine times, TPB six times, HBM four times, DOI three times, and the remaining models were used only once. This diversity of models reflects the complexity of telemedicine adoption process and underscores the importance of a different approach to understanding users' attitude toward this innovation.

The distribution of countries related to the included telemedicine studies varies. Developed countries contributed the most, with 14 studies, in contrast to developing countries, which had 6 studies. The USA had the highest number of studies, with 4, followed by Australia with 3 studies. Each of the following countries Netherlands, Canada, and Germany had 2 studies, while France had only one. In developing countries, China had 2 studies, whereas the remaining countries, including the Philippines, Malaysia, India, and Saudi Arabia, had one study each. This international diversity in research reflects the global significance and varied perspectives on telemedicine adoption especially in the developing countries.

In terms of research design, most of the studies relied on quantitative methodologies. One study employed a mixed-method approach, while 5 studies utilized qualitative research methods, including interviews (4 studies) and focus groups (1 study). There is a round balancing between publications focused on patients (12 studies) and those directed to providers (9 studies), covering a wide array of telemedicine applications. Details of the most important extracted data from these studies are presented in Table IV. For the complete set of extracted data, please refer to Appendix 3.

B. Telemedicine Applications

Telemedicine has a wide range of applications that are transforming the way healthcare services are accessed and delivered worldwide. Remote consultations, telerehabilitation, remote monitoring, mental health services, tele palliative care, teledermoscopy services, and teleneurology are examples of telemedicine applications that were covered and analyzed in the included studies. Among the included studies, a total of 5 studies focused on telemedicine services in general [13], [28]–[31], and the other 5 studies on telemonitoring [32]–[36]. A total of 3 studies were dedicated to remote mental health services - Telepsychotherapy [37]–[39], while 4 studies centered around telerehabilitation [40], [41], and tele palliative care [42], [43], each with 2 studies. The remaining studies were designed for teleneurology [44], teledermoscopy [45], teleconsultations [46], and group-based telemedicine [47], each with one study. Telemedicine is reshaping the landscape of healthcare, with a diverse array of applications that have been examined in recent studies.

Each application of telemedicine has its own factors that may influence its users' intentions, whether for the patient or provider. According to the analysis of the studies [13], [32]–[34], [39], [43], [45], [47], the most important factors shared by most studies of telemedicine applications are perceived usefulness and ease of use, which are key determinants of users' attitudes and behavioral intentions for the TAM model. This indicates that telemedicine applications that are easy to understand and bring desired benefits as needed have a higher chance of influencing user behavior to adopt the applications [13], [28], [35].

Other significant factors repeated in most studies are social influence and facilitating conditions from the UTAUT model. The literature highlighted that social influence is a significant predictor because the opinions of colleagues for providers and family or friends of patients strongly influence user behavior. Additionally, the availability of technological and organizational support shows a positive connection with usefulness and ease to use the technology [48], [49], thus, it has a better chance of influencing user behavior to adopt telemedicine. For an overview of the factors affecting each telemedicine application among healthcare providers and patients, see Appendix 4.

C. Overview of Telemedicine Adoption Factors

Telemedicine adoption is influenced by numerous factors, which are rooted in various theoretical frameworks to adopt the technology. Although some factors are synonyms to each other for example, insecure and perceived risk, it has been classified and counted its frequencies based on the identical general terminology (have the same meaning with different terms). These factors have been categorized into 5 groups, adapted

TABLE IV. MOST IMPORTANT EXTRACTED DATA FOR THE INCLUDED STUDIES

Study	Year	Country	Data collection method	Theory/Model	Constructs
Patient					
[28]	2020	KSA	survey	UTAUT, TTF	Awareness, Self-efficacy, Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Condition, TTF
[37]	2022	USA	survey	UTAUT, TAM3, TPB	Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Condition, Anxiety, Attitude
[29]	2021	China	survey	-	Trust the sponsor of a healthcare website, Gender, Age, Educational level, City Income level, Consumer type
[45]	2016	Australia	survey	TAM, TRA, DOI, UTAUT	Perceived Usefulness, Ease of Use, Trust, Attitude/intention, Subjective Norm, Compatibility, Facilitator
[40]	2019	Netherlands	interview	UTAU	Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Condition
[30]	2023	India	survey	Push-Pull Mooring(PPM), HBM,UTAUT2	Push Effects (Inconvenience and Perceived Healthcare Risk), Pull Effects (Opportunity for alternatives and Ubiquitous care) , Mooring effects (Trust in telemedicine) , Inertia (Habit, Switching Cost).
[46]	2023	Germany	survey	UTAU, External variables	Performance Expectancy, Effort Expectancy, Social Influence, Computer Proficiency, Knowledge about digital health care solutions (Awareness), Depressive symptoms.
[33]	2023	USA	survey	TAM	Perceived Usefulness, Ease of Use.
[47]	2019	USA	mixed method	TRAM, External variables	Perceived Usefulness, Ease of Use, Innovativeness, Optimism, Discomfort, Insecurity, Group readiness, HIV-related privacy concerns.
[34]	2019	USA	interview	FITT, UTAUT2,TAM, TAM3	Fit between individuals and task: (Motivation/ Engagement, Self-efficacy) Fit between individuals and technology: (Preference for device design, HIV status, Customized alert ,Ease of use) Fit between task and technology: (System functionality, Self-awareness).
[35]	2021	France	survey	UTAUT,HBM, UTAUT2	Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Condition, Perceived risk, Financial cost
[36]	2021	Not mentioned	survey	TAM, TPB, TAM3,HBM	Interpersonal Influence, Personal Innovativeness, Trustworthiness,Attitude ,Self-efficacy, Health Interest ,Perceived Value.
Healthcare Provider					
[32]	2021	Australia	survey	TAM, TPB	Perceived Usefulness, Ease of Use, Attitude.
[39]	2013	Canada	survey	TAM, TPB	Perceived Usefulness, Ease of Use, Attitude.
[38]	2020	Germany	Interview/focus group	DOI, External variables	Perceived benefit, Avaliability of designed room.
[42]	2022	Netherlands	survey	UTAUT,TPB,TAM3,DOI	Outcome expectancy, Effort expectancy, Facilitating Condition,Social Influence, Attitude, Anxiety, Self-efficacy,Personal Innovativeness.
[41]	2020	Australia	interview	TAM,HBM,External variables	Context of use,Perceived Benefits,Technical and connectivity issues Client capability and compatibility,Lack of physical presence,Balancing the service and user needs.
[31]	2017	China	survey	-	Authenticity and reliability of data, Awareness, Previous experience.
[44]	2022	Philippine	survey	UTAUT, TPB	Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Conditions, Attitude.
[13]	2014	Malaysia	survey	UTAUT, TAM, TPB,TAM3, External variables	Perceived Usefulness, Perceived Ease of Use, Attitude, Self-efficacy, Organizational Culture, Facilitating Conditions.
[43]	2020	Canada	interview	TAM	Perceived Usefulness, Perceived Ease of Use.

from [27], to study the adoption of telemedicine among healthcare providers and patients. These categories include individual factors, organizational factors, technological factors, security factors, and health factors. The following sections will discuss the factors of each category separately. Fig. 4 shows the factors influencing healthcare providers and patients based on the aforementioned classifications. The results show that, individual factors play a vital role in the successful adoption of telemedicine for both providers and patients as it occupied around half of the percentages among other factors. Additionally, technological, and organizational factors are considered more important for providers than patients.

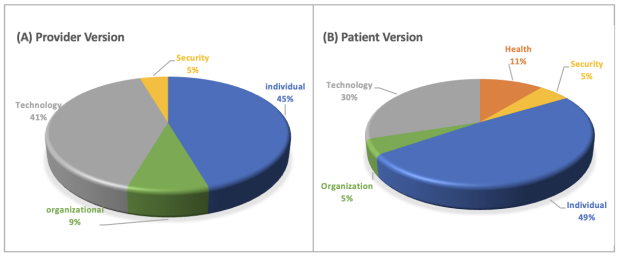


Fig. 4. Classify factors affecting the adoption of telemedicine from **(A)** provider's perspective. **(B)** patient's perspective.

1) Individual Factors: As a definition, individual factors refer to the personal attributes, beliefs, attitudes, and characteristics of an individual that influence their decision to accept, adopt, or resist the use of new technology [50]. Individual factors for telemedicine adoption represented 45% for providers and around half for patients. Fig. 5 summarizes the individual factors and analyzes their frequencies based on the number of using them in the included studies from both providers' and patients' perspectives. The findings of the analysis showed that social influence is the most important factor influencing patients which was repeated 6 times [28], [35], [37], [40], [45], [46]. The second most important factor is attitude, which was repeated 3 times [36], [37], [45], followed by personal innovativeness [36], [47], self-efficacy [34], [36], awareness [28], [46], and habit [30], [40], each of which was repeated 2 times. Furthermore, the remaining set of factors include group readiness [47], optimism [47], motivation, engagement [34], compatibility [45], and other demographic characteristics such as age, gender [29], [35], computer skills, education, and socioeconomic status [29], [34], [46], each of which was mentioned only once. Finally, the lack of control of technology was mentioned in one study as a barrier [47].

Contradictory to that, the most important factor influencing providers is the attitude which was mentioned 5 times [13], [32], [39], [42], [44]. Other important factors were repeated 2 times including social influence [42], [44], self-efficacy [13], [42], and experience [31], [44]. Additionally, personal innovativeness [42], awareness [31], voluntariness [44], and client capabilities [41] factors were mentioned one time for each as a success predictor. However, anxiety [42] and lack of physical presence [41] were reported once for each as barriers.In summary, the analysis of individual factors affecting telemedicine adoption reveals that social influence and attitude are key drivers for patients, while providers are primarily influenced by their attitude.

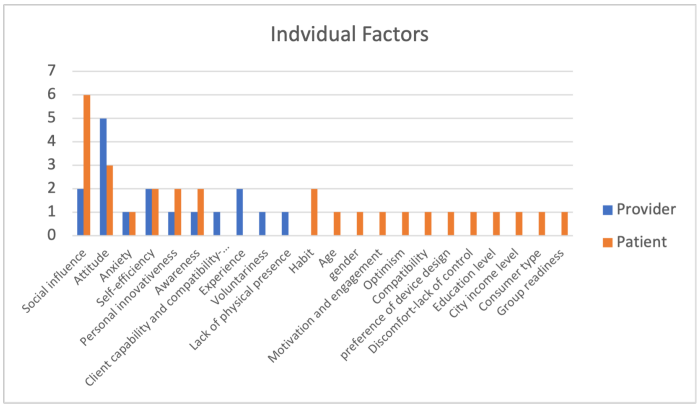


Fig. 5. Individual factors that influence the adoption of telemedicine for providers and patients.

2) Organizational Factors: Organizational factors refer to the various aspects and characteristics within healthcare institutions that influence the overall organization's operations, performance, efficiency, and quality of care provided. These characteristics including as examples infrastructure, resource allocation, and training programs [51].The review showed that the facilitating conditions factor is the significant facilitator for adopting telemedicine where it was repeated in the included studies 3 times for providers [13], [42], [44] and 5 times for patients [28], [35], [37], [40], [45]. Government policy, top management support, project team capacity, and external suppliers' capacity are examples of the facilitators mentioned in the included studies from the providers' perspective whereas technical and connectivity issues were mentioned as barriers for patients.

Another important factor for patients was the type of the hospital which was mentioned once [29]. On the other hand, the organizational culture factor influences providers' intention to adopt telemedicine and it was mentioned once in the included studies [13]. It is noteworthy that, the culture and values within an organization can either facilitate or hinder technology adoption and that rely on reinforcement of the top management of the organization and its policies to the adaptability, change, and support the innovation through the use of technology [13], [52]. Additionally, the findings of one study confirmed that patients tend to trust public hospitals more than private ones, thus affecting their intention to adopt telemedicine services provided by these hospitals [29]. In summary, organizational factors including facilitating conditions, government policy, and organizational culture shape the adoption of telemedicine in healthcare institutions.

3) Technological Factors: The technological factors for telemedicine adoption represented 41% for providers and 30% for patients. It refers to the technological components and considerations in the design, implementation, communication infrastructure, and related technologies that enable the delivery of remote healthcare. As shown in Fig. 6, the most common factors for patients were effort expectancy and performance expectancy [28], [35], [37], [40], [46], which were derived from the UTAUT model. They were mentioned in the included studies 5 times each. Other important factors that influenced the patients were perceived usefulness and ease to use [33],

[45], [47] which stem from the TAM model, and they were repeated 3 times for each. Task technology fit (TTF) and cost were also considered factors for patients which were repeated 2 times each. The remaining set of factors was repeated once in the included studies including perceived benefit [36], availability of the service [30], system functionality [34], customizing the functionality of the device [34], and computer proficiency [46].

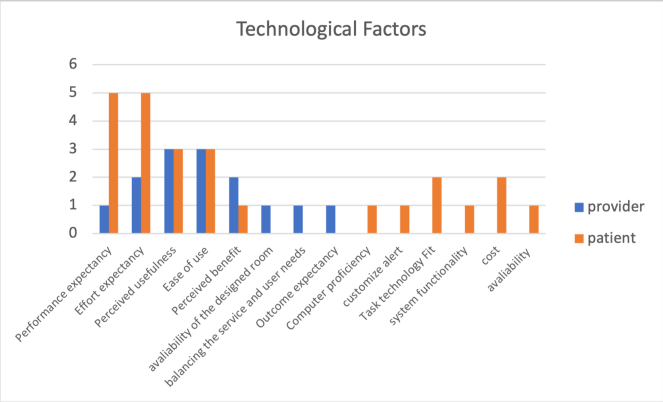


Fig. 6. Technological factors that influence the adoption of telemedicine for providers and patients.

In contrast, the most important factors for providers were perceived usefulness and ease to use which were mentioned 3 times [32], [39], [43]. Indeed, when the providers find telemedicine user-friendly and help them to improve their work with less effort then, they are more likely to integrate telemedicine into their practice and adopt it. Furthermore, other important factors from the providers' perspective were effort expectancy [42], [44] and perceived benefit [38], [41] were mentioned 2 times for each factor. Finally, performance expectancy [44], balancing the service and users' need [41], outcome expectancy [42], and availability of designed room [38] were facilitators that influenced the adoption and they were repeated once for each factor. In discussing the perceived benefits of telemedicine, the literature highlights that it offers a wide range of benefits that positively impact their practice, patient care, and overall healthcare delivery. These benefits contribute to improved efficiency of time and cost, help patients access support, and enhance the quality of care [38], [41], [53]. In summary, the adoption of telemedicine is significantly influenced by various technological factors that differ between providers and patients, with both groups placing a strong emphasis on the ease of use and perceived usefulness.

4) Security Factors: The security factors related to telemedicine adoption represented 5% for each provider and patient. It is defined as the level of protection and assurance that healthcare providers and patients have regarding the confidentiality, integrity, and availability of sensitive information and data transmitted, stored, or accessed during telemedicine interactions [54]. The most important security factor from the providers' perspective is the authenticity and reliability of data from remote patient monitoring, and it serves as a barrier that negatively influences telemedicine adoption [31]. Whereas the patients' behavioral intention was influenced by two factors: perceived trust and perceived risk which were repeated in the included studies four and two times respectively. The perceived trust ranges between privacy [47], trust in general [36], trust in the organization [29] and trust in telemedicine [45]. Moreover, the included studies identify the perceived risk in the context of losing to reach the desired outcome when using technology as it may not work properly [35], [47]. So, all security factors are considered barriers. This is due to the sensitive nature of healthcare data and the potential risks associated with remote communication and data exchange.

5) Health Factors: The factors related to health are likely to influence the behavioral intention of patients just, and it represented 11% among other factors. As it was defined, health factors refer to both the benefits and challenges that telemedicine brings to the healthcare landscape [55]. Health interests, perceived health risks, health conditions, and depressive symptoms are factors identified in the included studies once for each factor. The perceived health risk associated with telemedicine includes risk management, human resource, clinical risk, technology risk, and regulatory issues [30]. With regard to health interests, when individuals have a strong interest in maintaining their health and accessing healthcare services, they are more likely to adopt telemedicine as a convenient and accessible option [36]. Moreover, people with depressive symptoms are more likely to seek for professional treatment as telemedicine gives the option to receive treatment from home thus, adopting and accepting care through telemedicine [46].

IV. Discussion

Studying the behavioral intention of end users to adopt telemedicine has attracted research attention recently. The reasons behind that are the importance of understanding user perspectives, predicting adoption rates, and designing effective strategies to promote telemedicine adoption. By taking users' intentions and preferences into account, healthcare organizations can foster a positive telemedicine experience and maximize the potential benefits of this innovative healthcare delivery method.

The fundamental goal of this review is to summarize the different factors based on various acceptance theories that might influence healthcare providers and patients to adopt telemedicine through different applications. The findings of the review showed that the TAM and UTAUT models are the most important models in adopting and accepting telemedicine. This result is consistent with the study conducted to evaluate the different acceptance models and theories in the healthcare sector [56]. From a user-centered perspective, it has been noticed that the UTAUT model was used more frequently by patients than by providers to explain telemedicine acceptance. This is due to its inclusion of variables associated with social influence, which is likely to be more important for patients than for providers. In contrast, the TAM model was used for investigating the acceptance of providers more than for patients. As it was mentioned earlier, the functional characteristics of the technology to be free of effort and compatible with the desired need are good determinants for adopting telemedicine by users.

Furthermore, it has been found that some studies use a single theoretical model [33], [40], [43], while others use more than one model [13], [28], [32], [34]–[37], [42], [44],

[45]. Although the TAM and UTAUT are the based models in most literature as it was mentioned above, other factors from different models were integrated with them to represent the acceptance in different contexts. TRA, TPB, DOI, HBM, and TTF are the models that were integrated parts of their predictors with the based model. According to Barrettee, it is useful to combine more than one model as each model focuses on different aspects of technology adoption, such as individual perceptions, organizational factors, or external influences. By combining multiple models, researchers can obtain a holistic view of the adoption process, considering various factors that may influence technology acceptance [57].

Another important factor integrated with the base model in 4 studies is perceived trust [29], [30], [36], [45]. It was observed to be essential for patients to accept modern healthcare delivery. Accordingly, healthcare organizations and telemedicine providers must proactively address trust-related concerns and demonstrate a commitment to maintaining the highest standards of security, privacy, and quality care delivery. Building and maintaining trust in telemedicine services are key to establishing a sustainable and patient-centric approach to modern healthcare delivery. One of the most important factors added to the base model in 8 studies [13], [32], [36], [37], [39], [42], [44], [45] is attitude from TPB model. The TPB model suggests that an individual's attitude toward a specific behavior strongly influences their intention to perform that behavior [19]. This implies that a positive attitude toward using telemedicine services is likely to promote its adoption among healthcare providers and patients. So, if individuals believe that telemedicine is eased to use and offers tangible benefits, such as convenience, improved access to care, improve work efficiency, and time savings, they are more likely to have a positive attitude toward its adoption [13], [32].

Additionally, the findings revealed that self-efficacy significantly influences adoption. It refers to an individual's belief in their ability to use technology effectively [36]. As stated by Bandura in his social cognitive theory [58], individuals' behaviors are influenced by their own capabilities to perform a particular task. So, patients with high self-efficacy are more likely to actively engage and accept telemedicine to receive healthcare. Besides that, self-efficacy can influence the healthcare providers' willingness to adopt and integrate telemedicine into their practice, where high self- efficacy makes them feel more confident in their ability to use virtual communication tools, maintain patient engagement, and conduct remote consultations and treatment effectively [42], [59].

The review also explored different key factors that were extensively employed in the included literature with the base models to understand the acceptance including awareness, innovativeness, and habit. According to Chen et al. [31] , users with positive attitudes toward telemedicine reflect their great awareness of service benefits. Consequently, there is a direct relationship between users' awareness and their attitude, and thus their intention toward adopting telemedicine. Moreover, it is has found that innovativeness as a perceived advantage makes healthcare more accessible and efficient for a broader population [42], [47]. As a result, it improves the quality of healthcare delivery and maximizes patient engagement then its adoption. Regarding habit, authors in [30], and [40] indicate that building positive habits around telemedicine usage is crucial for its widespread adoption and long-term success. Consequently, telemedicine can become an established and habitual part of modern healthcare practices by addressing any barriers related to usability, trust, convenience, and positive outcomes.

A. Limitations and Future Research

The review has a few limitations, which gives an opportunity for further research. First, this review used only two databases for retrieving relevant studies. Including more databases could lead to richer results. Additionally, language bias could affect the results, papers not published in English or Arabic language were excluded. Third, the focus of the population in the included studies is for patients and healthcare providers including physicians and nurses. Future studies are encouraged to encompass different sets of the individual such as administrators, and health professionals as it may lead to in-depth knowledge and a full picture of the adoption process.

V. Conclusion

Telemedicine is an evolving field of healthcare that has revolutionized the way patients receive medical attention remotely. This systematic literature review identified the theoretical constructs associated with end-user adoption and acceptance of telemedicine. When reviewing included studies, it is obvious that TAM and UTAUT are the most widely technology acceptance models applied to the telemedicine context. Additionally, the constructs of TAM and UTAUT models were the most deployed factors to evaluate the acceptance and adoption of telemedicine. Adding to that other factors were integrated with the previous constructs including, attitude, self-efficacy, perceived trust, innovativeness, and habit.

Existing studies include various applications of telemedicine. Teleconsultations, telerehabilitation, telemonitoring, telepsychotherapy, tele palliative care, teledermoscopy services, and teleneurology are some examples of applications that are covered in this review. While some factors apply to all applications of telemedicine, others are validated to specific applications. Therefore, understanding the nuanced factors that impact the success of telemedicine in diverse healthcare contexts is crucial for optimizing patient care and healthcare delivery.

The review provided a classification analysis of the factors that influence telemedicine adoption among healthcare providers and patients. These categories include individual, organizational, technological, security, and health factors. In general, it has been noticed that individual factors occupied the largest percentage among other factors for both providers and patients. Social influence and attitude are the most significant factors at the individual level. At the organizational level, facilitating conditions is an essential factor for telemedicine adoption by both parties' providers and patients. Furthermore, perceived usefulness, ease of use, effort expectancy, and performance expectancy are important influencers at the technological level. Besides that, perceived trust was found to be a significant factor at the security level. Finally, at the health level, health interests, perceived health risks, health conditions, and depressive symptoms were identified to influence the patients' intention to adopt telemedicine.

The results of this review can provide insights to policymakers and healthcare organizations on the factors that influence end-user behavioral intention to adopt the modern healthcare delivery method. Understanding these factors is pivotal in crafting effective strategies for the widespread implementation of telemedicine. Moreover, recognizing the unique individual characteristics of end-users, such as their technological proficiency, awareness, and subjective norms, is essential. Tailoring telemedicine initiatives to meet the specific needs of diverse patient populations can significantly enhance the acceptance and utilization of this innovative approach to healthcare delivery. By acknowledging and accommodating these individual differences, healthcare systems can maximize the potential benefits of telemedicine, leading to improved patient outcomes and more efficient healthcare services.

REFERENCES

[1] Y. O'Connor, P. Andreev, and P. O'Reilly, "Mhealth and perceived quality of care delivery: a conceptual model and validation," *BMC Medical Informatics and Decision Making*, vol. 20, 02 2020.

[2] J. Shen, S. Ghatti, N. Levkov, H. Shen, T. Sen, K. Rheuban, K. Enfield, N. Facteau, G. Engel, and K. Dowdell, "A survey of covid-19 detection and prediction approaches using mobile devices, ai, and telemedicine," *Frontiers in Artificial Intelligence*, vol. 5, 12 2022.

[3] A. Mubaraki, A. Alrabie, A. Sibyani, R. Aljuaid, A. Bajaber, and M. Mubaraki, "Advantages and disadvantages of telemedicine during the covid-19 pandemic era among physicians in taif, saudi arabia," *Saudi Medical Journal*, vol. 42, pp. 110–115, 01 2021.

[4] S. Sood, V. Mbarika, S. Jugoo, R. Dookhy, C. Doarn, N. Prakash, and R. Merrell, "What is telemedicine? a collection of 104 peer-reviewed perspectives and theoretical underpinnings," *Telemedicine journal and e-health : the official journal of the American Telemedicine Association*, vol. 13, pp. 573–90, 11 2007.

[5] M. Alsaleh, V. Watzlaf, D. DeAlmeida, and A. Saptono, "Evaluation of a telehealth application (sehha) used during the covid-19 pandemic in saudi arabia: Provider experience and satisfaction," *Perspectives in health information management*, vol. 18, p. 1b, 10 2021.

[6] M. Hassibian and S. Hassibian, "Telemedicine acceptance and implementation in developing countries: Benefits, categories, and barriers," *Razavi International Journal of Medicine*, vol. Inpress, 08 2016.

[7] HealthIT.gov. (2019, Oct) What is telehealth? how is telehealth different from telemedicine? [Online]. Available: https://www.healthit.gov/faq/what-telehealth-how-telehealth-different-telemedicine

[8] F. Fatehi and R. Wootton, "Telemedicine, telehealth or e-health? a bibliometric analysis of the trends in the use of these terms," *Journal of telemedicine and telecare*, vol. 18, 12 2012.

[9] W. H. Organization. (2022, Oct) Ageing and health. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

[10] Y.-R. Hong, J. Lawrence, D. Williams, and A. M. Iii, "Population-level interest and telehealth capacity of us hospitals in response to covid-19: Cross-sectional analysis of google search and national hospital survey data," *JMIR Public Health and Surveillance*, vol. 6, 2020.

[11] H. Nadri, B. Rahimi, H. Lotfnezhad Afshar, M. Samadbeik, and A. Garavand, "Factors affecting acceptance of hospital information systems based on extended technology acceptance model: A case study in three paraclinical departments," *Appl Clin Inform*, vol. 09, pp. 238–247, 04 2018.

[12] M. Douglas, J. Xu, A. Heggs, G. Wrenn, D. Mack, and G. Rust, "Assessing telemedicine utilization by using medicaid claims data," *Psychiatric Services*, vol. 68, pp. 173–178, 02 2017.

[13] S. Zailani, M. Gilani, D. Nikbin, and M. Iranmanesh, "Determinants of telemedicine acceptance in selected public hospitals in malaysia: Clinical perspective," *Journal of medical systems*, vol. 38, p. 111, 09 2014.

[14] A. Alqudah, M. Al-Emran, and K. Shaalan, "Technology acceptance in healthcare: A systematic review," *Applied Sciences*, vol. 11, p. 10537, 11 2021.

[15] V. Venkatesh and F. Davis, "A theoretical extension of the technology acceptance model: Four longitudinal field studies," *Management Science*, vol. 46, pp. 186–204, 02 2000.

[16] V. Venkatesh and H. Bala, "Technology acceptance model 3 and a research agenda on interventions," *Decision Sciences - DECISION SCI*, vol. 39, pp. 273–315, 05 2008.

[17] V. Venkatesh, M. Morris, G. Davis, and F. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, vol. 27, pp. 425–478, 09 2003.

[18] R. EM, *Diffusion of Innovations (4th ed.)*. New York: The Free Press, 1995.

[19] I. Ajzen, "The theory of planned behaviour: Reactions and reflections," *Psychology and health*, vol. 26, pp. 1113–27, 09 2011.

[20] M. Fishbein, I. Ajzen, and A. Belief, "Belief, attitude, intention, and behavior: An introduction to theory and research," *Contemporary Sociology*, vol. 6, 03 1977.

[21] I. M. Rosenstock, "Historical origins of the health belief model," *Health Education Monographs*, vol. 2, no. 4, pp. 328–335, 1974. [Online]. Available: https://doi.org/10.1177/109019817400200403

[22] A. Alaboudi, A. Atkins, B. Sharp, A. Balkhair, M. Alzahrani, and T. Sunbul, "Barriers and challenges in adopting saudi telemedicine network: The perceptions of decision makers of healthcare facilities in saudi arabia," *Journal of Infection and Public Health*, vol. 9, 09 2016.

[23] A. Albahri, J. Alwan, Z. Taha, S. Fawzi, R. Amjed, A. Zaidan, O. Albahri, B. Bahaa, A. Alamoodi, and M. Alsalem, "Iot-based telemedicine for disease prevention and health promotion: State-of-the-art," *Journal of Network and Computer Applications*, 10 2020.

[24] D. Moher, A. Liberati, J. Tetzlaff, and D. Altman, "Research methods and reporting. preferred reporting items for systematic reviews and meta-analyses: the prisma statement," *Br Med J*, vol. 339, pp. 332–339, 07 2009.

[25] J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, and V. Welch, "Cochrane handbook for systematic reviews of interventions," *Cochrane Handbook for Systematic Reviews of Interventions*, 09 2019.

[26] M. Al-Emran, V. Mezhuyev, A. Kamaludin, and K. Shaalan, "The impact of knowledge management processes on information systems: A systematic review," *International Journal of Information Management*, vol. 43, pp. 173–187, 12 2018.

[27] M. Al-rawashdeh, P. Keikhosrokiani, B. Belaton, M. Alawida, and A. Zwiri, "Iot adoption and application for smart healthcare: A systematic review," *Sensors*, vol. 22, p. 5377, 07 2022.

[28] M. Yamin and B. Alyoubi, "Adoption of telemedicine applications among saudi citizens during covid-19 pandemic: An alternative health delivery system," *Journal of Infection and Public Health*, vol. 13, 10 2020.

[29] L. Liu and L. Shi, "Chinese patients' intention to use different types of internet hospitals: Cross-sectional study on virtual visits," *Journal of Medical Internet Research*, vol. 23, p. e25978, 08 2021.

[30] N. Dogra, S. Bakshi, and A. Gupta, "Exploring the switching intention of patients to e-health consultations platforms: blending inertia with push–pull–mooring framework," *Journal of Asia Business Studies*, vol. ahead-of-print, 01 2022.

[31] P. Chen, L. Xiao, Z. Gou, L. Xiang, X. Zhang, and P. Feng, "Telehealth attitudes and use among medical professionals, medical students and patients in china: A cross-sectional survey," *International Journal of Medical Informatics*, vol. 108, 09 2017.

[32] V. Brunelli, J. Fox, and D. Langbecker, "Disparity in cancer survivorship care: A cross-sectional study of telehealth use among cancer nurses in australia," *Collegian*, vol. 28, 01 2021.

[33] M. Finco, G. Cay, M. Lee, J. Garcia, E. Salazar, T.-W. Tan, D. Armstrong, and B. Najafi, "Taking a load off: User perceptions of smart offloading walkers for diabetic foot ulcers using the technology acceptance model," *Sensors*, vol. 23, p. 2768, 03 2023.

[34] H. Cho, G. Sanabria, M. Saylor, M. Gradilla, and R. Schnall, "Use of the fitt framework to understand patients' experiences using a real-time medication monitoring pill bottle linked to a mobile-based hiv self-management app: A qualitative study," *International Journal of Medical Informatics*, vol. 131, 08 2019.

[35] W. Ben Arfi, I. Nasr, T. Khvatova, and Y. Ben Zaied, "Understanding acceptance of ehealthcare by iot natives and iot immigrants: An integrated model of utaut, perceived risk, and financial cost," *Technological Forecasting and Social Change*, vol. 163, 11 2020.

[36] M. I. Hossain, A. Fadhil, A. Hussin, N. Iahad, and A. Sadiq, "Factors influencing adoption model of continuous glucose monitoring devices for internet of things healthcare," *Internet of Things*, vol. 15, p. 100353, 01 2021.

[37] V. Békés, K. Aafjes-Van Doorn, and B. Bőthe, "Assessing patients' attitudes towards telepsychotherapy: The development of the unified theory of acceptance and use of technology -patient version," *Clinical Psychology and Psychotherapy*, vol. 29, 06 2022.

[38] M. Haun, I. Stephan, M. Wensing, M. Hartmann, M. Hoffmann, and H.-C. Friederich, "Intent-to-adopt video-based integrated mental healthcare and characteristics of its supporters: A mixed-methods study among general practitioners applying diffusion of innovations theory," *JMIR Mental Health*, vol. 7, 09 2020.

[39] J. Monthuy-Blanc, S. Bouchard, C. Maïano, and M. Seguin, "Factors influencing mental health providers' intention to use telepsychotherapy in first nations communities," *Transcultural psychiatry*, vol. 50, 05 2013.

[40] C. Bakker, J. Huirne, F. Schaafsma, C. Geus, H. Bonjer, and J. Anema, "Electronic health program to empower patients in returning to normal activities after colorectal surgical procedures: Mixed-methods process evaluation alongside a randomized controlled trial (preprint)," 04 2018.

[41] T. Ownsworth, D. Theodoros, L. Cahill, A. Vaezipour, R. Quinn, M. Kendall, W. Moyle, and K. Lucas, "Perceived usability and acceptability of videoconferencing for delivering community-based rehabilitation to individuals with acquired brain injury: A qualitative investigation," *Journal of the International Neuropsychological Society*, vol. 26, pp. 47–57, 01 2020.

[42] R. Evering, M. Postel, H. van Os-Medendorp, M. Bults, and M. den Ouden, "Intention of healthcare providers to use video-communication in terminal care: a cross-sectional study," *BMC Palliative Care*, vol. 21, 11 2022.

[43] M. Nguyen, J. Fujioka, K. Wentlandt, N. Onabajo, I. Wong, R. Bhatia, O. Bhattacharyya, and V. Stamenova, "Using the technology acceptance model to explore health provider and administrator perceptions of the usefulness and ease of using technology in palliative care," *BMC palliative care*, vol. 19, p. 138, 09 2020.

[44] G. Pagaling, A. Espiritu, M. Dellosa, C. F. Leochico, and P. Pasco, "The practice of teleneurology in the philippines during the covid-19 pandemic," *Neurological Sciences*, vol. 43, 11 2021.

[45] C. Horsham, L. Loescher, D. Whiteman, P. Soyer, and M. Janda, "Consumer acceptance of patient-performed mobile teledermoscopy for the early detection of melanoma," *The British journal of dermatology*, vol. 175, 04 2016.

[46] A. Esber, M. Teufel, L. Jahre, J. Schmitten, E.-M. Skoda, and A. Bäuerle, "Predictors of patients' acceptance of video consultation in general practice during the coronavirus disease 2019 pandemic applying the unified theory of acceptance and use of technology model," *DIGITAL HEALTH*, vol. 9, p. 205520762211493, 01 2023.

[47] S. Marhefka, D. Turner, and E. Lockhart, "Understanding women's willingness to use e-health for hiv-related services: A novel application of the technology readiness and acceptance model to a highly stigmatized medical condition," *Telemedicine and e-Health*, vol. 25, 08 2018.

[48] L. Chen and A. Aklikokou, "Determinants of e-government adoption: Testing the mediating effects of perceived usefulness and perceived ease of use," *International Journal of Public Administration*, vol. 43, pp. 1–16, 09 2019.

[49] R. Wu, Z. Wu, J. Wen, Y. Cai, and Y. Li, "Extrinsic and intrinsic motivation as predictors of bicycle sharing usage intention: An empirical study for tianjin, china," *Journal of Cleaner Production*, vol. 225, 04 2019.

[50] S. Kaphle, S. Chaturvedi, I. Chaudhuri, R. Krishnan, and N. Lesh, "Adoption and usage of mhealth technology on quality and experience of care provided by frontline workers: Observations from rural india," *JMIR mHealth and uHealth*, vol. 3, p. e61, 05 2015.

[51] A. Carvalho and P. Santos, "Medication adherence in patients with arterial hypertension: The relationship with healthcare systems' organizational factors," *Patient Preference and Adherence*, vol. 13, p. 1761—1774, 10 2019.

[52] D. Bangert and R. Doktor, "Implementing store-and-forward telemedicine: Organizational issues," *Telemedicine journal and e-health : the official journal of the American Telemedicine Association*, vol. 6, pp. 355–60, 02 2000.

[53] J. Doran and J. Lawson, "The impact of covid-19 on provider perceptions of telemental health," *Psychiatric Quarterly*, vol. 92, 09 2021.

[54] V. Garg and J. Brewer, "Telemedicine security: A systematic review," *Journal of Diabetes Science and Technology*, vol. 5, no. 3, pp. 768–777, 2011, pMID: 21722592. [Online]. Available: https://doi.org/10.1177/193229681100500331

[55] J. Matusitz and G.-M. Breen, "Telemedicine: Its effects on health communication," *Health communication*, vol. 21, pp. 73–83, 02 2007.

[56] A. Alqudah, M. Al-Emran, and K. Shaalan, "Technology acceptance in healthcare: A systematic review," *Applied Sciences*, vol. 11, p. 10537, 11 2021.

[57] C. Barrette, "Usefulness of technology adoption research in introducing an online workbook," *System*, vol. 49, 04 2015.

[58] A. Bandura, "Self-efficacy: Toward a unifying theory of behavioral change," *Psychological Review*, vol. 84, pp. 191–215, 01 1977.

[59] L. Goldin, G. Cohen, C. Mueller, and A. Salles, "The relationship between self-efficacy and well-being among surgical residents," *Journal of Surgical Education*, vol. 76, 09 2018.

Appendices

Appendix 1: Database Search String

Search String (Scopus):

1.Population	patient* OR provider* OR doctor* OR physician OR nurse* OR HCP	11,296,563
	AND	
2.Intervention	Telemedicine OR "Remote Consultation" OR "virtual clinic" OR virtual-clinic OR "tele medicine" OR tele-medicine OR telehealth OR tele-health OR "mobile health" OR mHealth OR m-health OR eHealth OR e-health OR "remote consultation*" OR teleconsultation* OR tele-consultation* OR "tele consultation*" OR "video consultation*" OR "virtual consultation"* OR televisit* OR tele-visit* OR "tele visit*" OR eVisit* OR e-visit* OR "video visit*" OR "remote visit*" OR "virtual visit*" OR "video encounter*" OR "remote encounter*" OR "virtual encounter*" OR telediagnos* OR "tele-diagnos*" OR "tele diagnos*" OR "remote diagnos*" OR "virtual diagnos*" OR "video diagnos*" OR videoconferenc* OR "video conferenc*" OR "real-time video" OR "synchronous video"	7315
	AND	
3.Outcome	factor* OR influence OR determin* OR predict* OR "acceptance model" OR accept* OR satisfaction OR adopt* OR "adopt* model" OR "information system success model" OR theory OR theories OR framework* OR "behavioral intention" OR "intention to use"	31,302,427
Result	1 AND 2 AND 3	1804

Search String (PubMed):

1.Population	patient* OR provider* OR doctor* OR physician OR nurse* OR HCP	162,264
	AND	
2.Intervention	Telemedicine OR "Remote Consultation" OR "virtual clinic" OR virtual-clinic OR "tele medicine" OR tele-medicine OR telehealth OR tele-health OR "mobile health" OR mHealth OR m-health OR eHealth OR e-health OR "remote consultation*" OR teleconsultation* OR tele-consultation* OR "tele consultation*" OR "video consultation*" OR "virtual consultation"* OR televisit* OR tele-visit* OR "tele visit*" OR eVisit* OR e-visit* OR "video visit*" OR "remote visit*" OR "virtual visit*" OR "video encounter*" OR "remote encounter*" OR "virtual encounter*" OR telediagnos* OR "tele-diagnos*" OR "tele diagnos*" OR "remote diagnos*" OR "virtual diagnos*" OR "video diagnos*" OR videoconferenc* OR "video conferenc*" OR "real-time video" OR "synchronous video"	4,438
	AND	
3.Outcome	factor* OR influence OR determin* OR predict* OR "acceptance model" OR accept* OR satisfaction OR adopt* OR "adopt* model" OR "information system success model" OR theory OR theories OR framework* OR "behavioral intention" OR "intention to use"	139,233
Result	1 AND 2 AND 3	1949

Selected free full text, Books and Documents, Classical Article, Clinical Study, Controlled Clinical Trial, Corrected and Republished Article, Humans, Arabic, English.

Appendix 2: Quality Assessment Result

Quality assessment scores for included review studies.

Study ID	Q1	Q2	Q3	Q4	Q5	Score
1	1	1	1	1	1	5
2	1	1	1	1	1	5
3	1	0	1	1	1	4
4	1	1	1	1	1	5
5	1	1	1	1	1	5
6	0.5	1	1	1	1	4.5
7	1	0.5	1	1	1	4.5
8	1	1	1	1	1	5
9	1	1	1	1	1	5
10	1	1	1	1	1	5
11	1	1	1	1	1	5
12	1	1	1	1	1	5
13	0.5	1	1	1	1	4.5
14	1	0	1	1	1	4
15	1	1	1	1	1	5
16	1	1	1	1	1	5
17	0.5	1	1	1	1	4.5
18	1	1	1	1	1	5
19	1	1	1	1	1	5
20	0.5	1	1	1	1	4.5
21	0.5	1	1	1	1	4.5

Appendix 3: Complete form of the Extracted Data.

Study ID	Title	Author	year	Journal	country	Telemedicine application	Theory/model	Data Collection Design	Population		Telemedicine application	component of the theory/model	moderator components	statistical analysis
									NO. of sampling	Focus				
1	Adoption of telemedicine applications among Saudi citizens during COVID-19 pandemic: An alternative health delivery system	Yamin M, Alyoubi B	2020	Journal of Infection and Public Health	KSA	Telemedicine based wireless sensor network applications	UTAUT TTF	survey	348	patient	telemedicine covid-19	awareness self-efficacy UTAUT TTF awareness self-efficacy	non	structural equation modeling
2	Assessing patients' attitudes towards telepsychotherapy: The development of the unified theory of acceptance and use of technology-patient version	Békés V., Doorn K.A.-V., Bõthe B.	2022	Clinical Psychology and Psychotherapy	USA	telepsychotherapy	UTAUT	survey	107	patient	telepsychotherapy covid-19	Performance Expectancy Effort Expectancy Social Influence Facilitating Condition Anxiety Attitudes	non	Construct validity was assessed using exploratory factor analysis (EFA) in SPSS and confirmatory factor analysis (CFA) in Mplus 8.7. Reliability was assessed using two indicators (i.e., Cronbach's alpha and McDonald's) as using only Cronbach's alpha coefficients
3	Chinese patients' intention to use different types of internet hospitals: Cross-sectional study on virtual visits	Liu L., Shi L.	2021	Journal of Medical Internet Research	China	internet hospital	_	survey	1653	patient	telemedicine covid-19	_	non	logistic regression
4	Consumer acceptance of patient-performed mobile teledermoscopy for the early detection of melanoma	Horsham C., Loescher L.J., Whiteman D.C., Soyer H.P., Janda M.	2016	British Journal of Dermatology	Australia	teledermoscopy	TAM	survey	228	patient	teledermoscopy teledermoscopy for the early detection of melanoma	perceived usefulness ease of use trust attitude/intention subjective norms compatibility facilitators	non	not mentioned
5	Disparity in cancer survivorship care: A cross-sectional study of telehealth use among cancer nurses in Australia	Brunelli V.N., Fox J.A., Langbecker D.H.	2021	Collegian	Australia	telemedicine	TAM	survey	79	provider provider - nurse	telemonitoring cancer survivorship	perceived usefulness ease of use Attitude intention to use	non	logistic regression.
6	Electronic Health Program to Empower Patients in Returning to Normal Activities After Colorectal Surgical Procedures: Mixed-Methods Process Evaluation Alongside a Randomized Controlled Trial	den Bakker CM, Huirne JA, Schaafsma FG, de Geus C, Bonjer HJ, Anema JR.	2019	J Med Internet Res	Netherlands	using the "ikherstel" program	UTAUT	interview	14	patient	rehabilitation Empower Patients After Colorectal Surgery	Performance expectancy Effort expectancy Social influence Facilitating and inhibiting conditions	gender age experience voluntariness of use	_
7	Exploring the switching intention of patients to e-health consultations platforms: blending inertia with push–pull–mooring framework	Dogra N., Bakshi S., Gupta A.	2023	Journal of Asia Business Studies	India	telemedicine	push–pull–mooring (PPM) framework.	survey	413	patient	telemedicine	push effects (inconvenience and perceived risk) pull effects (opportunity for alternatives and ubiquitous care) mooring effects (trust) inertia (habit, switching cost)	non	structural equation modeling
8	Factors influencing mental health providers' intention to use telepsychotherapy in First Nations communities	Monthuy-Blanc J., Bouchard S., Maïano C., Séguin M.	2013	Transcultural Psychiatry	Canada	telepsychotherapy	TAM	survey	205	provider	telepsychotherapy mental health in First Nations communities	perceived usefulness ease of use	non	structural equation modeling
9	Intent to adopt video-based integrated mental health care and the characteristics of its supporters: Mixed methods study among general practitioners applying diffusion of innovations theory	Haun M.W., Stephan I., Wensing M., Hartmann M., Hoffmann M., Friederich H.-C.	2020	JMIR Mental Health	Germany	mental health specialist video consultations (MHSVC)	DOI Diffusion of Innovations Theory	focus group, interview	6	provider	telepsychotherapy mental health	perceived benefit availability	non	
10	Intention of healthcare providers to use video-communication in terminal care: a cross-sectional study	Evering RPostel Mvan Os-Medendorp H et al	2022	BMC Palliative Care	Netherlands	telemedicine	UTAUT DOI	survey	90	provider	app in palliative care terminal care	Outcome expectancy Effort expectancy Attitude Social influence Facilitating conditions Anxiety Self-efficacy Personal innovativeness	non	A multiple linear regression
11	Perceived Usability and Acceptability of Videoconferencing for Delivering Community-Based Rehabilitation to Individuals with Acquired Brain Injury: A Qualitative Investigation	Ownsworth T., Theodoros D., Cahill L., Vaezipour A., Quinn R., Kendall M., Moyle W., Lucas K.	2020	Journal of the International Neuropsychological Society	Australia	videoconferencing (VC)	TAM	interview	30	patient provider	rehabilitation for people with acquired brain injury (ABI)	the context or impetus for use perceived benefits potential problems and parameters around use balancing the service and user needs	non	Thematic analysis
12	Predictors of patients' acceptance of video consultation in general practice during the coronavirus disease 2019 pandemic applying the unified theory of acceptance and use of technology model	Esber A., Teufel M., Jahre L., in der Schmitten J., Skoda E.-M., Bäuerle A.	2023	Digital Health	Germany	telemedicine	UTAUT	survey	371	patient	teleconsultation covid-19	performance expectancy effort expectancy social influence computer proficiency	non	A hierarchical regression
13	Taking a Load Off: User Perceptions of Smart Offloading Walkers for Diabetic Foot Ulcers Using the Technology Acceptance Model	Finco MG, Cay G, Lee M, Garcia J, Salazar E, Tan TW, Armstrong DG, Najafi B.	2023	Sensors (Basel)	USA	Perceptions of Smart Offloading Walkers in adherence	TAM	survey	21	patient	telemonitoring Diabetic Foot Ulcers	perceived usefulness ease of use	non	Chi-squared tests
14	Telehealth attitudes and use among medical professionals, medical students, and patients in China: A cross-sectional survey	Chen P., Xiao L., Gou Z., Xiang L., Zhang X., Feng P.	2017	International Journal of Medical Informatics	China	telemedicine	not theory	survey	550	provider patient medical student	telemedicine general	advantage concern awareness usage	non	Structural equation modeling
15	The practice of teleneurology in the Philippines during the COVID-19 pandemic	Pagaling G.T., Espiritu A.I., Dellosa M.A.A., Leochico C.F.D., Pasco P.M.D.	2022	Neurological Sciences	Philippine	telemedicine	UTAUT	survey	147	provider	teleneurology	performance expectancy effort expectancy social influence facilitating conditions attitude behavioral intention	Experience Voluntariness	logistic regression analysis
16	Understanding Women's Willingness to Use e-Health for HIV-Related Services: A Novel Application of the Technology Readiness and Acceptance Model to a Highly Stigmatized Medical Condition	Marhefka S.L., Turner D., Lockhart E.	2019	Telemedicine and e-Health	USA	telemedicine	TRAM	mixed method	91	patient	Group-based telemedicine technology based group program of WLH	Innovativeness Optimism Discomfort Insecurity Perceived Usefulness Ease-of-Use group readiness HIV-related privacy concerns	_	logistic regression analysis
17	Use of the FITT framework to understand patients' experiences using a real-time medication monitoring pill bottle linked to a mobile-based HIV self-management app: A qualitative study	Cho H, Flynn G, Saylor M, Gradilla M, Schnall R.	2019	Int J Med Inform	USA	telemedicine- monitoring	FITT	interview	38	patient	telemonitoring improving medication adherence	fit between individuals and task: _motivation _self-efficacy _engagement fit between individuals and technology _preference for device design _HIV status _customized alert _ease of use fit between task and technology _system functionality _self-awareness	_	thematic analysis
18	Determinants of Telemedicine Acceptance in Selected Public Hospitals in Malaysia: Clinical Perspective	Zailani SGiiani MNikbin D et al.	2014	Journal of Medical Systems	Malaysia	telemedicine	extended TAM	survey	117	provider	telemedicine acceptance	government policy external suppliers' capacity project team's capacity top management support perceived usefulness perceived ease of use attitude self-efficiency	age job title job tenure	
19	Using the technology acceptance model to explore health provider and administrator perceptions of the usefulness and ease of using technology in palliative care	Nguyen MFujioka JWentlandt K et al	2020	Neurological Sciences	Canada	telemedicine	TAM	interview	18	provider	app in palliative care	perceived usefulness _ emote connection _ information-sharing platform ease of use _integration with existing IT systems _ user- friendly		
20	Understanding acceptance of eHealthcare by IoT natives and IoT immigrants: An integrated model of UTAUT, perceived risk, and financial cost	Ben Arfi WBen Nasr KKhatoova T et al	2021	Technological Forecasting and Social Change	France		UTAUT	survey	268	patient	telemonitoring IoT-based healthcare devices	performance expectancy effort expectancy social influence facilitating conditions perceived risk financial cost	age gender	Structural equation modeling
21	Factors influencing adoption model of continuous glucose monitoring devices for internet of things healthcare	Hossain MYusof AHussin A et al.	2021	Internet of Things (Netherlands)	not mentioned	telemedicine	TAM TRA TPB SE (Self-efficacy Theory)	survey	97	patient	telemonitoring Glucose Monitoring	Interpersonal Influence Personal Innovativeness Trustworthiness Attitude Toward a Wearable Device Self-Efficacy Health Interest Perceived Value	_	Partial least square and structural equation modelling

Appendix 4: List of the Factors Affecting Patients and Healthcare Providers in Different Telemedicine Applications.

Context of telemedicine	Focus		Factors					Frequency
			Individual	*Organizational and Environmental*	*Health*	*Technological*	*Security*	
Telerehabilitation	Provider (Ownsworth et al., 2020)	√	• Client capability and compatibility • Lack of physical presence	-	-	• Perceived benefit • Balancing the service and user needs. • The context for use • Technical and connectivity issues	-	
	Patient (den Bakker et al., 2019)	√	• Social Influence • Habit	• Facilitating Condition	-	• Performance Expectancy • Effort Expectancy	-	
Telemonitoring	Provider (Brunelli et al., 2021)	√	• Attitude	-	-	• Perceived usefulness • Ease to use	-	
	Patient (Hossain et al., 2021) (Finco et al., 2023) (Cho et al., 2019) (Ben Arfi et al., 2021)	√	• Motivation and engagement • Self-efficacy • Social Influence • Attitude • Personal innovativeness • Self-awareness • Age • Gender • Preference of device design	• Facilitating Condition • Financial cost	• Health condition statues (HIV) • Health interest • Perceived healthcare risk	• Perceived usefulness • Ease to use • Performance Expectancy • Effort Expectancy • System functionality • Perceived value • Customize alert	• Trust	ease to use (2) self-efficacy (2) Social Influence (2)
Telehealth applications in palliative care	Provider (Nguyen et al., 2020) (Evering et al., 2022)	√	• Social Influence • Attitude • Anxiety • Self-efficacy • Personal innovativeness	• Facilitating Condition	-	• Outcome expectancy • Effort Expectancy	-	
	patient	X	-	-	-	-	-	
Telepsychotherapy	Provider (Haun et al., 2020) (Monthuy-Blanc et al., 2013)	√	• Attitude	-	-	• Perceived usefulness • Ease to use • Availability of designed room. • Perceived benefit	-	
	Patient (Békés et al., 2022)	√	• Social Influence • Anxiety • Attitude	• Facilitating Condition	-	• Performance Expectancy • Effort Expectancy	-	
Teleneurology	Provider	√	• Social Influence	• Facilitating Condition	-	• Performance Expectancy	-	
	(Pagaling et al., 2022)		• Attitude • Experience • Voluntariness			• Effort Expectancy		
	patient	X	-	-	-	-	-	
Teledermoscopy	provider	X	-	-	-	-	-	
	Patient (Horsham et al., 2016)	√	• Subjective norms • Attitude • Compatibility	• Facilitating Condition		• Perceived usefulness • Ease to use	• Trust	
Teleconsultation	provider	X	-	-	-	-	-	
	Patient (Esber et al., 2023)	√	• Social Influence • Knowledge about digital health care solutions	-	• Depressive symptoms	• Performance Expectancy • Effort Expectancy • Computer proficiency	-	
Group-based telemedicine	provider	X	-	-	-	-	-	
	Patient (Marhefka et al., 2019)	√	• Optimism • Innovativeness • Discomfort • Group readiness	-	-	• Perceived usefulness • Ease to use	• Privacy • Insecurity	
Telemedicine	Provider (Chen et al., 2017) (Zailani et al., 2014)	√	• Attitude • Self-efficiency • Awareness • Previous experience	• Government policies • Top management support • Project team capacity • External suppliers' capacity • Health culture	-	• Perceived usefulness • Ease to use	• Authenticity and reliability of data from remote monitoring of patients	ease to use (2) perceived usefulness (2)
	Patient (Yamin & Alyoubi, 2020) (Liu & Shi, 2021) (Dogra et al., 2023)	√	• Social Influence • Awareness • Habit • Inconvenience • Age • Gender • Education level • City income level • Consumer type • Self-efficiency •	• Facilitating Condition • Cost	• Perceived healthcare risk • Hospital type	• Performance Expectancy • Effort Expectancy • Task technology fit • Ubiquitous care • Perceived benefit	• Trust in organization • Trust in telemedicine	

Attention-based Cross-Modality Multiscale Fusion for Multispectral Pedestrian Detection

Zhou Hui
School of Electrical and Information Engineering
TongJi University
Shanghai, China

***Abstract*—Multispectral pedestrian detection has wide applications in fields such as autonomous driving and intelligent surveillance. Mining complementary information between modalities is one of the most effective approaches to improve the performance of multispectral pedestrian detection. However, the inevitable introduction of redundant information between modalities during the fusion process leads to feature degradation. To address this challenge, we propose a multiscale differential fusion algorithm that leverages complementary information between modalities to suppress feature degradation caused by noise propagation along the network. We compare our algorithm with other cross-modal fusion pedestrian detection algorithms on the LLVIP and cleaned KAIST datasets. Experimental results demonstrate that our algorithm outperforms others, particularly in nighttime scenes where our algorithm achieves a 7.28% improvement in recall rate compared to the baseline on the cleaned KAIST dataset.**

***Keywords*—Pedestrian detection; multispectral pedestrian detection; attention mechanism; cross-modal fusion.**

I. Introduction

Pedestrian detection plays an important role in autonomous driving systems. In well-illuminated conditions, pedestrian detection achieves high precision. In poor lighting conditions, the appearance of pedestrians becomes blurred. Obstacles, overlapping figures, and varying distances contribute to these differences. As a result, nighttime pedestrian detection currently faces significant challenges [1].

Many advanced algorithms based on visible light images achieve notable performance improvements. Recent studies involving these images have validated their effectiveness, including in nighttime environments [2]. Due to the poor quality of nighttime visible light images, deep convolutional neural networks struggle to learn effective features. Image enhancement techniques show remarkable performance in enhancing the contrast between the foreground and background of an image. Some studies utilize enhanced image for feature extraction [3]. However, the majority of machine vision and deep learning models tend to perform poorly in highly challenging low-light scenarios [4]. Infrared images can highlight the thermal radiation characteristics of target objects, allowing for the capture of details such as human contours. Therefore, it possesses unique advantages in scenarios with insufficient lighting, adverse weather conditions, or concealed surveillance.

Despite the significant advantages of multimodal input data, effectively fusing information between modalities has become the core challenge and focus of algorithmic research. Li [5] et al. compared six fusion architectures which integrate color and thermal modalities at different position. Based on different fusion stages, it can be classified into early fusion, halfway fusion, and late fusion. Late fusion is currently the more commonly employed method, capable of mitigating the influence of modality and feature misalignment. However, it encounters challenges in network convergence and high computational complexity. We observed that discussions rarely address both the redundancy and complementarity of modalities. Crucially, the spread of redundant information can have detrimental effects in networks. This paper focuses on examining and mitigating these negative impacts by leveraging differential information of modalities in the backbone network to reduce redundancy.

The Non-Local neural network (Non-Local) [6] enhances inputs by calculating similarity in the channel direction. We conjecture that constructing an attention map by calculating similarity between pedestrian features could effectively allocate increased attention to those with blurred characteristics. In multispectral scenarios, there also exists a certain level of correlation both between channel dimensions and between spatial dimensions. Therefore, this paper proposes a dual-branch attention mechanism, named Dual Non-Local, which is based on both channel and spatial information. It establishes long-range dependencies between channels and spaces. Simultaneously, we utilize bright channel prior (BCP) algorithm to address low-light image compensation issues, and employ a multiscale feature fusion module to integrate visible and infrared modalities. Our work achieves superior results compared to some methods on the public available datasets KAIST and LLVIP.

We summarize the contributions of our work as follows:

1) A novel fusion approach is proposed for mining complementarity and reducing feature degradation. This technique involves the cross-fusion of complementary information from different modalities within the backbone network. The outputs of the backbone network for each modality is effectively integrated through this method.
2) A dual-branch attention mechanism based on channel and spatial attention. We embed positional information into attention map, and reduced the computational complexity of spatial attention. Ultimately, we build a dual-branch 3D attention mechanism that collaborates between spatial and channel dimensions.

II. Related Work

A. Pedestrian Detection

Pedestrian detection has high practical value in various applications, eg., autonomous driving and video surveillance. It receives extensive research attention in the field of computer vision. Pedestrian detection has undergone a significant transformation from handcrafted features to depending on deep convolutional networks for feature extraction [7]. Based on channel features or Deformable Part Models (DPM), there are two approaches to pedestrian detection that rely on handcrafted features. In 2009, P. Dollar et al. offered a fresh approach Integral Channel Features (ICF) [8], which utilized integral images for rapid feature computation. By combining channel feature pyramids with a cascaded classifier, they achieved faster detection results. ICF was the basis of channel features. Filtered Channel Features (FCF) [9] was optimization methods derived from ICF. Conventional algorithms were contingent upon manual design and frequently yielded diminished levels of detection accuracy. With Convolutional Neural Networks (CNNs) demonstrating outstanding feature extraction capabilities across various object detection tasks, pedestrian detection methods focused on leveraging deep learning techniques to enhance detection performance recently. The emergence of single-stage and two-stage algorithms, such as Faster R-CNN [10], [11], was a substantial potential for advancing accuracy and speed in pedestrian detection. Once in all weather conditions, especially during nighttime scenes, visible-light-based detection methods struggle to be effective. Simultaneously, infrared images complements visible light images, enabling the capture of pedestrian contours even in nighttime conditions. Detecting pedestrians in all weather conditions using multispectral images of color-thermal pairs has become a research hotspot.

B. Multispectral Pedestrian Detection

Effectively integrating infrared and visible light modalities is a challenging problem. In 2015, Hwang [12] et al. collected multispectral datasets, KAIST. The authors proposed the multispectral Aggregated Channel Features (ACF) method, incorporating intensity and gradient information from the thermal channel as additional channel information. An increasing number of multispectral pedestrian detection algorithms emerged based on this dataset. Liu [13] confirmed that multimodal pedestrian detection outperforms single-modal detection in terms of performance. Liu also investigated four fusion architectures: early fusion, mid-fusion, late fusion, and confidence fusion. They concluded that halfway fusion is the most effective fusion architecture. Inspired by Faster R-CNN, Konig [14] et al. proposed an effective multispectral RPN (Region Proposal Network)+BDT (Enhanced Decision Tree) model. In addition to investigating the fusion stages of multispectral images, another research approach involved using an illumination-aware network to weight the two modalities. Illumination-aware Faster R-CNN (IAF R-CNN) [5] introduced an illumination-weighting mechanism, forming a unified detection framework with separate subnetworks for visible light and infrared, along with a weighting layer. That means in low-light conditions, the network emphasized the features learned from the infrared sub network. In well-illuminated conditions, it focused on the visible light subnetwork. Our work is closely related to the conclusions drawn in [14]. We employed YOLOv7 [15] as our baseline and investigated the positive impact of low-light image enhancement techniques on the performance of multispectral pedestrian detection.

C. Attention Mechanism

In deep learning, the attention mechanism emulates the human visual and cognitive system, enabling neural networks to focus attention on relevant parts. Due to its outstanding performance, the attention mechanism is widely utilized in machine vision. Squeeze-and-Excitation Networks (SENet) [16] achieved adaptive channel-wise feature recalibration by modeling interdependencies between channels. Convolutional Block Attention Module (CBAM) [17] combined a channel attention module with a spatial attention module, allowing channel attention and spatial attention to operate sequentially. This enabled the network to simultaneously learn dependencies between channels and positional information. Non-Local neural network [6] combined self-attention with the general non-local mean method, establishing a long-range dependency model for transmitting long-range information. Non-Local maintained consistent feature scales between input and output, so it can be employed without modifying the network architecture. Criss-Cross Attention Network (CCNet) [18] and Global Context Network (GCNet) [19] were improvements derived from Non-Local. Similarity-based Attention Module (SimAM) [20] suggested that attention in the human brain often work in synergy, thus a unified attention mechanism was more in line with the working mechanism of neurons in the human brain. This paper introduces a new attention mechanism that combines the ideas of SimAM and Non-Local.

III. Proposed Method

The structure of this paper is depicted in Fig. 1. This model utilizes YOLOv7 as the baseline and integrates it with an illumination compensation module, a multiscale fusion module, and a detection module, forming a unified detection architecture. Our model consolidates the methods of image enhancement and differential fusion into a cohesive framework, thoroughly addressing the redundancy and complementarity across different modalities.

A. Illumination Compensation Network

The atmospheric scattering model is commonly employed to represent the degradation process of hazy images and is sometimes used for image enhancement tasks in low-light conditions as well [21]. Original image captured by the camera can be expressed as:

$$I = tJ + A\,(1 - t) \tag{1}$$

where, I is original image, J is restoration function of the image, A is environmental light description function, and t is medium propagation description function.

Wang [22] demonstrated that well-exposed images have at least some pixels with high illumination, unless these pixels are in shadow or covered by a black object. We visualize the Bright Channel on the KAIST dataset in Fig. 2. Most visible light images on the KAIST dataset are in underexposed scenes. We adopt unsupervised low-light Image enhancement network

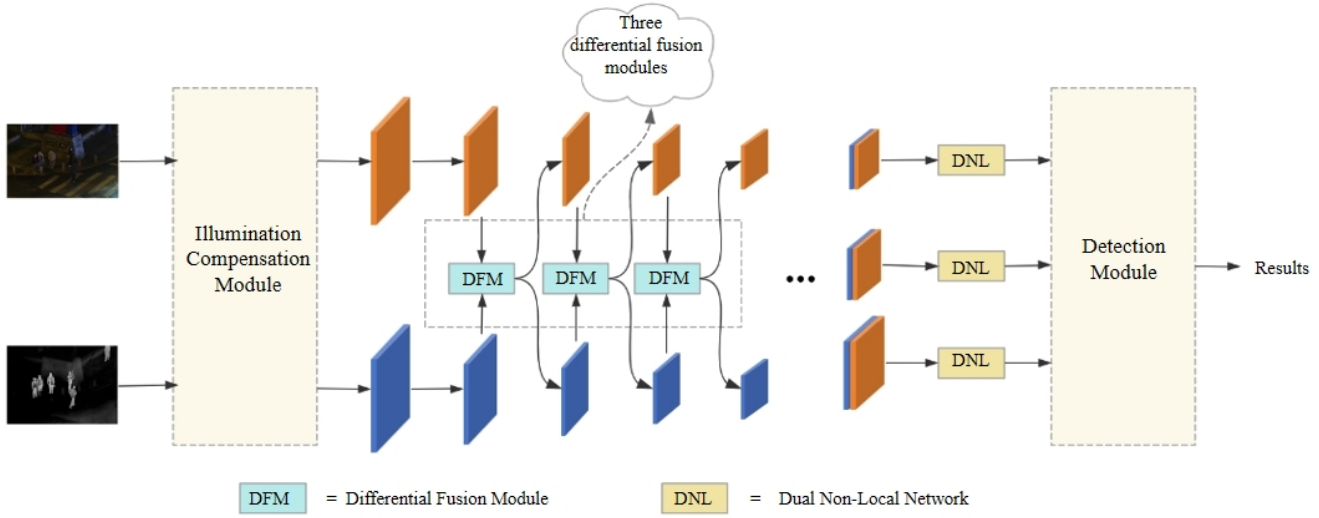


Fig. 1. The overall structure of proposed method. The network takes multispectral images of color-thermal pairs as inputs.

Fig. 2. The bright channel visualization of KAIST. We selected several low-light images and calculated the brightest pixels in R, G, B channels for each image, denoted as bright channel.

for outputing adjusted image that is guided by a unsupervised loss L_{BCP}. The parameters in Eq. (1) are reinterpreted as Eq. (2):

$$I_p = t_p J_p + A\,(1 - t_{\mathrm{p}}\,) \tag{2}$$

where, A and t_p is environment light and the illumination map, respectively. J_p represents enhanced output images and I_p is observed images. According to BCP [22], We adopted Eq. (3) as the brightest intensity:

$$J_p^{bright} = \max_{c\in r,g,b}\left(\max_{q\in\Phi(p)} J_q^c\right) \tag{3}$$

where, J_p^{bright} represents the brightest intesity in r,g,b channels, q is the pixels centered at p, and c is different channels in RGB images. Additionally, the brightest intesity becomes $J_p^{bright} \to 1$. Assuming that A is known, $\widetilde{t}$ represents the illumination map, which is considered as a constant within a patch.

By taking the maximum operator between the left and right sides of Eq. (3), we obtain an initial illumination map formulation as shown in Eq. (4):

$$\tilde{t}_p = 1 - \max_{c,q}\left(\frac{1 - I_q^c}{1 - A^c}\right). \tag{4}$$

where $\tilde{t}_p$ is the illumination value at pixel p, I_q^c is observed image, q is pixel centered at p, and c is different channels in RGB images. Under the supervision of the initial illumination map, we obtain enhanced illumination map t_p through Illumination Compensation Network. Substituting t_p into Eq. (2), we get enhanced output images as Eq. (5):

$$J_p = \frac{I_p - A}{t_p} + A \tag{5}$$

The darkest pixels in the bright channel of the image can be considered as environment light. To adjust dark spots and black objects in real life, we take the average value of the darkest 0.1% pixels (denoted as K) in the bright channel of the image as the environment light, as shown in Eq. (6):

$$A = \frac{1}{|K|}\sum_{p\in K} I_p \tag{6}$$

To address oversaturation, this paper similarly utilizes the output from Eq. (7) as the attention map:

$$T_{attention} = T^{\gamma} \tag{7}$$

where T is thermal image, and $\gamma(\gamma > 1)$ controls the curvature of the attention map.

The enhancement network alters the feature scale and utilizes the attention map to optimize spatial weights. In summary, our final illumination compensation network is illustrated in Fig. 3. The visible light features are compensated by the infrared images, effectively enhancing the distinguishability of the RGB images.

B. Multiscale Fusion Module

In all weather conditions, visible light images provide more information about pedestrians in well-illuminated conditions while in low-light conditions, thermal images provide more information. Most multispectral approaches extract features from

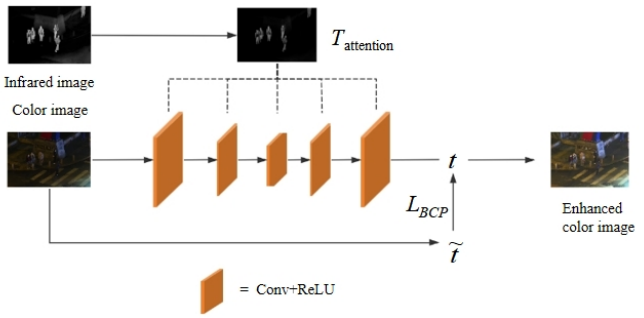


Fig. 3. The structure of Illumination Compensation module. $T_{attention}$ fed to five convolutional layers to adjust visible light feature. $\tilde{t}$ is an initial illumination map.

two streams and directly combine them either by element-wise addition or by channel concatenation. However, these mehods overlook the complementarity between the two modalities. The propagation of redundant information between modalities through the network also have adverse effects.

Inspired by the differential modality information [23], we propose a fusion module: Differential Fusion Module(DFM), to enhance the mutual suppression and enhancement, as shown in Fig. 4. The features obtained by element-wise subtraction of the two modalities reflect their complementary information, ingeniously excluding redundant information from feature fusion. This element-wise subtraction also prevents interference from features learned from another modality in the previous fusion from affecting the next fusion. Integrated within the architecture of YOLOv7, we perform multiscale feature fusion at the position illustrated in Fig. 1. In multispectral pedestrian

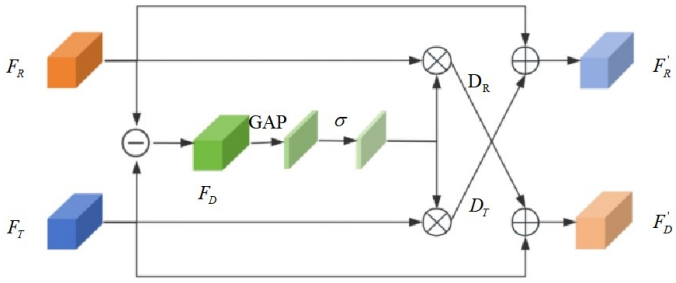


Fig. 4. The structure of differential fusion module. F_R and F_T are visible light features and thermal features. We obtain F_D by subtracting element-wise.

detection task, it is crucial to effectively integrate valuable information between modalities and mitigate interference caused by redundant information.

We applied DFM at Conv3, Conv4, Conv5 layers of the Backbone. The outputs feed into multiscale feature fusion network. To further enhance crucial features, we employ the Dual Non-Local attention mechanism before pyramid feature network. This helps to improve pedestrian feature expression effectively. DFM involves using a differencing mechanism, where F_R and F_T are subtracted element-wise to obtain the feature F_D. The equation for F_D is as follows:

$$F_{D1} = F_R - F_T \tag{8}$$

$$F_{D2} = F_T - F_R \tag{9}$$

where, F_R and F_T represent the extracted features from the visible light image and infrared image, respectively. F_D is the difference between F_R and F_T.

Subsequently, F_D is obtained through a global average pooling layer(GAP) and a tanh activation layer in order to get the attention map in the channel direction. That attention map is multiplied with the input visible light features and infrared features separately, producing D_R and D_T. The cross addition is applied to F_T and D_R, as well as F_R and D_T. Differential features yields the output features.

GAP computes the mean of the two-dimensional images within each channel, obtaining an attention map in the channel direction that contains global information. D_R is present in the visible light features but absent in the infrared image features while D_T is present in the infrared image features but absent in the visible light features. The formulation of differential feature is as follows:

$$F_R^{'} = F_R + D_T \tag{10}$$

$$F_T^{'} = F_T + D_R \tag{11}$$

where $F_R^{'}$ is the output of F_R, and $F_T^{'}$ is the output of F_T.

After cross-complementary feature fusion at three scales in the Backbone, the deep semantic information is concatenated. The deep semantic information needs to be fed into the Dual Non-Local module to enhance crucial information before the feature pyramid network. There is a high degree of correlation between pedestrian features, so establishing long-range dependencies is beneficial for modeling the similarity relationships between pedestrian features.

C. Dual Non-Local Attention Mechanism

Capturing long-range dependencies is crucial in pedestrian detection task. Long-range dependencies in image can only be formed through the successive convolutional layers in deep neural networks, named large receptive field. Inspired by SENet and Non-Local, we proposed a 3D attention mechanism called Dual Non-Local, as illustrated in the Fig. 5. Does long-range dependency work effectively in multi-pedestrian scenes? Undoubtedly, there is some correlation among the extracted features of pedestrians. The feature similarity matrices between each pixel and the feature similarity matrices between each channel assist in providing blurred pedestrian features with the weights of clear pedestrian features. Spatial attention and channel attention were applied concurrently for enhancing pedestrian feature. Dual Non-Local Network consists of a Spatial Attention Module (SAM) and a Channel Attention Module (CAM), which share the same input, denoted as $x \in \mathbb{R}^{C\times H\times W}$. There is a similar attention map computed for each position in Non-Local Network [19], thus, we use global attention maps to reduce computational cost.

In spatial attention module, we reshape the input into m, $m \in \mathbb{R}^{1\times C\times (H*W)}$. A 1×1 convolutional operation is imposed to x for getting global information of channels, denoted as n,

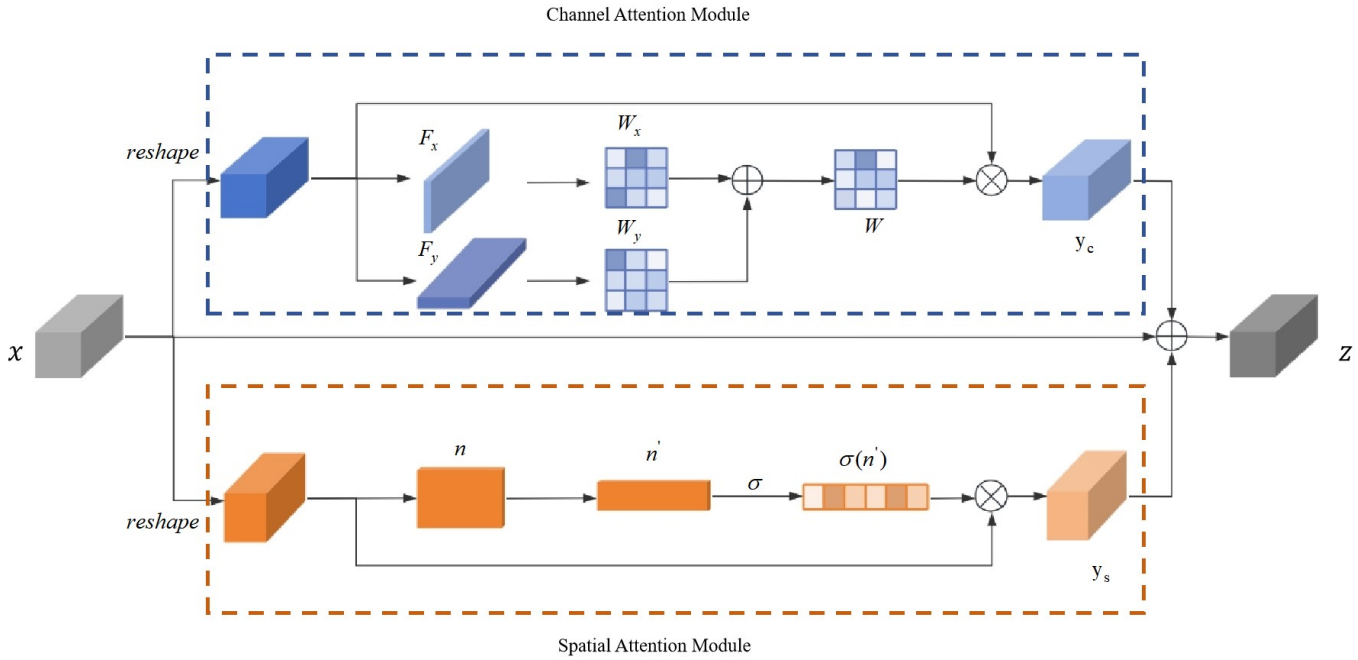


Fig. 5. The structure of dual non-Local network consists two branch attention mechanisms. SAM uses a shared attention map $\delta(n^{'})$ globally for reducing computation, while CAM calculates attention maps in different dimensions.

$n \in \mathbb{R}^{1\times H\times W}$. Before multipling with m, $n^{'}$ passes through a softmax function. Global channel information was reshaped to $n^{'} \in \mathbb{R}^{1\times(H*W)\times 1}$. In order to get the similarity between $H*W$ pixels, we multiply m and $n^{'}$. y_s is the output of spatial attention module, $y_s = m \bigotimes \delta(n^{'})$, where δ is softmax function. y_s is formulated as follows:

$$\mathbf{y}_{\mathrm{s}} = \sum_{j=1}^{N} \mathbf{x}_j \delta\left(\mathbf{n}'\right) \tag{12}$$

where, N is the number of pixels and y_s is shared globally as a spatial attention map.

In channel attention module, this paper innovatively performs pooling operations separately in the x and y dimensions, injecting positional information into channel attention map. After pooling in the x dimension, features are denoted as F_x, $F_x \in \mathbb{R}^{C\times H\times 1}$, and in the y dimension are denoted as F_y, $F_y \in \mathbb{R}^{C\times 1\times W}$, correspondingly. F_x performs a 1×1 filter, followed by a reshape function, to obtain θx, φx, where $\theta x \in \mathbb{R}^{C\times 1\times H}$ and $\varphi x \in \mathbb{R}^{C\times H\times 1}$. Dot product is carried out between θx and φx to obtain the weights between channels refered to W_x, $W_x \in \mathbb{R}^{C\times C}$ in dimension y. In a similar way, W_y, $W_y \in \mathbb{R}^{C\times C}$, represents the weights between channels in dimension x.

We suggest x and y dimensions play the same important role in channel attention, so the total channel weight distribution is considered to be the sum of the weight distribution in both dimension x and dimension y, denoted as W, $W = W_x + W_y$. Followed by a 1×1 filter and reshape operation, x generates g_x, $g_x \in \mathbb{R}^{C\times(H*W)}$, in order to obtain the final channel attention output y_c by applying the learned channel attention weights (W). To recover the features to their original input dimensions, we use a 1×1 filter to generate weights W_z. The final channel attention output is formulated as follows:

$$y_c = W_z(W \bigotimes g_x) \tag{13}$$

Finally, there is a shortcut between input and output as a residual structure. The network retains the input x, only learns the difference between output and input, and the data flows across layers to avoid the gradient disappearing during the training process. We denoted the final response of x as z, and we summarized the formulation between every pixel as follows:

$$\begin{aligned} z_i =& W_z \sum_{j=1}^{N_C} \left(\frac{f\left(C_{W_i}, C_{Wj}\right) + f\left(C_{Hi}, C_{Hj}\right)}{N_C} \right) \left(g_x \cdot x_j\right) \\ &+ m \bigotimes \delta(n^{'}) + x_i \end{aligned} \tag{14}$$

where i represents a position in image, and j is all possible positions. N_C is number of channels, $f\left(C_{W_i}, C_{Wj}\right)$ and $f\left(C_{Hi}, C_{Hj}\right)$ are the similarity between channels calculated by dot product.

The total loss is defined as (16):

$$\begin{aligned} L_{BCP} =& \frac{1}{N} \sum_{p} \{ \left(t_p - \tilde{t}_p\right)^2 \\ &+ \lambda \sum_{i,j\in\Phi(p)} w_{ij} \left(t_i - t_j\right)^2 \} \end{aligned} \tag{15}$$

$$L = L_{BCP} + \alpha L_{YOLOv7} \tag{16}$$

where N is the number of pixels, $\Phi(p)$ is the pixels within a 3×3 patch centered at p, w_{ij} represents affinity matrix between $\Phi(p)$, and λ controls balance between the data term and the smoothing term. Integrating detection and enhancement

loss during training is beneficial for obtaining image with prominent pedestrian features for pedestrian detection.

IV. Experiments

In order to demonstrate the effectiveness of the model we proposed, we present the detection results on two datasets, cleaned KAIST and LLVIP.

A. Dataset

1) KAIST: The KAIST dataset, proposed by Hwang [12] et al., consists of multispectral pedestrian data captured by specialized hardware with a beam splitter. It comprises 95,328 pairs of color and thermal images. However, this dataset is derived from consecutive frames of a video causing a high similarity in adjacent images, so we perform data clean. Finally, we get 7601 pairs of images as training set, and 2252 pairs of images as testing set. Additionally, we adopted the re-annotated labels by Li [24] and Hangil [25] for the training and test sets, respectively, to enhance label quality.

2) LLVIP: The LLVIP [26] dataset consists of rigorously aligned pairs of images in both time and space, which is used for pedestrian detection in low-light conditions. The entire dataset comprises 15,488 pairs of color-thermal images.

B. Evaluation Metrics

We use the Recall and Average Precision (AP) as evaluation metrics to evaluate the proposed model effectively. Here, we use TP (True Positive), FP (False Positive) to represent true positive predictions and false positive predictions, respectively. Recall is the ratio of detected pedestrians in ground truth. $Recall = \frac{TP}{TP+FP}$.

C. Implementation Details

In this paper, we built our network based on YOLOv7 and added a illumination compensation network at the input, which enhances the visible light by using the bright channel prior. In the Backbone, differential fusion module was performed on the feature inputs of Conv3, Conv4, and Conv5 to reduce redundancy in modal fusion. Finally, an innovative attention mechanism was added for long-term dependencies, facilitating direct transmission of high-level semantics.

The experiments were conducted on an NVIDIA GeForce RTX 4080 GPU, Intel(R) Core(TM) i7-13700F CPU, using the PyTorch framework and public code YOLOv7. We set the batch size to 8, epoch to 100, and resize input images to 640×640. K-means clustering provided nine anchor boxes for the KAIST dataset: [44,65], [26,111], [33,141], [41,117], [43,153], [58,116], [52,146], [59,178], [71,152]. We used some training tricks such as mosaic augmentation and random cropping to enhance the network's generalization.

D. Results Analysis

We conduct a comparison between our algorithm, Halfway Fusion and IAF R-CNN on the cleaned KAIST dataset. Here, we primarily discuss the potential advantages of our method, such as how our framework utilizes a Halfway Fusion architecture for integration, and we identify key methodologies that are beneficial in enhancing detection performance. The pedestrian detection results are presented in Table I. For the cleaned KAIST dataset, proposed method achieves the best detection performance in terms of Recall 64.17%.

TABLE I. Comparison on Cleaned KAIST Dataset in Terms of Recall

Method	DAY	NIGHT	ALL
Halfway Fusion [13]	59.98	50.77	58.27
IAF R-CNN [5]	65.22	56.62	62.14
YOLOv7 [15]	66.11	49.89	60.98
Ours	68.23	57.17	64.17

Compared to IAF R-CNN, our method is equally competitive, maintaining a high recall rate while our inference time is only 0.096s/image, as opposed to 0.210s/image for IAF R-CNN. We record comparison of inference time using an NVIDIA GeForce RTX 4080 GPU in Table II. This advantage is attributed to the real-time nature of the single-stage object detection algorithm, but the improvement in recall rate is due to our differential fusion module effectively mining the complementary features of pedestrian characteristics, reducing redundant noise interference in feature propagation. However, the effectiveness of DFM is limited by the requirement that the input pairs of visible and infrared images must be strictly aligned. Misaligned image pairs transmit incorrect differential information, and the noise is amplified by the network. This limitation calls for the use of more sophisticated image acquisition instruments to be adequately addressed.

TABLE II. Comparision of Inference Time using an NVIDIA GeForce RTX 4080 GPU

Method	IAF R-CNN [5]	YOLOv7 [15]	Ours
Time(s.)	0.210	0.041	0.096

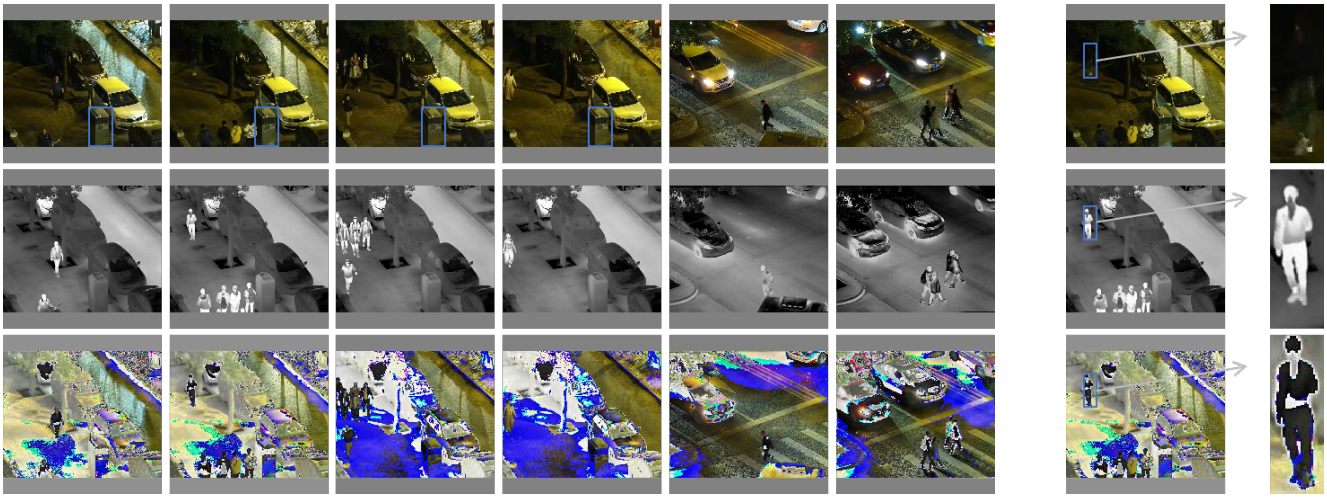
In Table III, IV and V, we compare our algorithm with YOLOv7. Moreover, we explored three versions input of YOLOv7: a) RGB branch; b) thermal branch; c) concat thermal image and visible light image as input. Directly concatenat-

TABLE III. Comparison on Cleaned KAIST Dataset for Nighttime Scenes in Terms of Average Precision, Recall, and Accuracy.

Method	AP/%	Recall/%	Accuracy/%
YOLOv7 RGB	39.73	35.54	88.03
YOLOv7 T	49.24	18.74	92.65
YOLOv7 T+RGB	55.58	49.89	80.35
Ours	64.20	57.17	86.50

TABLE IV. Comparison on Cleaned KAIST Dataset for Daytime Scenes in Terms of Average Precision, Recall, and Accuracy

Method	AP/%	Recall/%	Accuracy/%
YOLOv7 RGB	60.97	60.26	80.05
YOLOv7 T	44.57	14.68	89.89
YOLOv7 T+RGB	66.05	66.11	72.73
Ours	63.83	68.23	81.47



(a) some examples (b) detail comparsion

Fig. 6. The visualizations of enhanced features from Illumination compensation network a) some examples; b) detail comparison.

TABLE V. COMPARISON ON CLEANED KAIST DATASET FOR ALL WEATHER SCENES IN TERMS OF AVERAGE PRECISION, RECALL, AND ACCURACY

Method	AP/%	Recall/%	Accuracy/%
YOLOv7 RGB	54.30	52.44	81.63
YOLOv7 T	46.01	15.96	90.90
YOLOv7 T+RGB	62.60	60.98	76.01
Ours	63.94	64.17	82.96

ing the visible light image and thermal image did not lead to a significant improvement. We achieve improvements of 3.19%, 2.12%, and 7.28% on the all weather, daytime, and nighttime test sets in terms of Recall, respectively. Our method demonstrate better performance in both accuracy and Recall, indicating the effectiveness of our fusion strategy. Illumination compensation network enhances pedestrian features in low-light scenarios. Thus, we obtained the highest performance in nighttime scenes.

There are some visualizations for enhancced images as the outputs of Illumination Compensation Network in Fig. 6. We observed that obstacles in blue boxes have a high similarity to pedestrians, especially in low-light scenarios. The illumination compensation network is advantageous in suppressing background features and enhancing foreground characteristics. That enables the network to concentrate more on pedestrian targets, free from background interference. In the third line of Fig. 6b, pedestrian feature is clearer in blue box. However, the multispectral images of color-thermal pairs must be aligned. When misalignments occur, our model leads to worse results, which requires more sophisticated image acquisition instruments.

In addition to the quantitative analysis,we also provide several qualitative results on the cleaned KAIST dataset in Fig. 7. Upon observation, it is evident that our method excels in generating precise bounding boxes and accurately detecting pedestrians, especially in challenging scenarios when compared to the baseline model.

Gradient-weighted Class Activation Mapping (Grad-CAM) is a method for visualizing the attention mechanisms of deep neural networks. Our Dual Non-Local module constructs a unified attention framework based on the similarity of channel and spatial features, making it particularly suitable for single-object detection scenarios. The feature similarity between different types of targets may cause confusion in the attention map. In single object detection tasks, our Dual Non-Local module demonstrates superior performance compared to Non-Local and SimAM. These three similar attention mechanisms have consistent input and output dimensions. We removed all other modules in Fig. 1, retaining only the attention module. We replaced this position with different attention mechanisms and trained using the RGB images from LLVIP.

Using Grad-CAM, we visualized the outputs of Dual Non-Local, Non-Local, and SimAM, as shown in Fig. 8. Our Dual Non-Local model focuses more attention on the entirety of pedestrians, while Non-Local and SimAM distribute attention more precisely, but they both have the issue of some pedestrian regions not receiving attention. Comparatively, although the attention regions of Dual Non-Local are less precise, all pedestrian regions receive attention intensely. This also validates that features of clear pedestrians can rectify those pedestrians with blurred features.

E. Ablation Experiment

Our model achieves a leading performance. Nevertheless, the specific contributions of each module to the results remained uncertain. To address this, we design some ablation experiments to verify it. The comparsion results are presented in Table VI.

1) Illumination Compensation Module(IC): To figure out the impact of the illumination compensation network, we design a new network that uses the concatenated outputs of the illumination compensation network as the input of backbone, by removing the multiscale fusion network and

(a) visible light images (b) baseline detection results (c) our detection results (d) ground truth

Fig. 7. The visualizations of baseline and our algorithm. It contains a) visible light images; b) baseline detection results; c) our detection results; d) ground truth. According to the results, our method generate more target boxes correctly. Our method performs better when visible light images are in low-light condition.

TABLE VI. ABLATION RESULTS ON THE CLEANED KAIST DATASET IN TERMS OF PRECISION

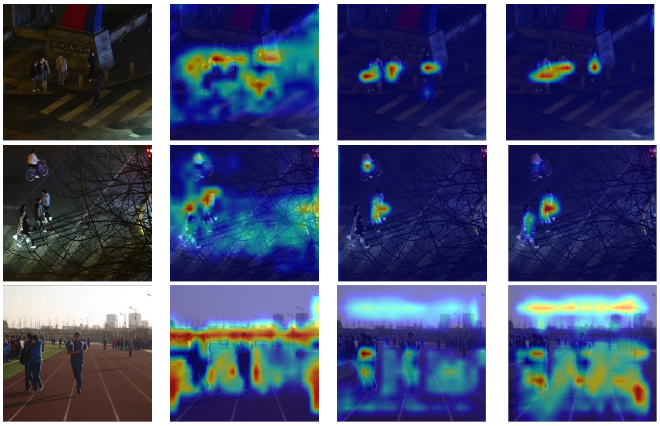
Method	DAY/%	NIGHT/%	ALL/%
YOLOv7_4c	72.73	80.35	76.01
YOLOv7+IC	73.98	84.09	79.90
YOLOv7+MFM	74.20	84.61	78.45
YOLOv7+DNL	73.10	80.53	76.56
YOLOv7+IC+MFM	79.87	86.11	81.39
YOLOv7+IC+DNL	74.41	84.05	80.04
YOLOv7+DNL+MFM	74.57	85.26	79.48
YOLOv7+DNL+MFM+IC(Ours)	81.47	86.50	82.96

attention mechanism. Thanks to the (15), the proposed L_{BCP} loss also has contributions to performance improvement, by distinguishing foreground from background as effectively as possible.

2) Multiscale Fusion Module (MFM): As shown, the one-branch methods are undoubtedly inferior to the two-branch approach. However, the crucial factor is fusion stage while halfway fusion structure achieves the best performance. For the two-branch method, we created two separate backbones to process visible and infrared images. It's worth noting that, during this experiment, we omitted the illumination compensation network. Both modalities were fed directly into their respective backbones. According to the results, it is evident that DFM plays a crucial role in improving detection performance, which resonates with our initial conjecture. Although DFM requires strictly aligned image pairs as input, this outcome provides strong experimental support for future research on pedestrian detection in more challenging environments.

F. Other Dataset

To demonstrate the generalization capability of our algorithm, we conducted experiments not only on the cleaned KAIST dataset, but also on another multispectral pedestrian detection benchmark called LLVIP. The majority of the LLVIP dataset were captured in low-light nighttime conditions. We



(a) original image (b) Dual Non-Local (c) Non-Local (d) SimAM

Fig. 8. The Grad-Cam visualizations between non-local, SimAM, and our dual non-local.

recorded the performance of the LLVIP dataset in Table VII, with mAP as the evaluation metric.

TABLE VII. COMPARISON ON LLVIP DATASET FOR ALL WEATHER SCENES IN TERMS OF AVERAGE PRECISION, RECALL, AND ACCURACY

Method	AP/%	Recall/%	Accuracy/%
YOLOv4 T+RGB	50.90	57.10	74.00
YOLOv7 T+RGB	79.60	71.89	94.34
Ours	83.76	78.16	97.81

V. CONCLUSION AND FUTURE WORK

In this paper, we investigated to integrate color-thermal image pairs effectively, leveraging the complementarity and exclusivity between modalities to enhance detection performance. We proposed an algorithm based on multiscale feature fusion. Specifically, we performed image enhancement on the input visible light image and simultaneously improved the Backbone network through integrating two modalities using differential information in Conv3, Conv4, and Conv5 convolutional layers. Our approach demonstrated outstanding performance on the cleaned KAIST and LLVIP datasets. Particularly in nighttime scenarios, we achieved a improvement of 7.28% in terms of Recall compared to the baseline on the cleaned KAIST dataset. We suggested that the proposed Dual Non-Local attention mechanism is also effective for other single object detection tasks, which is part of our future work. The findings of this paper offer a novel approach to combine image enhancement techniques and feature fusion for multispectral pedestrian detection, with potential applications beyond pedestrian detection. In our future work, we aim to further explore the complementarity between modalities and reduce redundant information between modalities in more challenging weather conditions, such as rain and snow scenarios.

REFERENCES

[1] F. Xu and K. Fujimura, "Pedestrian detection and tracking with night vision," in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 1, 2002, pp. 21–30 vol.1.

[2] A. Ćorović, V. Ilić, S. Đurić, M. Marijan, and B. Pavković, "The real-time detection of traffic participants using yolo algorithm," in *2018 26th Telecommunications Forum (TELFOR)*, 2018, pp. 1–4.

[3] W. Wang, Y. Peng, G. Cao, X. Guo, and N. Kwok, "Low-illumination image enhancement for night-time uav pedestrian detection," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5208–5217, 2021.

[4] S. S. S. Kruthiventi, P. Sahay, and R. Biswal, "Low-light pedestrian detection from rgb images using multi-modal knowledge distillation," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 4207–4211.

[5] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster r-cnn for robust multispectral pedestrian detection," *Pattern Recognition*, vol. 85, pp. 161–171, 2019.

[6] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[7] T. Liu, K.-M. Lam, R. Zhao, and G. Qiu, "Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 315–329, 2021.

[8] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," 2009.

[9] S. Zhang, R. Benenson, B. Schiele *et al.*, "Filtered channel features for pedestrian detection." in *CVPR*, vol. 1, no. 2, 2015, p. 4.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[11] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *IEEE transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2017.

[12] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.

[13] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *arXiv preprint arXiv:1611.02644*, 2016.

[14] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 49–56.

[15] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464–7475.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[17] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[18] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.

[19] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[20] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "Simam: A simple, parameter-free attention module for convolutional neural networks," in *International conference on machine learning*. PMLR, 2021, pp. 11 863–11 874.

[21] H. Lee, K. Sohn, and D. Min, "Unsupervised low-light image enhancement using bright channel prior," *IEEE Signal Processing Letters*, vol. 27, pp. 251–255, 2020.

[22] Y. Wang, S. Zhuo, D. Tao, J. Bu, and N. Li, "Automatic local exposure correction using bright channel prior for under-exposed images," *Signal processing*, vol. 93, no. 11, pp. 3227–3238, 2013.

[23] K. Zhou, L. Chen, and X. Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 787–803.

[24] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," *arXiv preprint arXiv:1808.04818*, 2018.

[25] H. Choi, S. Kim, K. Park, and K. Sohn, "Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks," in *2016 23rd International conference on pattern recognition (ICPR)*. IEEE, 2016, pp. 621–626.

[26] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "Llvip: A visible-infrared paired dataset for low-light vision," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3496–3504.

Deep Learning-Powered Mobile App for Fast and Accurate COVID-19 Detection from Chest X-rays

Rahhal Errattahi, Fatima Zahra Salmam, Mohamed Lachgar, Asmaa El Hannani, Abdelhak Aqqal
Chouaib Doukkali University of El Jadida, National School of Applied Sciences, Laboratory of Information Technologies, El Jadida, Morocco

***Abstract*—The COVID-19 pandemic has imposed significant challenges on healthcare systems globally, necessitating swift and precise screening methods to curb transmission. Traditional screening approaches are time-consuming and prone to errors, prompting the development of an innovative solution - a mobile application employing machine learning for automated COVID-19 screening. This application harnesses computer vision and deep learning algorithms to analyze X-ray images, rapidly detecting virus-related symptoms. This solution aims to enhance the accuracy and speed of COVID-19 screening, particularly in resource-constrained or densely populated settings. The paper details the use of convolutional neural networks (CNNs) and transfer learning in diagnosing COVID-19 from chest X-rays, highlighting their efficacy in image classification. The trained model is deployed in a mobile application for real-world testing, aiming to aid healthcare professionals in the battle against the pandemic. The paper provides a comprehensive overview of the background, methodology, results, and the application's architecture and functionalities, concluding with avenues for future research.**

Keywords*—*COVID-19 diagnosis; computer vision; deep learning; X-ray images; mobile application.

I. Introduction

The COVID-19 pandemic has brought unprecedented challenges to healthcare systems worldwide. Early and efficient screening is crucial to prevent the spread of the virus. However, traditional screening methods such as manual assessments and laboratory tests are time-consuming and often have a high error rate [1]. Therefore, there is an urgent need for a more effective and rapid screening solution. To address this challenge, we propose an innovative solution - a mobile application that leverages machine learning for automatic screening for COVID-19. The application uses computer vision and deep learning algorithms to analyze X-ray images of patients and detect symptoms of the virus in real-time. The proposed application has the potential to improve the accuracy of screening results and reduce the time required for patient diagnosis by automating the screening process.

The proposed solution aims to overcome the limitations of traditional COVID-19 screening methods, especially in settings with limited resources or high population density. By leveraging the capabilities of machine learning, the screening process is broken down into smaller components that can be easily updated and maintained. The ultimate objective of this solution is to enhance the speed and accuracy of COVID-19 screening and support healthcare professionals in their efforts to combat the pandemic.

Over the last few decades, Computer Vision (CV) and Deep Learning (DL) have been widely used in various Computer Aided Diagnosis (CAD) applications especially in radiology, where CAD system is used to help radiologists analyze and interpret medical images like mammography, X-ray, CT-scan, etc. Recently, radiologists have noted that chest X-ray images show distinctive marks of pneumonia caused by COVID-19 [2]. Several studies investigated automatic COVID-19 detection from chest X-ray images [2], [3]. However, despite the promising results achieved in the reported works, the deployment of the developed DL models has not been comprehensively studied [4]. Deploying DL models in mobile applications will allow the scientific community to experiment those models in real conditions by reaching a large population. On the other hand, the proposed application will be highly portable, simple to use, and inexpensive. Thus, such software will allow primary screening for COVID-19 in areas where medical expertise and test kits are not available or insufficient (rural and remote areas, hot spots, and clusters of at-risk populations).

Our study had the objective of assessing the use of deep learning, in particular convolutional neural networks (CNNs), in the context of chest X-rays for patients suspected of having COVID-19 pneumonia. This included its capability for directly diagnosing COVID-19 and its ability to distinguish COVID-19 from other community-acquired types of pneumonia. The overarching goal was to introduce an effective tool that could aid in COVID-19 pneumonia diagnosis, either by offering a second opinion to radiologists or by providing a preliminary assessment when a radiologist is not readily accessible.

To achieve these objectives and considering the existing body of research that consistently demonstrates the superior performance of CNNs in image classification tasks, we undertook the training and cross-validation of an Inception-V3 based architecture. We assessed the model's performance using two distinct sets of independent datasets. This model was then deployed in a mobile application to assist in the differentiation of COVID-19 cases.

The paper is structured as follows: Section II presents background and related works. Materials and methods are presented in Section III followed by results and discussion in Section IV. Section V describes the software, including its architecture and functionalities. Finally, Section VI concludes the paper and discusses future directions for research.

II. Background and Related Works

Multiple studies have been conducted to explore the automatic detection of COVID-19 from CT-scan and chest X-ray images, showing encouraging outcomes. The initial set

of research focuses on CT-scan images [3], [5] which can effectively identify to detect deep-rooted changes in lung tissue. Despite their promising detection rates, these methods involve higher costs and expose patients to potentially harmful radiation doses [6].

The second group of studies revolves around the investigation of chest X-ray radiographs [7], [8], [9]. These images are the most commonly used radiographs in the medical field. Besides detecting changes in the lungs, they are valuable for swiftly assessing breathing or heart-related issues without causing discomfort to the patient.

Due to these advantages, our paper specifically focuses on the second group of approaches that utilize chest X-ray images. This type of medical images is prioritized because it offers the lowest cost and are widely accessible in medical institutions for detecting thoracic pathologies.

To date, numerous methods have been developed and introduced for the automatic detection of COVID-19 using X-ray Images. In our prior publication [10] we presented a comprehensive summary of the most significant works in this area. It is evident that the majority of the proposed approaches rely on Transfer Learning (TL), where models are pre-trained on external large datasets and adapted to the target task. The use of this training technique is mainly due to the lack of COVID-19 data available for the community.

Recently, Cohen et al. [11] conducted a study investigating the cross-domain performance using Densenet [12]. They found that the model's performance was unsatisfactory when evaluated on datasets from different sources. On the other hand, several researchers have utilized various datasets from multiple origins. For example, Narin et al. [9] proposed a pre-trained convolutional neural network based on Res-Net50 with a binary classification approach (COVID-19 and Normal). The dataset they used was obtained from the open-source GitHub repository proposed by Cohen et al. [13] and the "ChestX-ray8" database proposed by Wang et al. [14]. For evaluation, they employed a 5-fold cross-validation method and achieved an accuracy of 98%.

Apostolopoulos et al. [8] conducted a study where they gathered X-ray images from different sources [13], [15]. They assessed the performance of state-of-the-art convolutional neural network architectures using a Transfer Learning strategy. They demonstrated that VGG19 and MobileNetV2 could extract significant biomarkers related to COVID-19 with accuracies of 98.75% and 96.78%, respectively, using binary classification.

In another work, Wang et al. [16] introduced a tailored convolutional network architecture called COVID-Net for detecting COVID-19 cases from chest X-ray images. They collected a dataset of 13,975 chest X-ray images from 13,870 patient cases to train and evaluate their model, achieving an accuracy of 93.3Meanwhile, Zhang et al. [17] conducted their experiments on two combined datasets [13], [14]. They proposed a deep learning model for COVID-19 detection, comprising a backbone network for high-level feature extraction, a classification head, and an anomaly detection head. Their model achieved a sensitivity of 96% and a specificity of 70.65%.

Consequently, existing studies have focused on using transfer learning for the COVID-19 detection from medical images. Moreover, they present potential limitations, including:

- **Limited Data Availability:** Many existing works suffer from a scarcity of labeled data, making it challenging to train accurate and robust models. COVID-19 X-ray images are relatively rare, and obtaining large datasets for training can be difficult;
- **Data Imbalance:** The datasets used for training often suffer from class imbalance, with a disproportionate number of negative cases (non-COVID-19) compared to positive cases (COVID-19). This can lead to biased models that are better at detecting the majority class but perform poorly on COVID-19 cases;
- **Generalization Issues:** Models developed in one region or using specific equipment may not generalize well to other regions or types of X-ray machines. There can be significant variability in how X-ray images are captured and processed;
- **Evaluation Bias:** Some studies may suffer from evaluation bias, where models are tested on the same datasets from which they were trained or on datasets that are too similar. This can result in overly optimistic performance metrics;
- **Limited Clinical Validation:** While models may achieve high accuracy in identifying COVID-19 from X-ray images in research settings, their clinical validation and real-world performance may be less certain. More rigorous clinical testing and validation are needed;
- **Ethnic and Age Bias:** Some models may not perform equally well across different ethnic groups or age ranges. Bias in the training data can lead to disparities in model performance;
- **Overfitting and Noise:** Overfitting to noise or irrelevant features in the data can be a problem, particularly when dealing with small datasets. Models may learn to exploit artifacts in the images rather than true COVID-19 patterns;
- **Resource Intensive:** Deep learning models for image classification, particularly those used in healthcare can be computationally intensive and require substantial hardware resources, limiting their practical deployment in resource-constrained settings;
- **Continuous Updates:** The evolving nature of the COVID-19 virus and the emergence of new variants may require continuous updates and retraining of models to maintain their effectiveness.

Addressing these limitations is crucial for the development of reliable and clinically viable automatic COVID-19 detection solutions from X-ray images. Along this direction, and to overcome these limitations, the main contributions of this work can be summarized as follows:

- The experiments in this paper utilize a fairly large and balanced dataset in terms of classes, collected from

different data sources, including various ethnic groups and age ranges;

- The model's performance was evaluated using two distinct sets of independent datasets to assess its generalization capabilities.
- The CLAHE histogram equalization method was applied to enhance data quality;
- The trained model was deployed in a mobile/web application, offering the potential for rigorous clinical testing and validation. This application will also facilitate the collection of other clinical data related to COVID-19.

III. Materials and Methods

A. Dataset

In order to assess the model's predictive ability and generalization capability, we established two distinct datasets. The first dataset gathers 3429 X-ray images from the COVID-19 Radiography Database [18], [19], and Chest Imaging data collection [20] sources. This dataset is divided into three subsets: training, development, and test1. While for the second dataset, we used images from the RSNA Pneumonia Detection Challenge [21] and Chest X-rays Radiopaedia [22] including 758 X-ray images. This set, named test2, will be used to make a double evaluation of the model in order to check the model's ability to generalize.

For both datasets, we consider three classes: Normal, COVID-19, and other pneumonia, given that the last class gathers bacterial and viral pneumonia. Data distribution over training, evaluation, and both test sets is presented in Table I.

TABLE I. Label Distributions in Training, Development and Test Sets

	Train	Dev	Test1	Test2	Total
NORMAL	804	112	227	236	1379
COVID-19	799	116	230	132	1277
OTHER PNEUMONIA	798	115	228	390	1531

B. Model Architecture

The experimental dataset used in this study consists of 4,187 X-ray images, which is considered insufficient for training a neural network with a deep architecture. Consequently, the COVID-19 detection system proposed in this research relies on the Inception-V3 model architecture that is pre-trained on the ImageNet dataset [23].

The original Inception-V3 architecture is composed of two main sections: feature extraction and classification. The feature extraction part includes five convolutional layers, each followed by batch normalization, two pooling layers, and 11 inception modules. Each inception module consists of a combination of parallel convolutional and pooling layers. The second part of the architecture comprises fully-connected and softmax layers, responsible for the final classification task.

To customize the Inception-V3 model for our specific task, we made modifications by removing its original classification part. Instead, we constructed a custom classification section to suit the number of classes in our dataset. The customized part consists of an average-pooling layer, followed by two fully connected layers. To prevent overfitting during training, each of these three added layers is accompanied by a dropout layer, randomly deactivating activations with a probability of 0.4.

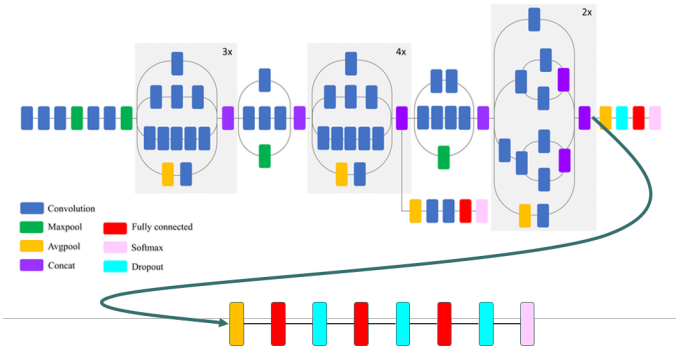


Fig. 1. The customized Inception-V3 architecture.

Finally, as the output for classification, a softmax layer was incorporated, providing the probabilities for each class. The model architecture is visually depicted in Fig. 1.

C. Performance Evaluation

In addition to measuring classification accuracy, we also employed sensitivity in Eq. (1) and specificity in Eq. (2) as metrics for evaluating the COVID-19 detection performance. Sensitivity, also known as True Positive Rate or Recall, represents the model's ability to correctly identify COVID-19 cases, and can be calculated using Eq. (1):

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

While Specificity measures the model's ability to correctly identify normal (non-COVID-19) cases and can be calculated using Eq. (2):

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

Where:

- TP (True Positives) refers to the number of correctly identified COVID-19 cases;
- FN (False Negatives) refers to the number of actual COVID-19 cases that were incorrectly classified as non-COVID-19 cases by the model;
- TN (True Negatives) refers to the number of correctly identified normal cases;
- FP (False Positives) refers to the number of normal cases that were incorrectly classified as COVID-19 cases by the model.

IV. Results and Discussion

In the first set of experiments, we investigated the fine-tuning of the Inception-V3 model. Table II shows the customized Inception-V3 model's performance of the test set with and without adaptation of the Inception-V3 convolution part. In the first line, the Inception-V3 model was pre-trained on ImageNet dataset and only the customized layers were trained on the chest X-ray images as explained in Section III. While in the second line, the pre-trained Inception-V3 model is fine-tuned by gradually unfreezing the top layers of the convolutional part of the model on the chest X-ray images and the customized layers are trained from scratch.

The results, in Table II, show that using Inception-V3 convolution parts without any adaptation on the new task give acceptable results with an accuracy of 95.62%. However, when the model is fine-tuned (adapted) on the new task the results are even better with an accuracy of 98.98% and a sensitivity and specificity of 98.25% and 99.34%, respectively.

TABLE II. COVID-19 Detection Results using Inception-V3 with and Without Adaptation on the Test1 Subset

	Accuracy	Sensitivity	Specificity
Inception-V3 without adaptation	95.62	92.11	97.37
Inception-V3 with adaptation	98.98	98.25	99.34

In the next experiments, we look at the effect of data preprocessing on the model's generalization while preserving fine-tuning when training our model. We apply the well-known Contrast Limited Adaptive Histogram Equalization (CLAHE) on our X-ray images while analyzing the model's performance on the second test set (Test2). The results reported in Table III.

TABLE III. COVID-19 Detection Results using Inception-V3 with and without CLAHE Preprocessing on Test2 Subset

	Accuracy	Sensitivity	Specificity
Inception-V3 with adaptation	16.62	40.91	11.50
Inception-V3 with adaptation + CLAHE	81.27	93.18	78.75

It can be seen by comparing the classification accuracies achieved when evaluating the model on the second test set that without data preprocessing, the model shows a pore generalization, where the classification accuracy drops from 98.98% on the first test set to only 16.62% on the second one. On the other hand, adopting data preprocessing leads to a significant improvement in performance generalization on the second test set, with an accuracy of 81.27%.

This improvement in the model's generalization could be explained by the consistent quality of X-ray images from each class coming from the same sources. Therefore, the model is very sensitive to image quality, and when re-evaluated in a real-world test set (test2) where it encounters images of varying quality, it may not perform as well as it did during the initial evaluation.

In summary, experiments show that the adaptation of the CNN layers of the pre-trained Inception-V3 model on COVID-19 X-ray images leads to better results as compared to using the model without any adaptation. Furthermore, the results confirmed that image preprocessing is essential in medical image analysis as it enhances the quality of medical images, improves the performance of classification models, and ensures the accuracy and reliability of diagnostic and analytical results.

V. Software Description

We embedded the trained Inception-V3 model into a web and Android mobile application for automated COVID-19 screening. The proposed application is called CovidChecker and it is publicly available on GitHub [24].

The app is designed to identify anomalies and estimate the extent of lung infection in real-time, and automatically detect high-risk patients with pneumonia or other COVID-19 related pulmonary symptoms.

The app enables users to complete a questionnaire and submit chest X-ray images for COVID-19 screening. The images and accompanying metadata will be used to augment the training dataset of DL models, thereby improving the accuracy and reliability of COVID-19 tests over time. To protect privacy, the data collected from users will be stored on secure servers for research purposes only.

A. Software Architecture

The software architecture, as shown in Fig. 2, illustrates the implementation of the web and mobile application utilizing Flask, MongoDB, and the DL module in the backend, jQuery Ajax in the web frontend, and Retrofit for the Android mobile application. The architecture is designed as follows:

Web Frontend: The web frontend is developed using jQuery Ajax to communicate with the Flask backend. Through the web user interface, the users can capture images using their computer's webcam and complete a questionnaire to assist the decision-making.

Backend: The backend is built using Flask, a RESTful API for the Python web server. It receives requests from the frontend and process them. MongoDB is used as a database to store images, classification results, and user data. All data transferred to the application is fully anonymized, including metadata. It has been hypothesized that an individual is not identifiable from their chest radiographs.

DL Module: The best DL module trained in section IV is loaded into the Flask backend and used to process images submitted by users.

Android Mobile Application: The Android mobile application uses Retrofit to communicate with the Flask backend. The mobile user interface allows the user to take images via their smartphone camera and fill out a questionnaire, similar to the web application.

The architecture is designed to be fast, scalable, and user-friendly for both web and mobile users. It can also handle scenarios where the mobile application loses connectivity to the backend.

B. Data Privacy Policy

The information that is collected will be transmitted from the user's phone, when they are connected to the Internet, to a secure database server. Once the data has been successfully transmitted, it will be deleted from the user's device. We do not

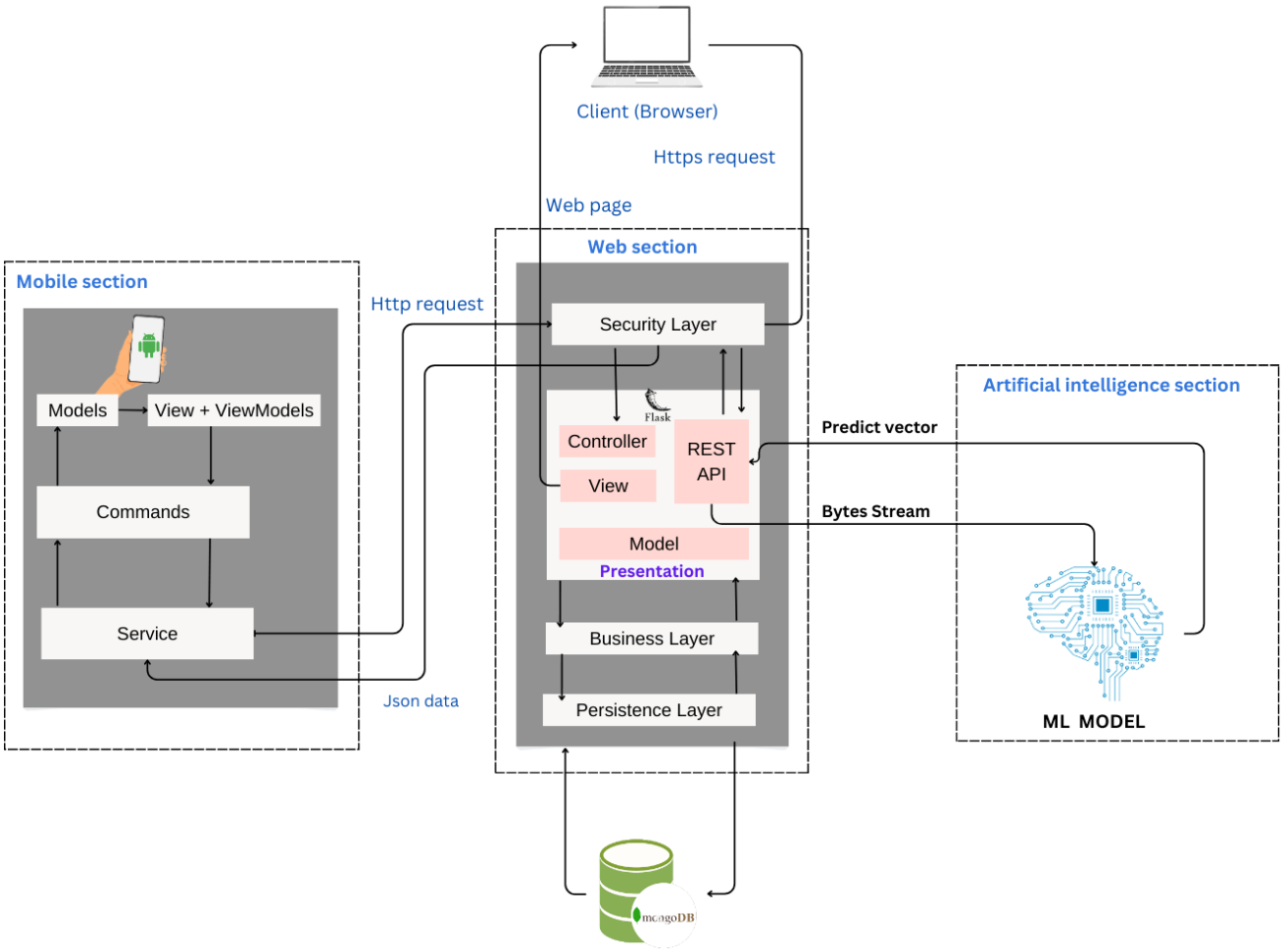


Fig. 2. Software architecture.

collect any personally identifiable information such as email addresses or other explicit personal identifiers. Since the data collected is sensitive and contains demo-graphic information, all data transferred to the database server is fully anonymized, including metadata. Furthermore, we established a sharing agreements for the data under the terms and conditions agreement of the mobile app.

C. Illustrative Example

The main functionality of this app is self-screening, which provides an automatic and rapid computer-aided diagnosis of COVID-19 from chest X-ray images. During this stage, the app will collect some basic demographic and medical history data, as well as the X-ray image through a quick questionnaire (see Fig. 3).

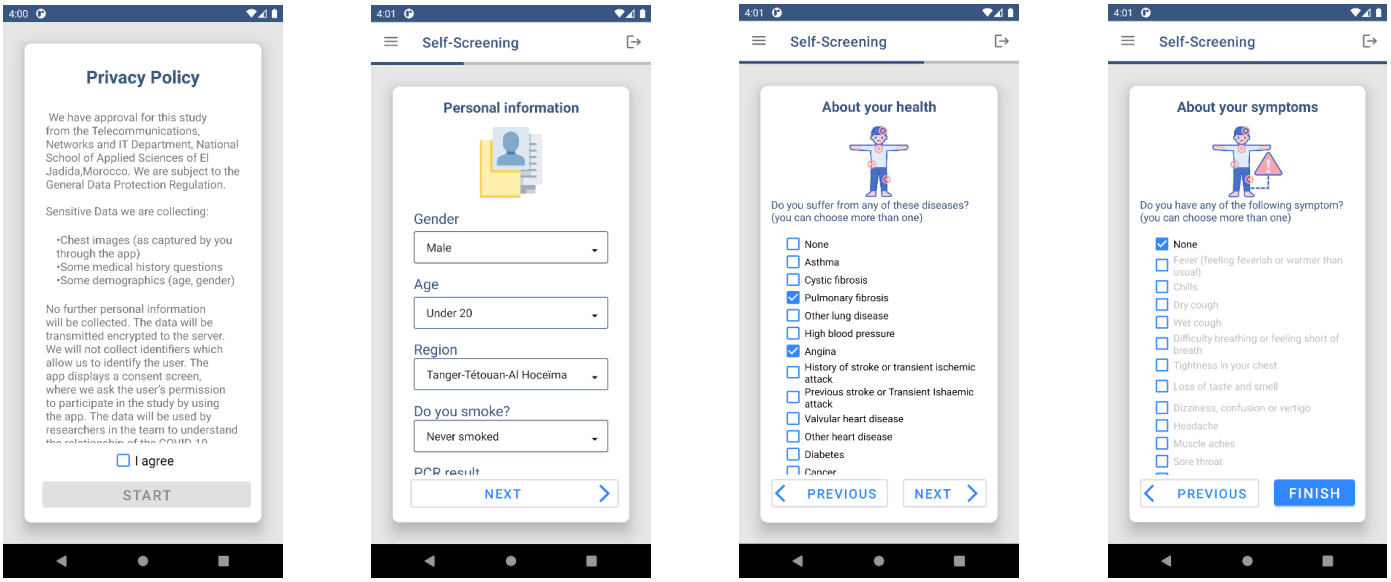
Before proceeding with the self-screening, users are required to accept the terms and conditions governing the app's use and agree to the data privacy policy that outlines how data is collected, stored, and managed. Second, after providing basic demographic information, such as their gender, age, region, and smoking history, the users will be asked whether they have previously tested positive for COVID-19. Then, they will be prompted to enter their medical history and any current symptoms they may be experiencing.

Finally, the user uploads his/her X-ray image and would be able to find the results within a matter of seconds. The uploaded image is sent to the backend server for analysis, where it undergoes preprocessing before being passed to a pre-trained image classification model. The model determines whether the input image shows signs of COVID-19 or not, and sends the results back to the mobile app in real-time. The app user receives the screening results and guidelines based on the X-ray analysis results. The results are presented as the probability of having COVID-19, which is determined by the output of the DL module. If the probability is less than 50%, the results page will appear in green to indicate safety (see Fig. 4), while a probability exceeding 50% will cause the results page to turn red (see Fig. 5).

D. Software Impact

The study on using transfer learning for COVID-19 detection from X-ray images can have a significant impact on various parties involved in healthcare.

For patients, this system could provide a faster and more accessible diagnosis, which could lead to earlier treatment and



(a) (b) (c) (d)

Fig. 3. Mobile App Self-screening survey interfaces.

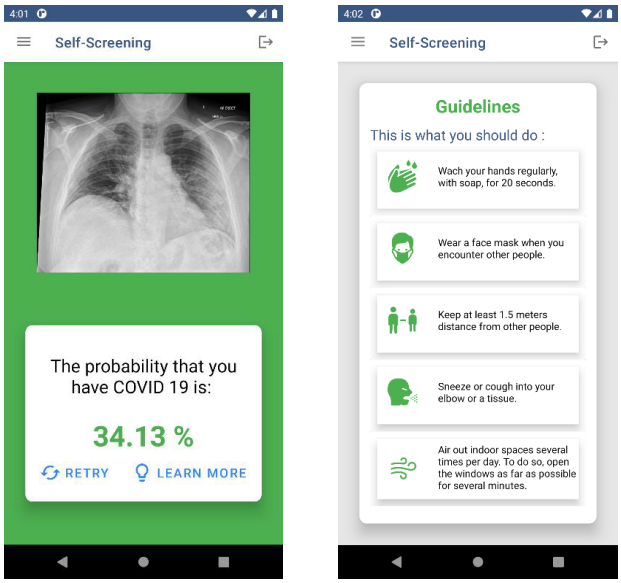


Fig. 4. Screening result and guidance page: Non-infected case.

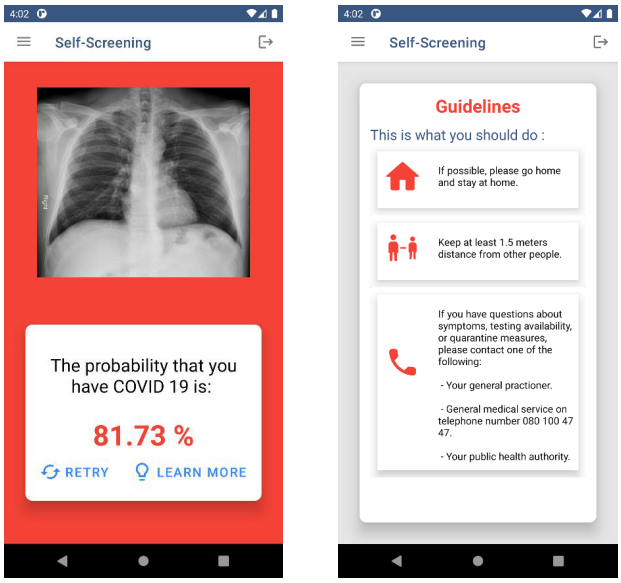


Fig. 5. Screening result and guidance page: Infected case.

better outcomes. Additionally, the use of X-ray images can be less invasive than other diagnostic methods, such as RT-PCR testing, and can be done onsite, reducing the need for patients to travel to a medical facility.

For healthcare providers, the proposed system could serve as an aid in the diagnosis process, reducing the time and effort required for manual analysis of chest radiographs and allowing healthcare professionals to focus on tasks that are more critical. This could also help to improve the accuracy and consistency of COVID-19 diagnoses, as the system can provide a second opinion when a radiologist is not immediately available.

Moreover, the successful implementation of this system in the larger healthcare industry could lead to the wider adoption of computer vision and deep learning techniques in the diagnosis of various diseases. This could significantly improve the efficiency and accuracy of medical diagnoses, ultimately leading to better patient outcomes and reduced costs for the healthcare system.

Overall, the proposed system has the potential to improve the efficiency and accuracy of COVID-19 diagnoses, providing benefits to patients, healthcare providers, and the broader healthcare industry.

VI. Conclusion

In conclusion, this research paper presents an innovative solution in the form of a mobile application that employs machine learning, computer vision, and deep learning to analyze X-ray images for the rapid detection of COVID-19 symptoms. The study focuses on the use of convolutional neural networks (CNNs) and transfer learning to diagnose COVID-19 from chest X-ray images, showcasing their effectiveness in image classification. The results demonstrate that adapting the Inception-V3 model with fine-tuning leads to superior accuracy, sensitivity, and specificity in COVID-19 detection.

The paper addresses various challenges and limitations encountered in the development of automatic COVID-19 detection solutions, such as data availability, data imbalance, generalization issues, and the need for clinical validation. It also highlights the importance of data preprocessing, such as CLAHE histogram equalization, in enhancing the quality of medical images and improving classification model performance.

The research introduces "CovidChecker", a mobile application that enables users to self-screen for COVID-19, submit chest X-ray images, and provide valuable data for improving the accuracy of COVID-19 tests over time. The application is designed with a strong focus on data privacy, ensuring that user data is anonymized and securely stored.

Acknowledgment

This work was supported by the CNRST Morocco under contract no UCD-CNRST-5.

References

[1] R. N. Binny, P. Priest, N. P. French, M. Parry, A. Lustig, S. C. Hendy, O. J. Maclaren, K. M. Ridings, N. Steyn, G. Vattiato *et al.*, "Sensitivity of reverse transcription polymerase chain reaction tests for severe acute respiratory syndrome coronavirus 2 through time," *The Journal of infectious diseases*, vol. 227, no. 1, pp. 9–17, 2023.

[2] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia, "Correlation of chest ct and rt-pcr testing for coronavirus disease 2019 (covid-19) in china: a report of 1014 cases," *Radiology*, vol. 296, no. 2, pp. E32–E40, 2020.

[3] J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, Q. Chen, S. Huang, M. Yang, X. Yang *et al.*, "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography," *Scientific reports*, vol. 10, no. 1, p. 19196, 2020.

[4] Z. Chen, Y. Cao, Y. Liu, H. Wang, T. Xie, and X. Liu, "A comprehensive study on challenges in deploying deep learning based software," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020, pp. 750–762.

[5] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Q. Ni, Y. Chen, J. Su *et al.*, "A deep learning system to screen novel coronavirus disease 2019 pneumonia," *Engineering*, vol. 6, no. 10, pp. 1122–1129, 2020.

[6] L. J. Kroft, L. van der Velden, I. H. Girón, J. J. Roelofs, A. de Roos, and J. Geleijns, "Added value of ultra–low-dose computed tomography, dose equivalent to chest x-ray radiography, for diagnosing chest pathology," *Journal of thoracic imaging*, vol. 34, no. 3, p. 179, 2019.

[7] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "Covidgan: data augmentation using auxiliary classifier gan for improved covid-19 detection," *Ieee Access*, vol. 8, pp. 91 916–91 923, 2020.

[8] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and engineering sciences in medicine*, vol. 43, pp. 635–640, 2020.

[9] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," *Pattern Analysis and Applications*, vol. 24, pp. 1207–1220, 2021.

[10] R. Errattahi, S. F. Zahra, A. El Hannani, A. Abdelhak, H. Ouahmane, S. Mohamed *et al.*, "Investigating generalization in automatic covid-19 detection using deep learning," in *2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC)*. IEEE, 2022, pp. 1–6.

[11] J. P. Cohen, M. Hashir, R. Brooks, and H. Bertrand, "On the limits of cross-domain generalization in automated x-ray prediction," in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 136–155.

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[13] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," *arXiv preprint arXiv:2006.11988*, 2020.

[14] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[15] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[16] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific reports*, vol. 10, no. 1, p. 19549, 2020.

[17] J. Zhang, Y. Xie, Y. Li, C. Shen, and Y. Xia, "Covid-19 screening on chest x-ray images using deep learning based anomaly detection," *arXiv preprint arXiv:2003.12338*, vol. 27, no. 10.48550, 2020.

[18] M. E. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al Emadi *et al.*, "Can ai help in screening viral and covid-19 pneumonia?" *Ieee Access*, vol. 8, pp. 132 665–132 676, 2020.

[19] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. A. Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughaier, M. S. Khan *et al.*, "Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images," *Computers in biology and medicine*, vol. 132, p. 104319, 2021.

[20] Chest imaging data collection. [Online]. Available: https://twitter.com/ChestImaging/status/1243928581983670272

[21] Radiological Society of North AmericaRsna pneumonia detection challenge. (2018) Radiological society of north americarsna pneumonia detection challenge. [Online]. Available: https://www.kaggle.com/c/rsna-pneumonia-detectionchallenge

[22] F. Gaillard *et al.*, "Radiopaedia: building an online radiology resource." European Congress of Radiology-RANZCR ASM 2011, 2011.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[24] Errattahi et al. (2023) Covidchecker: Covid-19 diagnosis mobile application. [Online]. Available: https://github.com/errattahi/CovidChecker

Explicit Knowledge Database Interface Model System Based on Natural Language Processing Techniques and Immersive Technologies

Luis Alfaro[1], Claudia Rivera[2], José Herrera[3], Antonio Arroyo[4], Lucy Delgado[5], Elisa Castañeda[6]
Universidad Nacional de San Agustín de Arequipa, Arequipa, Perú[1,2,4,5,6]
Universidad Nacional Mayor de San Marcos[3]

***Abstract*—This work is focused on the proposal and development of an interface system model, based on natural language processing, immersive technologies and natural user interfaces, for the interaction with Explicit Knowledge databases. Five phases were proposed: The user testing characterization, the establishment of the state of the art and the theoretical foundation, the design and development of software, the system implementation and the functional tests and evaluation of the usability of the interface model. In order to establish the user testing characterization and the corresponding theoretical framework, the expert guide on Knowledge Management and Virtual Reality was followed, based on the approach of Usability and Computer Ergonomics compatible with the ISO 9241 standard. The traditional interfaces and the proposal in this work were evaluated for each of the metrics defined by the ISO 9241 standard, considering the dimensions of effectiveness, efficiency and satisfaction. The statistical test of "T-Student" established that there is enough evidence to confirm the existence of the following significant differences: Effectiveness is lower using the proposed interface model; efficiency and satisfaction is higher using the proposed interface model. Based on the conducted tests, it can be established that the proposed interface model is superior to the traditional interface in terms of the "Efficiency and Satisfaction" dimensions and inferior in terms of "Effectiveness." Consequently, it can be concluded that the scientific article exploration model using VR and NLP is superior to the traditional model.**

***Keywords*—Knowledge management; explicit knowledge databases; natural language processing; natural user interfaces; Immersive technologies**

I. Introduction

Knowledge Management (KM) is one of the emerging academic disciplines which responds to the growing demands of the information society [1]. It is the result of the confluence or intersection of the areas of business administration, communication technologies and information systems, and is defined as the process of creation, storage and exchange of organizational knowledge. With the existence of a wide variety of modern organizations, objectives and cultures, KM becomes an extremely interdisciplinary area of operation which develops its own objectives and professional approach.

The knowledge-based vision asserts that knowledge management can facilitate innovative practices by transforming knowledge assets into new products and services through a series of management processes and activities. It can promote the exchange and distribution of knowledge necessary for innovation activities, thereby stimulating the generation of new ideas and ultimately enhancing innovation performance.

For [2], an important activity of the organizations is the integration of individual knowledge into collective knowledge. Today, there is a need to properly manage tacit and explicit business knowledge [3], which, in the latter case, may be document-based, having a direct relationship with the productivity of the workforce and the efficiency of the organization. In many organizations, much of the business documents are not structured and do not have an adequate management system. These documents take many forms, such as office documents, images, reports, e-mails, drawings [4].

The contemporary organizations have a large number of documents which can be located in various containers or repositories. These documents can be lost, misplaced or even stored by someone without the knowledge of other collaborators. Since the files may have been stored on several separate computers, locating and retrieving them can be a complex task. To this end, Web-based applications can be equipped with the ability to manage various types of documents, such a business, process and research documents, which can be explored with immersive technology resources [5], as well as other types of documents in public and private organizations. Likewise, web-based applications are much more accessible and share some common characteristics. They are distributed, which means that storage and access takes place in different physical locations. Similarly, for [6], Scientific data repositories have a key role in science.

Tasks of exploration, search, and information collection are carried out using computer systems with traditional interfaces, which may include forms, menus, windows, and hyper medial or hypertextual elements. Typically, these tasks involve using the author's name, keywords, publication year, and/or digital files, along with some additional search parameters. The results are presented in a text list format [7].

The information indexed in a repository can grow rapidly, and when using traditional interfaces, there can be difficulties in exploring and selecting different types of documents. This can impact the search processes conducted by collaborators, making them tedious and resulting in unexpected outcomes.

The originality and motivation of this research lie in an attempt to contribute to the design and exploration of tools for optimizing the review tasks of explicit documents within private, public, and research-based organizational settings. These

tools aim to facilitate the work of executives, collaborators, and researchers through novel forms of representation and interaction with textual information. The approach involves utilizing efficient and agile system models for searching and selecting from the extensive variety and diversity of explicit documents stored in organizational repositories.

In this paper, Section II, describes methodology, Section III describes the theoretical framework; Section IV explores and analyses immersive technologies; Section V explores and analyzes the works related to the development of interaction systems with Textual Web-based repositories using Virtual Reality (VR) and NLP; Section VI describes the proposed model; Section VII evaluates the model in relation to a traditional one; Section VIII evaluates the results. In conclusion, the Section IX presents the conclusions and recommendations for future work.

II. Methodology

For the development of the theoretical framework, the required elements for a systematic literature review and meta-analysis (PRISMA) were also utilized. Consequently, a search string was defined concerning the topics of interest, identifying relevant keywords to construct the search string. This search string was applied to recognized databases such as IEEE Xplore, Science Direct, Web of Science, and Scopus, considering that the studies were articles from journals and conferences published in the last 20 years. Duplicate studies were removed, those that did not contribute to the study were excluded, and applying inclusion and exclusion criteria, a total of 43 studies were used in the research.

Five phases were proposed for the development of the model, as described in Fig. 1.

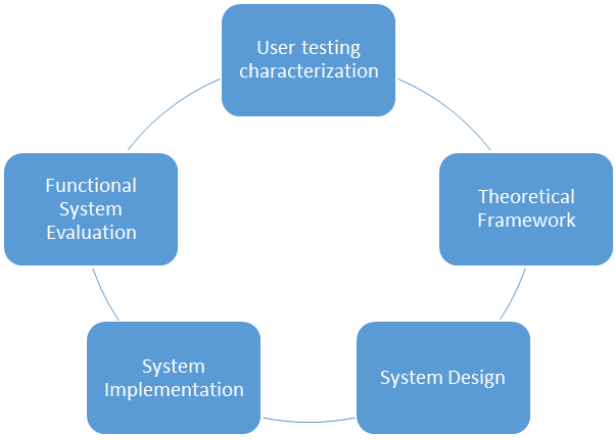


Fig. 1. Proposed methodology.

In the "Using Testing Characterization" phase, the theoretical framework was established, incorporating Knowledge Management, Web-Based Repositories, Natural Language Processing, and Immersive Reality along with their technological resources. This phase also involved the design and development of software, as well as the criteria required for evaluating the model, referred to as the usability interface.

In the "System Design" phase, the system was designed by establishing the background and reviewing related works. It was noted that there are various approaches to addressing and resolving the problem.

Moving on to the "System Implementation" phase, the different components of the system were implemented in the following steps: Physical environment setup for immersive experiences, Natural Interaction System Model, Implementation of Web Crawler for information retrieval from the model, Speech recognition and text-to-speech engine and Implementation of the syntactic analyser.

For the categorization of user tests, the expert guidelines in Knowledge Management (KM) and Virtual Reality (VR) were considered. These guidelines were based on the usability and ergonomics of computer science, aligning with the ISO 9241 standards. After configuring the interconnected devices and initializing the software modules, the evaluation of the model for exploring scientific articles using VR and NLP was conducted, alongside the traditional model of the scientific database ALICIA. Five users participated in the experiment, a number recommended by Nielsen [43] for evaluating interfaces, a criterion considered for obtaining the required dataset.

Participants were assigned the task of exploring scientific articles for a duration of 30 minutes using the interfaces of both models. During the experiment, users could request assistance from an expert who observed and documented the proceedings. Subsequently, the results were tabulated based on the metrics of ISO 9241, specifically efficiency, effectiveness, and satisfaction. These results are discussed in Section IX.

III. Theoretical Framework

In this section, a historical review and state-of-the-art analysis are conducted to explore the theoretical foundation required for the research.

A. Knowledge in the Organizations

The Knowledge Management System (KMS) is focused on information and enterprise social media, as well as on the connection and communication between employees. These two perspectives correspond to the categorizations of knowledge theory and KM. KM recognizes that virtually all jobs involve "knowledge work," and employees are "knowledge workers" to some extent. Therefore, it places more emphasis on knowledge than on manual skills. Consequently, the creation, sharing, and use of knowledge are among the most important activities in nearly every organization.

Existing research provides the theoretical framework for carrying out these activities, and it is proposed that knowledge sharing can contribute to companies achieving innovation objectives, thereby improving organizational performance [8]. The positive impact of KM capabilities on organizational performance has also been studied [9], considering that KM contributes to a company's innovation capacity [10].

Demuner-Flores and Nava-Rogel [11] emphasize a close relationship between knowledge and action, which also involves understanding and comprehension. The knowledge that an individual possesses is a product of their experience and incorporates the norms for evaluating novel contributions from their environment. For [12], knowledge involves framed experience, values, contextual information, and informed intuition.

It provides a framework for evaluating and incorporating new experiences and information. Knowledge originates and is applied in the minds of knowledge workers, but it can also be integrated into organizations through documents, repositories, practices, routines, and organizational standards.

KM is based on the concept of collecting, organizing, analyzing, and sharing knowledge in a way that employees can easily access and use it [13]. Therefore, a competitive organization relies, among other things, on the effectiveness of its workforce in creating new knowledge, sharing it within the organization, and using it to gain a competitive advantage.

Ultimately, KM aims for a positive impact on business value generated from knowledge. This is achieved through the creation of processes, culture, and technology in an environment where knowledge exchange is considered essential and occurs seamlessly. Thus, this research work can contribute to the exploration of textual documents in repositories of explicit web-based knowledge.

On the other hand, in the KMS, the following types of knowledge are defined:

1) Explicit Knowledge [14]: It classified as structured or unstructured, can be symbolically transferred, meaning it can be conveyed through linguistic or computational mediation, and can be documented and stored in databases. Some examples include written procedures, instructional manuals, lessons learned, best practices, as well as research results in the form of indexed scientific articles in repositories such as Web of Science, Scopus, Scielo, etc., or in academic data repositories of universities and other institutions, for the exploration, search, and collection of information [15]. Some databases, document types, and spreadsheets are examples of structured data and information, which are organized to allow for future retrieval. This work is focused on the exploration of emails, images, training courses, and audio and video selections, which are examples of unstructured knowledge because there is no clear way to reference them for retrieval.

2) Tacit Knowledge: It is knowledge hidden from the consciousness of the individual who possesses it [15]. It resides in the minds of human beings, and capturing or encoding it is not easy because it involves a complex process. Knowledge in its entirety resides in a tacit dimension. It is expressed through human actions, stories, as well as competencies, skills embedded in an individual's worldview, experiences, attitudes, evaluations, and perspectives that are often taken for granted. It can be observed through action and is less concrete compared to explicit knowledge. It can be described as a "tacit understanding" of something, which is difficult to document in writing or in databases. It is considered more valuable because it provides context for people, ideas, and experiences.

Managing both types of knowledge requires applying various methods and approaches, taking into consideration their unique characteristics. Some other types of knowledge in the context of corporate KM: (1) Individual knowledge, in which only one entity owns it; (2) Collective knowledge, whose transfer is focused on interaction and goal achievement.

Finally, the aim of KM's contribution to positive business value generated from knowledge and its management will be achieved by creating an environment that encompasses processes, culture, and technology for seamless knowledge sharing. Hence, the importance of this research work, which can contribute to the investigation and exploration of textual documents in web-based repositories of explicit knowledge.

B. Web Based Repositories

Information systems designed to preserve and organize commercial, scientific, and academic materials are referred to as repositories. They are used to support operations, research, and learning, while also ensuring access to information [16]. Institutional repositories also consist of interoperable web-based services, dedicated to disseminating the resources of the results and analysis of institutional, scientific, academic, or business operations of institutions, based on the listing of a specific set of data (metadata), so that these resources can be collected, catalogued, accessed, managed, disseminated and preserved. Respect for compatibility and interoperability standards makes it possible for the content of a repository to be more easily retrievable, not only for the institution itself, but also for the scientific community and society.

Scientific databases consist of components that include information related to library materials such as books and other types of documents, including journals and other scientific publications. These materials encompass scientific articles, research papers, conference proceedings, books, among others. Scientific databases are typically in electronic format and are accessible via the internet. They contain abstracts, bibliographic citations, references, and often the full text or links to the full text.

The government institution of the National System of Science and Technology and Technological Innovation of Perú (CONCYTEC), has subscriptions to various scientific databases, to which it provides free access and makes available to the scientific and academic communities and general population. The repository of scientific and technological information and innovation, called ALICIA, provides free access to intellectual production and research in science, technology and innovation, developed in the public sector and private institutions, with public funds.

The bibliographic review carried out in this part of the work, allows us to establish that the research in the design and development of friendly interfaces, from a unique and original perspective, resulting from the confluence of the use of methodologies and techniques, such as NLP, Natural Users Interfaces and VR, that facilitate the search and exploration in Explicit Knowledge database, can be very useful for the collaborators of the business organizations, researchers, academics, scientists and professionals in general, since it will not only facilitate the tasks that these carry out, but will also shorten the time spent, as well as diminish the costs, a fact that can undoubtedly contribute with the capacity of research.

C. Natural Language Processing (NLP)

NLP refers to the hardware and/or software elements of a computer system focused on analyzing or synthesizing spoken or written language [17]. NLP contributes to the retrieval of texts for user reading, which can vary in size from a paragraph to a book. Innovations in emerging technologies have made tasks such as storage, searches, and the retrieval

of complete texts or parts of them from online documents relatively straightforward [18].

Many academic libraries have implemented various services to support users in their learning and research activities through chat. Chat serves as an important channel for accessing resources and services in university repositories. Repositories and libraries have accumulated a significant amount of data in the form of chat transcripts. Analyzing the content of these transcripts can help library officials understand user information needs, allocate library resources more efficiently, and improve the quality of chat reference services [3].

Conversational agents (chatbots) are programs that utilize natural language processing with a question-and-answer system, aiming to simulate intelligent dialogue with human interlocutors. This can be implemented through text messages or voice as a tool created to support the customer-company relationship, enabling virtual interaction through technology in the most human-like manner possible [19].

Recent applications of explicit database interfaces include two works by [20], who propose a Natural Language Query (NLQ) database query system (NADAQ) as an alternative solution. In their design, they introduce new models for translation, seamlessly merging deep learning with traditional techniques for database analysis and developing new techniques to allow the neural network to reject queries irrelevant to the content of the target database, and recommend that candidate queries be translated back into a tuned natural language. According to [21], the slow development of Natural Language Interfaces for Databases (NLIDB) stems from linguistic issues (such as language ambiguity) as well as domain portability. Brad et al. [21]assert that there is a demand for non-expert users to make queries to relational databases using their natural language instead of utilizing specified attribute values for the database domain. According to [22], the use of natural language instead of SQL-based queries enabled the development of NLIDB, a new type of processing that represents an approach to developing intelligent database systems aimed at improving the performance of flexible queries in databases.

Among the most important applications of NLP, we can mention human-machine interaction by voice, becoming one of the most successful applications with multiple uses, information retrieval and extraction, morphological, syntactic and semantic tagging, NLP for automatic responses to questions, text sentiment analysis, automatic text summarization, document classification, among others.

The review of the literature in NLP allows establishing the requirements of the users who demand the use of interfaces in their natural languages, making the research and development of research projects necessary, which contribute with the development of a model of interface systems for the interaction with Explicit Knowledge databases, which simplify and reduce the costs referred to the search and research tasks.

IV. Immersive Technologies

A. Virtual Reality (VR)

VR as defined by [23], is a system used to describe 3D environments generated by computers that can be explored and allow interaction with users. Users become part of a virtual world in which they are immersed and can interact with and manipulate objects, as well as perform various actions. VR is a simulated reality constructed using digital elements through computational systems. The construction and visualization of alternative reality require powerful hardware and software resources (e.g., immersion headsets, head-mounted displays, exoskeletons, power globes, 3D software, etc.) to make the creation of realistic immersive experiences in VR environments possible [24]. After many years of research and development, VR hardware and software are now available for use by the general public, researchers, and entrepreneurs.

VR can be classified into various categories, including virtual worlds, augmented reality (AR), mixed reality (MR), and extended reality (XR) [25]. Several researchers argue that VR and associated technologies contribute to stimulating creativity and enhancing various competencies. Furthermore, their presence has increased in the entertainment industry, generating interest in research areas such as education, knowledge management, psychology, engineering, architecture, among others. VR can be associated with Artificial Intelligence (AI) techniques to drive innovations in various fields, especially to enhance learning experiences by enabling direct experiences of real or imaginary phenomena [**?**]. As the design and development of VR and AI become more accessible to institutions, the levels of usage and dissemination could increase.

Liu et al. [27] argue that the essential elements characterizing VR are: Sensory immersion, through which users immerse themselves in a virtual world and can see their surroundings by using a Head Mounted Display. Interactive simulation involves recreating a virtual world and digitally representing physical objects through a computer, enabling implicit interaction that allows users a degree of control over their experiences using sensors and input devices like joysticks, keyboards, head-mounted displays, among others.

Immersive VR applications, where the user's perception is mediated entirely by visual and audio devices used in the virtual world. In this case, external information must be entirely isolated for the experience to be fully immersive. This technology can be costly and has some drawbacks, such as lower-definition images, computer overload, and environmental issues associated with simulators [28].

VR allows users to feel as if they were inside a simulated virtual world. Psotka [29] asserts that this contributes to a significant emotional factor, which facilitates cognition and improves information retention. The user's range of movement can also be expanded, allowing for navigation (useful for virtual tours). Immersion can be achieved in various configurations, such as using a large screen and anaglyph glasses, or in CAVE environments with a room containing four walls and a projector for each. Head-mounted displays (HMDs) are also used to project images that will be viewed by the user.

In this research, a VR immersion headset is used, which is comfortable for immersion due to its minimal space requirements compared to other devices. It also provides a greater sense of presence that contributes to memory retention and various cognitive processes [30]. However, it's important to note that in some users, VR can cause motion sickness, which is considered a significant disadvantage.

B. Natural User Interfaces (NUIs)

NUIs study new forms of human-computer interaction [31]. They are characterized by interactions that users find natural, common, and familiar. The design includes the use of new devices that make this type of interaction possible, not limited to just the use of a mouse and keyboard. They allow for gesture-based, touch-based, transduction-based, or body movement-based interactions, as well as voice command communication, among others. Innovations in input peripherals demand changes in the way we interact with digital screens and traditional interfaces like the mouse and keyboard, which are replaced by touch-based and motion-based interfaces [32].

NUIs represent a revolution in the field of computing, not only because they replace existing and widely used traditional interfaces but also because they enable the creation of new types of applications and original and innovative forms of interaction. These can be applied in various fields such as engineering, management, marketing, library science, and more.

C. Natural Devices Interaction

For [33], the emergence of new devices and interaction paradigms enables new forms of human-computer interaction, allowing users to interact with systems using body movements or transduction. Users can operate systems through intuitive actions such as gestures, touches, haptics, or speech. This offers a significant opportunity to enhance user experiences.

This literature review and the establishment of the state of the art have made it possible to recognize the immense potential of resources and elements in immersive technologies. Their use in proposals for interfaces with explicit knowledge databases will enable the conception and design of original and innovative models.

V. Related Works of Interaction Model Systems based on Immersive Technologies

The lack of flexibility in information systems has led users to employ their own strategies to carry out their daily activities in alignment with business objectives. Users fulfill their functions, enhance their performance, and save resources, especially time [34]. With the aim of formalizing the intentions of users of enterprise information systems for the execution of their daily tasks in line with their job roles and to improve their performance, Khodabandelou et al. [35] propose a method to discover the strategies users employ to achieve their personal objectives, aligned with business objectives.

The method is based on a supervised machine learning algorithm developed using general business activities and specific business rules. According to the authors, the model could be used to determine user behavior and enhance their performance. An advantage of this method is that it provides the option to work with flat or structured files. Additionally, the method allows for the structuring of a knowledge base that could be generalized for various types of businesses.

From another perspective, there is a considerable number of applications for exploring explicit knowledge databases that are designed to interact with traditional resources such as scientific databases [36]. However, there are some difficulties associated with the traditional environment. In the literature, several problems that hinder the search for scientific articles are reported, such as:

1) Traditional Devices: mostly have limited display space. It is assumed that newer generation LCD monitors have improved features [37], and they are currently the most commonly used means. Additionally, the visibility of LCD screens is lower when using an anti-reflective filter, so it is recommended for use when there is glare, a risk of visual fatigue, or when the user is performing tasks that do not depend on the detection of stimuli with low RGB values [38].

2) RepoVis: adopts the metaphor of visualizing source code listings hanging on a wall from a distance [39]. The complete visual description is the central piece of the structure (folders, files, and code) at a given moment in a software project. Source code files are represented as boxes with rows of colors that represent one or more lines of code.

3) Al-Amawi et al. [40]: propose the creation of a database to store explicit knowledge derived from lectures at a university. This database resembles a cognitive memory that grows and evolves over time, serving as a reference and repository of knowledge generated in the educational process. This is part of an e-learning system that provides benefits to students.

In the literature review, various works with different approaches were found, some of which make proposals that can contribute to the development of this research project.

VI. Model System Development

Below, the processes used for developing the different elements for building the model of the web repository interaction system based on immersive technologies and natural user interfaces are described.

A. Physical Environment Setup for Immersive Experiences

The equipment is configured, and the first step can be seen in Fig. 2 [15]. The Leap Motion was connected to the Oculus Rift, and then both devices were connected to a laptop using USB and HDMI ports. The installation software used was the Unity3D graphics engine.

(a) Schema setup (b) Actual operation

Fig. 2. Immersive environment configuration [5]

B. Usability for the Development of the Virtual Interface

Some usability principles should be considered for the design of the model in order to facilitate its use and proper handling of errors in its usage. Taking these principles into account in the design according to the defined and identified

target audience should contribute to maintaining product usability quality to achieve maximum user satisfaction. One of the most important criteria for evaluating the usability of a virtual environment is the sense of presence. VR environments have the advantage of generating simulations that can act on the imagination of the participant, creating the sensation of psychologically transporting him/her to another place, which can be real or imaginary. Immersion in a VR environment with a high degree of presence will be an experience close to reality and even more pleasant. Tasks with a 2D interface use interactive output devices such as: keyboard and mouse, and interact with 2D objects and various tools (Marsh, 1999). Interaction in a VR system is carried out in a different way, with resources whose ergonomic evaluation criteria of the interfaces they use, is different from what most users are used to, helmets, glasses, gloves, virtual rooms, to mention a few.

Jordan [41] established usability principles, describing how and why they influence usability. Klevjer (2006) argues that the sense of presence in an environment can be enhanced through the use of an Avatar, which is a digital character that can resemble a real person, an animal, or take on another appearance as required, considering the model's requirements and user preferences [15]. The image shown in Fig. 3 corresponds to the avatar developed for this work, a resource that can enable user engagement, allowing them to perceive it as an assistant with whom they can interact using voice commands. The avatar assumes basic functions for exploring explicit knowledge documents contained in web-based databases.

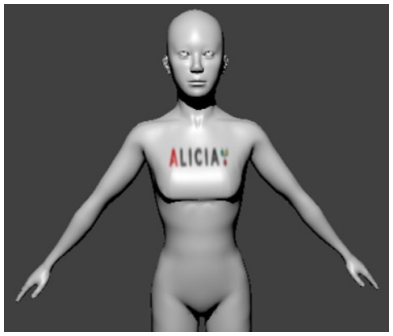


Fig. 3. Avatar of the model system.

C. Natural Interaction System Model

After the immersive environment was set up, different techniques for interaction in immersive environments were investigated and explored:

1) Direct Manipulation: Technique implemented using a virtual hand to accurately emulate human hand movements. The selection of an element involves holding an object in the same way as in the real world. In this technique, elements are required to possess solid and physical properties for better interaction. Fig. 4, shows the selection of a node through direct manipulation; which as can be seen, is similar to holding a physical object in the real world.

2) Gestural manipulation: Immersive VR elements can be selected with virtual hand gestures. Holding an object with the pinch metaphor involves performing the 'Pinch' gesture and releasing it by performing the 'Stop' gesture using the open palm. In Fig. 5a, the node is selected by pinching, and in Fig. 5b, the palm is opened to stop selecting. This technique is not completely intuitive, because both gestures may mean something different to the user.

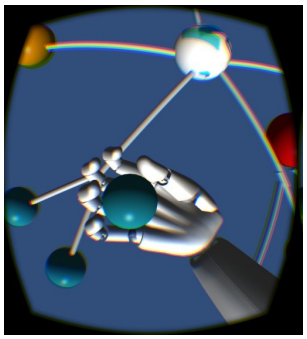
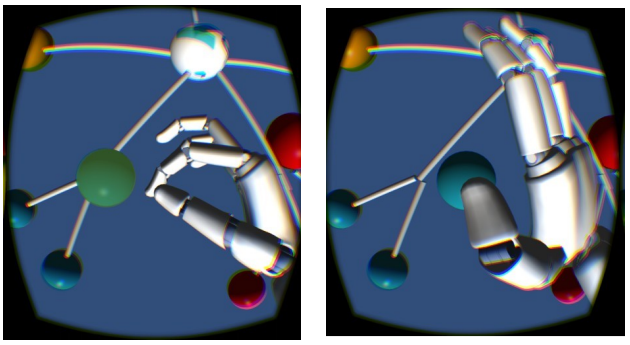


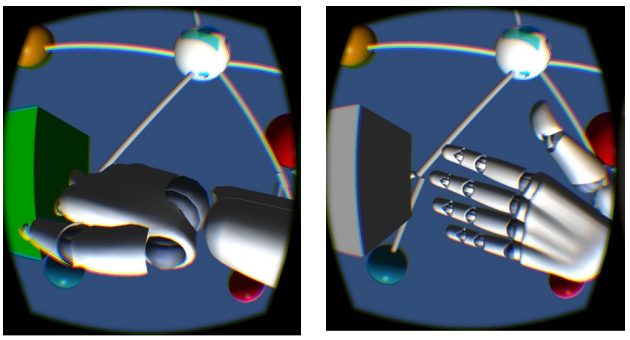
Fig. 4. Direct manipulation node selection.



(a) Item Selecting (b) Element releasing

Fig. 5. Gesture of a node selection example.

3) Artificial Manipulation: Winn [42], mentions the concepts of "Dimension", "Transduction" and "Reification", which allows the development of the corresponding metaphor, such as the crossing of objects. To leave the selection of the element, it is only necessary to go through it again. In Fig. 6a, a selected cube is visualized, changing color, and in Fig. 6b, the cube is visualized returning to its previous state.



(a) Item Selecting (b) Element releasing

Fig. 6. Artificial node manipulation example.

D. Implementation of Web Crawler for Information Retrieval

In order to carry out the tests, the interface model was connected to the ALICIA of Concytec scientific database and to enable the information retrieval, a program of the tracking type was developed and implemented, which allows automatic communication with the ALICIA and extracts the information required in the VR based interface through analysis of the source code of viewing concerning the respective visualization. The implementation involved investigating the pattern for performing the tracking and extraction for automated recovery.

E. Speech Recognition and Text-to-speech Engine

Intel Perceptual Computing features a library of raw audio-to-text conversion functions used to develop the recognition engine. To achieve accurate voice recognition, a certified wireless microphone was integrated into the prototype. A C# .Net library was developed to continuously capture user audio and then transmit it to Unity3D via Socket. A listening port was implemented to receive user voice commands.

F. Implementation of the Syntactic Analyzer

The Freeling library was used to perform the interpretation of the syntactic analysis of voice commands, taking into consideration the quality of the attributes of the functions for the treatment of the Spanish language. This process of syntactic analysis first involves breaking down the words in a sentence or phrase, to perform content analysis and add type labels: noun, verb, adverb, adjective, article, pronoun, etc. Fig. 7 shows an example of the analysis of a sentence, generating its respective syntactic tree.

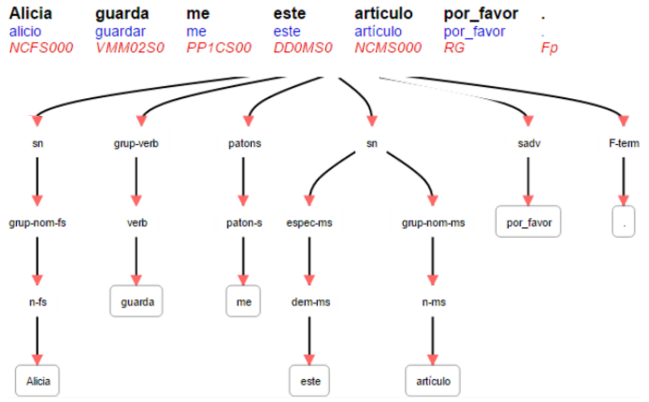


Fig. 7. Syntactic tree for a voice consultation generation.

The process diagram of the Proposed Model for document exploration in explicit knowledge databases, based on VR and NLP techniques, is presented in Fig. 8. The proposed model starts with a verbal query by a user who wants to search for a document in a web-based explicit knowledge database, then the speech recognition engine captures the speech signals and converts them into plain text. After the information is processed by the semantic analyzer, a syntactic tree is generated. Interaction and query commands are differentiated by the analyzer and then sent to the crawler, which establishes a connection with the virtual assistant to retrieve the information requested by the user. The results are visualized so that the user can begin

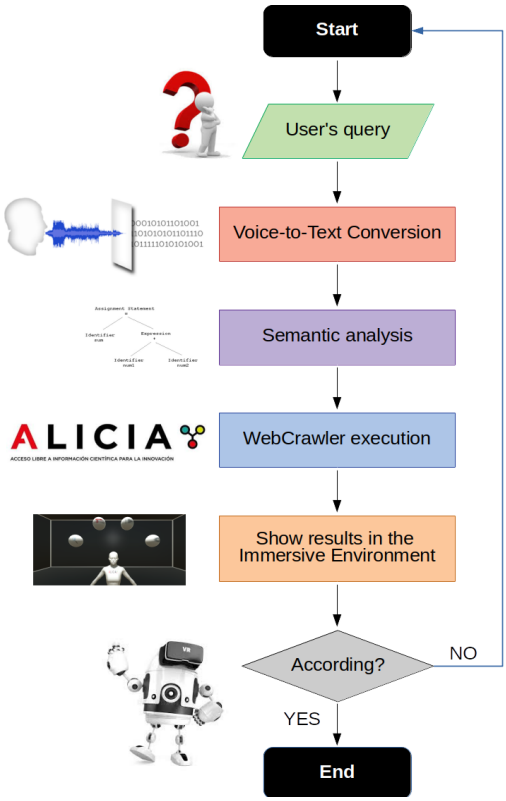


Fig. 8. Process diagram of the proposed model [5].

to interact. The process is concluded when the user agrees; otherwise, a new search is initiated.

VII. Evaluation of Proposed Model Regarding the Traditional

Usability is defined as the degree of satisfaction with which a product can be used by specific users to achieve specific goals, linked to effectiveness, efficiency, and satisfaction, within a particular context of use. The results of tests obtained using the proposed model in comparison to traditional models will be evaluated using the ISO 9241 standard, which pertains to interface usability. The dimensions to be evaluated include user-interface interaction, effectiveness, efficiency, and satisfaction.

A. Choice and Measurement of Variables

The following metrics are used:

1) Effectiveness: The accuracy achieved by users in relation to specified objectives also implies the absence of system errors and ease of learning and remembering. The metrics used include the number of important tasks executed, the number of relevant functions used, the number of tasks completed on the first attempt, the number of support requests, the number of accesses to help, and the number of functions learned.

2) Efficiency: This is the relationship of the resources employed (time, effort, etc.) with the accuracy and integrity with which the users achieve the specified objectives. The metrics used are: Time spent in the first attempt, time spent in

relearning functions, productive time, time to learn characteristics, time to relearn, time spent in error correction.

3) Satisfaction: it is a subjective factor that takes into account the positive attitude towards the use of the product. The metrics used are: qualification of satisfaction with important features, qualification of ease of learning, qualification of error handling and the rate of voluntary system usage.

4) Preparation of Experiments with users: Once the devices are interconnected and the software modules are initiated, the evaluation of the document scanning model for explicit knowledge databases using the VR and NLP model, and the model with traditional interfaces of ALICIA (Web) were conducted. These evaluations were performed with five users, following Nielsen's [43] recommendations for interface evaluation. Each user was assigned the task of searching for scientific articles for thirty minutes using both interface models. During the experiment, users could request assistance from an expert, who was also responsible for documenting the experiment.

The first model evaluated was the exploration of scientific articles using the VR, NLP, and NUI-based model. In this model, the user immersed in the interface communicates with the article search mechanism through voice commands and can also interact with the elements being queried through touch. Fig. 9 shows a user utilizing the interface model, and Fig. 10 displays the information content presented to the user on the video monitor.

Fig. 9. Evaluation of the VR, NLP and NUI interface model.

Fig. 10. View of the user with the model.

The second model to be evaluated is the traditional document exploration model, utilizing the Alicia scientific database interface. In this case, the user has a monitor for query visualization and uses the mouse and keyboard solely for interactions. Fig. 11 illustrates a user utilizing the traditional Alicia-Web interface, and what the user sees on the monitor is shown in Fig. 12.

Fig. 11. Evaluation of the traditional model of interface - Alicia Web.

Fig. 12. View of the monitor with the information shown by the traditional interface.

VIII. Results

The observations made on users using both types of interfaces allowed for the assessment of the effectiveness of each interface using the metrics outlined in ISO 9241. Effectiveness is defined as the achievement of goals or task completion, and it is represented by the number of goals accomplished. Various metrics count the goals achieved, tasks completed, functions used, etc. The results are tabulated in Table I.

The data from the results of using both interfaces for evaluating their effectiveness according to ISO 9241 are presented. Effectiveness is defined as the optimal use of the necessary resources to achieve a goal, where time is an important factor for the user of a software interface. The longer the time taken by the user to complete tasks using the interface, the lower the effectiveness. Conversely, if the time is shorter, the interface is considered more effective. The results are shown in Table I.

The satisfaction evaluation for each interface, considering the metrics defined by ISO 9241, takes into account that it is the degree or level of conformity of the user; it is their personal appreciation when evaluating the software interface. In order to measure software satisfaction in different aspects, the user must choose from five assessment alternatives: (1) Not at all satisfied, (2) Not very satisfied, (3) Fairly satisfied, (4) Very satisfied, (5) Completely satisfied. After each alternative of satisfaction there is a percentage of valuation corresponding to each alternative of satisfaction; these are respectively: 0.2, 0.4, 0.6, 0.8, 1.0. By averaging the levels of satisfaction, we can ultimately obtain the percentage of overall satisfaction required

TABLE I. COMPARISON OF EFFECTIVENESS USING THE TWO INTERFACES [15]

Metrics: Effectiveness Measure - ISO 9241	**Proposed model (Mean Quantity)**	**Web (Mean Quantity)**
Tasks performed	2.0	4.2
Functions used	3.8	7.8
Completed successfully Tasks at the first attempt	1.2	2.2
Important tasks performed	2.2	2.8
Calls for support	3.4	0.6
Access to help	0.4	0.2
Metrics: Measure of efficiency - ISO 9241	**Proposed model (Mean Quantity)**	**Web (Mean Cuantity)**
First attempt time	35.6	47.3
Time to relearn functions	31.2	16.6
Productive time	54.0	78.6
Learn characteristics time	16.6	31.2
Relearn characteristics time	54.0	78.6
Error correction time	55.8	54.6
Metrics: Measure of satisfaction - ISO 9241	**Proposed model (Mean %)**	**Web (Mean %)**
Qualification of satisfaction	0.72	0.32
Qualification of ease of learning	0.64	0.44
Error treatment qualification	0.24	0.28
Voluntary product use rate	0.92	0.64

in each metric.

IX. DISCUSSION

After processing the data obtained in the evaluation of the interfaces for each of the metrics following the ISO 9241 standards, allowing the assessment of the system in the dimensions of effectiveness, efficiency, and satisfaction, the "T-Student" statistical test was applied. This test revealed sufficient evidence that significant differences exist in the comparison of the evaluated dimensions:

- The proposed VR and NLP-based model showed low effectiveness compared to the traditional model of the scientific database ALICIA.
- The VR and NLP-based model exhibited higher efficiency compared to the traditional ALICIA database.
- Users expressed greater satisfaction with the VR and NLP-based model.
- The efficiency of the VR and NLP-based model was higher compared to the traditional ALICIA model.

The results obtained from the empirical analysis, based on user experiences with the VR and NLP-based model, generated considerable interest among users. They perceive it as a novel approach to interacting with explicit documents in institutional repositories, a proposal with significant development potential. Users faced initial difficulty with both types of models. Specifically, users who had never used ALICIA only utilized basic functions and not the tools and features available in the system. In contrast, users utilizing the VR and NLP-based interface initially struggled to adapt to the three-dimensional nature of immersion headsets. However, after a few minutes of adaptation, they expressed a desire to explore all available options and functionalities.

For exploration and document retrieval tasks, users employing the VR and NLP-based interface completed these tasks in less time than those using traditional web-based interfaces. It is noteworthy that a significant issue with VR immersion headsets is user discomfort.

X. CONCLUSION

An interface system model for explicit knowledge databases has been proposed and developed, based on natural language, natural users interface and immersive technologies processing techniques, which were successfully tested and found to be successful.

The scientific database ALICIA, was used for the evaluation of the proposed traditional model, using the ISO 9241 standard, considering Effectiveness, Efficiency and Satisfaction. The statistical test T-Student was applied to evaluate the data obtained by means of inferential statistics. The results allow to establish that the proposed model based on VR, NLP and NUI is superior to the traditional one in the dimensions of "Efficiency and Satisfaction" and is inferior in the dimension of "Effectiveness". Therefore, it is concluded that the model of exploration of scientific articles with VR, NLP and NUI is better than the traditional model.

It is important to point out that currently, users of large databases have traditional interfaces based mainly on the mouse and keyboard. In the opinion of the authors, a greater effort should be made to research new interface models that are more efficient and effective and that have ergonomic and usability features to facilitate user interaction with the systems.

ACKNOWLEDGMENT

The authors would like to thank both the Universidad Nacional de San Agustín (UNSA), Arequipa – Perú, and the Universidad Nacional Mayor de San Marcos (UNMSM), Lima - Perú.

REFERENCES

[1] N. Dmytrenko, *Knowledge management in Erasmus + mobility projects teams*. (Master Thesis), University of Gothemburg, Sweden, 2015.

[2] M. Franco and F. Di Virgilio and L. Di Pietro, *Management of Group Knowledge and the Role of E-WOM for Business Organizations*. In Knowledge Management and Competitive Advantage: Issues and Potential Solutions, IGI Global, 2014, pp. 71-89.

[3] Y. Wang, *Using Machine Learning and Natural Language Processing to Analyze Library Chat Reference Transcripts*. Information Technology and Libraries, vol. 41, no. 3, 2022.

[4] G. Vladova and A. Ullrich and J. Bahrs and B. Bender, *Digitalisation and Enterprise Knowledge (net)Working*. Proceedings of the 51st Hawaii International Conference on System Sciences, 2018.

[5] R. Linares and J. Herrera and L. Alfaro, *AliciaVR: Exploration of scientific articles in an immersive virtual environment with natural user interfaces*. 2016 IEEE Ecuador Technical Chapters Meeting (ETCM), Guayaquil, 2016.

[6] M. Assante and L. Candela and D. Castelli and A. Tani, *Are Scientific Data Repositories Coping with Research Data Publishing?*. Data Science Journal, vol. 15, no. 6, 2016.

[7] T. Sakai and B. Flanagan and J. Zeng and T. Nakatoh and S. Hirokawa, *Search Engine Focused on Multiple Features of Scientific Articles*. 2012 IIAI International Conference on Advanced Applied Informatics, Fukuoka, Japan, 2012, pp. 214-217.

[8] S. Singh and S. Gupta and D. Busso and S. Kamboj, *Top management knowledge value, knowledge sharing practices, open innovation and organizational performance*. Journal of Business Research, vol. 128, pp. 788–798, 2021.

[9] A. Attia and I. Eldin, *Organizational learning, knowledge management capability and supply chain management practices in the Saudi food industry*. Journal of Knowledge Management, vol. 22, pp. 1217–1242, 2018.

[10] A. Mohamad and T. Ramayah and L. May, *Sustainable knowledge management and firm innovativeness: The contingent role of innovative culture*. Sustainability, vol. 12, no. 17, pp. 6910, 2020.

[11] M. Demuner-Flores and R. Nava-Rogel, *Gestión del Conocimiento al Interior de las Instituciones de Educación Superior*. GECONTEC: Revista Internacional de Gestión del Conocimiento y la Tecnología, vol. 6, no. 1, pp. 68-81, 2018.

[12] E. Hajric, *Knowledge Management System and Practices. A Theoretical and Practical Guide for Knowledge Management in Your Organization*. New York Press, 2018.

[13] C. Archer-Brown and J. Kietzmann, *Strategic knowledge management and enterprise social media*. Journal of Knowledge Management, vol. 22, no. 6, pp. 1288-1309, 2018, doi:10.1108/jkm-08-2017-0359

[14] P. Federico and M. Wagner and A. Rind and A. Amor-Amoros and S. Miksch and W. Aigner, *The Role of Explicit Knowledge: A Conceptual Model of Knowledge-Assisted Visual Analytics*. 2017 IEEE Conference on Visual Analytics Science and Technology (VAST), 2017.

[15] L. Alfaro and R. Linares and J. Herrera, *Scientific Articles Exploration System Model based in Immersive Virtual Reality and Natural Language Processing Techniques*. International Journal of Advanced Computer Science and Applications, vol. 9, no. 7, pp. 254-263, 2018, doi:10.14569/ijacsa.2018.090736

[16] E. Duperet and D. Pérez and M. Rodríguez and A. Ramírez and L. Montoya, *Importance of repositories for preserving and recovering information*. Medisan, vol. 19, no. 10, pp. 1283-1290, 2015.

[17] P. Jackson and I. Moulinier, *Natural Language Processing for Online Applications*. Text retrieval, extraction and categorization, vol. 5, 2007.

[18] D. Lewis and K. Jones, *Natural language processing for information retrieval*. Communications of the ACM, vol. 39, no. 1, pp. 92-101, 1996, doi:10.1145/234173.234210

[19] J. Gratch and J. Rickel and J. Cassell and E. Petajan, *Creating Interactive Virtual Humans: Some Assembly Required*. IEEE Intelligent Systems, vol. 17, no. 4, pp. 54-63, 2002.

[20] B. Xu and R. Cai and Z. Zhang and X. Yang and Z. Hao and Z. Li and Z. Liang, *NADAQ: Natural Language Database Querying Based on Deep Learning*. IEEE Access, vol. 7, pp. 35012-35017, 2019, doi:10.1109/access.2019.2904720

[21] F. Brad and R. Iacob and I. Hosu and T. Rebedea, *Data set for a Neural Natural Language Interface for Databases (NNLIDB)*. 8th International Joint Conference on Natural Language Processing, Taipei, Taiwan, 2017.

[22] N. Nihalani and S. Silakari and M. Motwani, *Natural language Interface for Database: A Brief review*. International Journal of Computer Science Issues, vol. 8, no. 2, 2011.

[23] A. Bhat and G. Bhagwat and J. Chavan, *A Survey on Virtual reality platform and its Applications*. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 4, no. 10, pp. 3775-3778, 2015.

[24] J. Martín-Gutiérrez and C. Mora and B. Añorbe-Díaz and A. González-Marrero, *Virtual technologies trends in education*. EURASIA Journal of Mathematics Science and Technology Education, vol. 13, no. 2, pp. 469-486, 2017.

[25] L. Alfaro and C. Rivera and J. Luna-Urquizo and E. Castañeda and J. Zuñiga-Cueva and M. Rivera-Chavez, *New Trends in e-Technologies and e-Learning*. 2021 IEEE World Conference on Engineering Education (EDUNINE), 2021.

[26] L. Alfaro and C. Rivera and J. Luna-Urquizo and S. Alfaro and F. Fialho, *Knowledge construction by immersion in virtual reality environments*. International Journal of Advanced Computer Sciences and Applications, vol. 10, no. 12, 2019, doi: 10.14569/ijacsa.2019.0101278

[27] D. Liu and K. Kumar and Y. Gao and T. Chang and R. Huang, *The Potentials and Trends of Virtual Reality in Education*. Springer Singapore, 2017, pp. 105–130.

[28] A. Alqahtani and L. Daghestani and L. Ibrahim, *Environments and system types of virtual reality technology in STEM: A survey*. International Journal of Advanced Computer Science and Applications (IJACSA), vol. 8, no. 6, pp. 77-89, 2017.

[29] J. Psotka, *Immersive training systems: Virtual reality and education and training*. Instructional Science, vol. 23, no. 5-6, pp. 405-431, 1995, doi:10.1007/bf00896880

[30] E. Lombardo and S. Agostinelli and M. Metge, *Could an interactive and total immersive device with hmd improve memory and give the presence sensation?*. Issues in Information Systems, vol. 14, no. 1, pp. 315-321, 2013.

[31] S. Mann, *Intelligent Image Processing*. John Wiley & Sons, Inc., 2001.

[32] A. Câmara, *Natural User Interfaces*. In Human-Computer Interaction – INTERACT 2011. Berlin: Springer, Berlin, Heidelberg, 2011.

[33] F. García-Peñalvo and L. Moreno, *Special issue on exploring new Natural User Experiences*. Universal Access in the Information Society, vol. 18, no. 1, pp. 1-2, 2019, doi:10.1007/s10209-017-0578-0

[34] O. Díaz-Rodríguez and M. Pérez, *Minería de intenciones a partir de una base del conocimiento y aprendizaje automático supervisado*. 3C TIC. Cuadernos de desarrollo aplicados a las TIC, vol. 10, no. 3, pp. 65-101, 2021.

[35] G. Khodabandelou and C. Hug and C. Salinesi, *A novel approach to process mining: Intentional process models discovery* International Conference on Research Challenges in Information Science, pp. 78-89, 2014, https://doi.org/10.1109/RCIS.2014.6861040

[36] C. Jiménez Noblejas and A. Perianes Rodríguez, *Recuperación y visualización de información en Web of Science y Scopus: una aproximación práctica*. Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información, vol. 28, no. 64, pp. 15-31, 2014, doi:10.1016/s0187-358x(14)70907-4

[37] M. Ghodrati and A. Morris and A. Price, *The (un) suitability of modern liquid crystal displays (LCDs) for vision research*. Frontiers in Psychology, vol. 6, no. 303, 2015, https://doi.org/10.3389/fpsyg.2015.00303

[38] A. Zunjic and L. Ristic and D. Milanovic, *Effects of screen filter on visibility of alphanumeric presentation on CRT and LCD monitors*. Work, vol. 41, pp. 3553-3559, 2012.

[39] J. Feiner and K. Andrews, *RepoVis: Visual Overviews and Full-Text Search in Software Repositories*. 2018 IEEE Working Conference on Software Visualization (VISSOFT), Madrid, Spain, 2018.

[40] A. Al-Amawi and S. Alsmarai and M. Maraqa, *Creating a Knowledge Database for Lectures of Faculty Members, Proposed E-Module for Isra University*. International Journal of Advanced Computer Science and Applications, vol. 6, no. 11, pp. 69-77, 2015.

[41] P. Jordan, *An Introduction To Usability*. CRC Press, 1998.

[42] W. Winn, *A conceptual basis for educational applications of virtual reality*. (Technical Report TR-93-9). Seattle, Washington: Human Interface Technology Laboratory, University of Washington.

[43] J. Nielsen, *Designing web usability: The practice of simplicity*. New Riders Publishing, 1999.

CESSO-HCRNN: A Hybrid CRNN With Chaotic Enriched SSO-based Improved Information Gain to Detect Zero-Day Attacks

Dharani Kanta Roy
Department of Computer Science and Engineering,
Central Institute of Technology Kokrajhar
Kokrajhar, India

Ripon Patgiri
Department of Computer Science and Engineering
National Institute of Technology Silchar
Silchar, India

***Abstract*—Hackers use the vulnerability before programmers have a chance to fix it, which is known as a zero-day attack. Zero-day attackers have a variety of abilities, including the ability to alter files, control machines, steal data, and install malware or adware. When a series of complex assaults uses one or more zero-day exploits, the result is a zero-day attack path. Timely assessment of zero-day threats might be enabled by early detection of zero-day attack pathways. To detect this zero-day attack, this paper introduced a Chaotic Enriched Salp Swarm Optimization (CESSO) with the help of a hybrid Convolutional Recursive Neural Network (HCRNN) is implemented. The input data is retrieved from two datasets called IDS 2018 Intrusion CSVs (CSE-CIC-IDS2018) and NSL-KDD. The data is pre-processed with the help of data cleaning and normalization. A unique hybrid feature selection method that is based on the CESSO and Information Gain(IG) is introduced. The CESSO is also used to improve the Recursive Neural Network (RNN) performance to produce an optimized RNN. The selected features are classified, and prediction is performed using the hybrid Convolutional Neural Network (CNN) with RNN called HCRNN. The implementation of the zero-day attack is performed using MATLAB software. The accuracy achieved for dataset 1 is 98.36%, and for dataset 2 is 97.14%.**

***Keywords*—Hackers; vulnerability; zero-day attack; chaotic enriched salp swarm optimization; data cleaning; normalization; and MATLAB software.**

I. Introduction

An online assault that targets a software vulnerability that neither the programmed developer nor antivirus vendors are aware of is known as a zero-day (0-day) exploit. Before anybody else interested in resolving the issue can identify the software vulnerability, the attacker does, writes an exploit rapidly and then utilizes it to launch an attack [1], [2]. Particularly in today's environment, a zero-day vulnerability poses a major and possible threat to many enterprises. When a software flaw is a zero-day vulnerability, it goes beyond the expected immediate or almost daily instantaneous detection and remains unnoticed for a considerable amount of time [3]. Before the manufacturer even notices the issue or rushes to patch it, this security flaw is frequently exploited by hackers [4], [5]. This exploit is, hence, often referred to as a zero-day assault. Zero-day attacks might evade detection by traditional defenses for a considerable amount of time since network managers' primary focus is to stop assaults before they happen [6]. This makes dangerous agents more likely to penetrate the networks and makes the administrator's job more difficult. Preventing zero-day attacks is one of the most challenging aspects of risk management.

Risk management requires knowledge of the potential threats to be faced as well as the strength of the attack surface [7]. Zero-day attacks make use of vulnerabilities that have already been identified but of which the vendor or developer of the application is ignorant. The vulnerability's recent discovery may take weeks for identification and patching, leaving the program vulnerable to exploits. All kinds of intrusion detection systems are thought to face their greatest threat from zero-day (unknown) assaults [8], [9]. Various studies have dealt with the challenge of identifying unidentified assaults. Applying unsupervised anomaly detection algorithms is one way to find new attack kinds. Despite how difficult this issue is, resolving it would significantly improve the security of our computer system [10]. When additional safeguards are put in place to identify zero-day attacks, the difficulties listed below must be addressed. Not all zero-day vulnerabilities result in zero-day attacks [11].

The players without adequate security measures to thwart all potential attacks lose and are taken advantage of [12]. These weaknesses in security mechanism implementation by an organization or a person allow hackers an exploitation window that might result in zero-day attacks. Zero-day vulnerabilities can be challenging to identify since they can manifest in a variety of ways, including a lack of data encryption, a lack of authorizations, a weakness in an algorithm, a vulnerability in the password security system, etc. [13], [14]. Due to the nature of these security flaws, detailed information on zero-day exploits may only be accessed after the exploit has been discovered [15]. An enterprise may notice suspicious scanning activity or unexpected traffic originating from a customer or service due to a zero-day vulnerability.

The foremost contribution of the paper is as follows-

- The filter technique IG is used before initializing the population in the SSO model. Then, the CESSO approach is used to choose the optimum feature subset.
- Salp Swarm Optimization is improved by incorporating the Chimp Optimization algorithm's strategy,

i.e., its Chaotic Map, named Chaotic Enriched SSO (CESSO).

- This CESSO model is used to improve the performance of the RNN to produce an optimized RNN.
- The optimized RNN is a hybrid with the CNN called HCRNN, which will enhance the prediction accuracy of the zero-day attack.

The organization of the paper is as follows: Section II explains the literature review of the paper, Section III describes the problem statement, Section IV is the detailed description of the proposed methodology, Section V discusses the result and discussion, and finally, Section VI concludes the paper with a detailed conclusion.

II. Literature Review

In this section, the recent papers related to the zero-day attack are discussed, and their drawbacks are also discussed.

In 2016, Zhang et al. [16] suggested a security metric for assessing network resilience to Zero-Day Attacks. By creating and assessing several diversity measures, this research takes the first step toward formally characterizing network diversity as a security parameter.

In 2018, Sun et al. [17] proposed the probabilistic identification of zero-day attack paths, and Bayesian networks were used. This research suggested a probabilistic method and developed a prototype system called ZePro for identifying zero-day attack paths.

In 2019, Afek et al. [18] suggested the extraction of zero-day signatures for high-volume attacks. The key contributions of this research are the method created to extract the necessary signatures, together with the concept of the string-heavy hitters' issue and the technique for solving it.

In 2021, Zoppi et al. [19] presented the strategy and application of unsupervised algorithms for detecting zero-day attacks. The article uses a question-and-answer format to highlight common problems encountered when performing quantitative studies for zero-day detection, and it demonstrates how to build up and test unsupervised algorithms using the proper equipment.

In 2022, Mbona and Eloff [20] recommended machine learning approaches for the semi-supervised detection of zero-day intrusion attacks. The method suggested in this paper shows that Benford's rule, the law of anomalous numbers, is a workable technique that may successfully discover major network aspects that are suggestive of abnormal behavior and can be utilized for identifying zero-day assaults.

In 2022, Popoola et al. [21] examined federated deep learning for IoT-edge devices to detect zero-day botnet attacks. To protect IoT edge devices from leaking personal data, this study presented the federated DL (FDL) approach for zero-day botnet attack detection. This approach uses an ideal deep neural network (DNN) model to classify network traffic.

In 2022, Bar and Hajaj [22] proposed the SimCSE for the prediction of encrypted traffic and zero-day attacks. The simple contrastive learning of sentence embedding (SimCSE) proposed in this study serves as the embedding model for a novel method for traffic identification at the packet level.

In 2021, Kumar and Sinha [23] proposed an effective method for detecting zero-day cyberattacks. This study suggests an innovative, robust, and intelligent cyber-attack detection model that makes use of the heavy-hitter idea and a graphical approach to identify zero-day assaults.

III. Problem Statement

Traditional intrusion detection approaches are no longer enough for identifying zero-day attacks, which may exploit security vulnerabilities that may exist in a network due to the fast expansion of network-based cyberattacks. With the advancement and new development of systems, cyber security vulnerabilities, and technology, existing approaches in intrusion detection have not been sufficient to defend systems from cyber security vulnerabilities. There is a need for more proactive ways to identify known and new vulnerabilities to stop zero-day attacks from taking advantage of such network security weaknesses. If a hacker is successful in exploiting a vulnerability in software before the developers of the affected system can find a solution, this is known as a zero-day attack [24]. Zero-day vulnerabilities can take on practically any form since they can be disguised as any type of more generic software flaw.

IV. Proposed Methodology

Detecting zero-day threats is one of the most challenging responsibilities of an IDS. An attack that makes use of a network vulnerability that has not yet been found is known as a zero-day attack. The first step in the life of a zero-day assault is the identification of a software vulnerability. Once an exploit has been made, the vulnerability is utilized to attack the targets. After the initial assaults are carried out, the vendors distributing the weak software learn about it and produce a patch to fix it. This project will design a fresh, optimized deep learning system to address the difficulties of detecting zero-day ransomware. Fig. 1 shows the proposed approach for detecting zero-day attacks.

From Fig. 1, the input data is received from the two datasets called CSE-CIC-IDS2018 and NSL-KDD. Then, the input data is cleaned and normalized in the pre-processing stage. The statistical features and higher-order features are extracted in the feature extraction stage; the required features are selected using the improved IG using the CESSO model. Then, the same CESSO will enhance the features of the RNN, and it is hybrid with the CNN to produce HCRNN, which is used for zero-day attack detection.

A. Dataset Description

In the Zero-day attack model, two datasets are used.

1) IDS 2018 Intrusion CSVs (CSE-CIC-IDS2018): This dataset was created by the University of New Brunswick initially for research on DDoS data. This dataset won't be modified in the future because it was totally compiled in 2018. The dataset was indeed based on logfiles kept by the institution, which revealed several DoS assaults during the course of the period that were made accessible to the public. Because

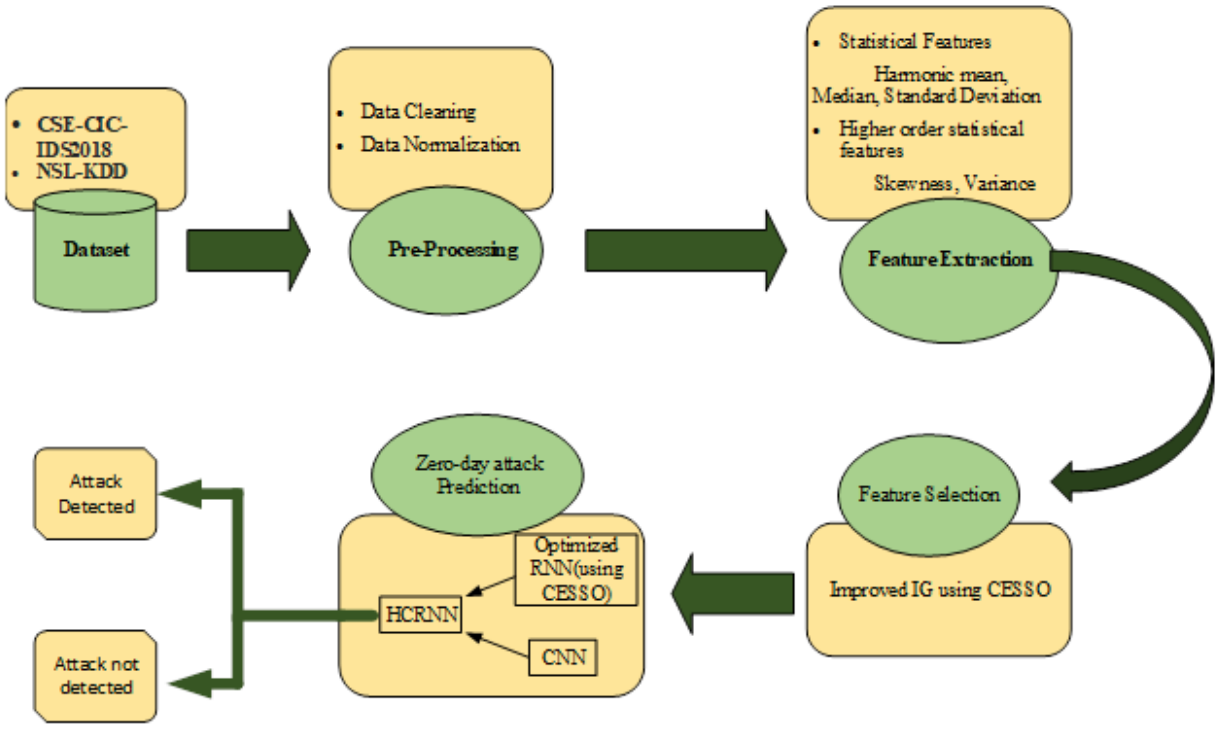


Fig. 1. Block diagram of the Zero attack detection mechanism.

it specifies whether or not the sent packets are malicious, the Label column is possibly the essential information for building machine learning algorithms for this dataset.

Based on date, data is separated into several files. Each file is imbalanced; the notebook creator is responsible for partitioning the dataset into a balanced shape for more accurate predictions. This dataset contains a total of 80 columns, each of which corresponds to a record in the IDS logging system used by the University of New Brunswick. Both columns for forward and backward traffic are present because their system distinguishes between the two. The following are the columns in this dataset that are most important: Label, Destination port, Protocol, Flow Duration, Total forward packets, Total backward packets, and Total forward packets [25].

2) NSL-KDD: As compared to the original KDD data set, the NSL-KDD data set is better in the following aspects:

The classifiers won't be skewed in favor of a more prominent impact since similar data are excluded from the train set. Due to the lack of duplicate data in the proposed test sets, methods with greater diagnostic accuracy for frequent records have no impact on the learners' outcomes. The proportion of records selected from each class of difficulty levels in the original KDD data set is inversely proportional to the total number of entries in each class.

As a result, there is a wider range of variance in the classification rates of different machine learning techniques, making it simpler to assess different learning approaches correctly. With a manageable number of records in both the train and test sets, it is possible to run the tests on the complete set without having to choose a subset at random. Therefore, the assessment results from different research initiatives will be comparable and consistent [26].

B. Pre-processing

The raw dataset gathered during the prior step must be cleaned and normalized during this stage. It is possible to access the dataset's raw log data set. To assist in identifying online assaults, the data is normalized, encrypted, and stripped of duplicate and missing information during this phase of the evaluation process.

1) Data Cleaning: Finding duplicate data is a common need of data cleaning operations. Duplicates should be eliminated since deep learning produces the same patterns when non-duplicate variables are used. The cleaned dataset has been obtained after some columns have been cleared of duplicate and missing values. This dataset contains no duplicates.

2) Normalization: By transforming them into unique patterns, data are normalized. The process of converting raw data into a format that can be handled effectively is known as data transformation. Normalization focuses mostly on reducing or even getting rid of redundant data. Because of this crucial problem, managing data in relational databases that store similar data in several places is getting harder.

C. Feature Extraction

The features such as statistical features (Harmonic mean, median, standard deviation), higher order statistical features

(skewness, Variance), Improved Information gain (Proposed), and based features will then be retrieved from the pre-processed data.

1) Statistical Features:

a) Harmonic Mean: Numerical averages include the harmonic mean. It is determined by multiplying the number of observations, or series elements, by the inverse of each integer in the series. The harmonic mean is, therefore, the inverse of the arithmetic mean of the reciprocals, as shown in Eq. (1)

$$\mu_{\mathcal{H}} = \frac{T}{\sum_{i=1}^{n} \frac{1}{\alpha_i}} \tag{1}$$

where, α_i a series with T numbers.

b) Median: The median, which is the midpoint in a list of numbers arranged either ascending or descending, may be more representative of the set of data than the average.

$$Median = \begin{cases} \frac{l+1}{2}, & for\ odd\ numbers \\ \frac{l}{2}, & for\ even\ numbers \end{cases} \tag{2}$$

where, l is the last number.

c) Standard deviation: The term "standard deviation" (σ) refers to a measurement of the data's dispersion from the mean. The values are centered all around the mean whenever the standard deviation is small and widely scattered when it is high.

$$SD(\sigma) = \sqrt{\frac{\sum_{i=1}^{n}(\alpha_i - \mu)^2}{N}} \tag{3}$$

where α_i is the input value, μ is the mean and N is the total number of elements.

2) Higher-order statistical features:

a) Skewness: A distribution's skewness can be measured to see how asymmetrical it is. A distribution is asymmetric whenever its left and right sides do not reflect each other in the same way. A distribution may have right (or positive), left (or negative), or zero skewness.

$$Skewness = \frac{3(\mu - Median)}{\sigma} \tag{4}$$

b) Variance: Variance is the analytical evaluation of the numerical variation inside a data set. More specifically, variance determines how far apart every integer in the set is from the mean and, thus, from the other numbers in the set.

$$Variance = \frac{\sum_{i=1}^{n}(\alpha_i - \mu)^2}{N} \tag{5}$$

D. Feature Selection

This section introduces a unique hybrid feature selection method that is based on the CESSO algorithm and IG. Let's start by outlining the IG algorithm and then describe how the CESSO technique was implemented in detail.

E. Improved Information Gain

Information gain uses correlations between features and classifications to distribute feature weights statistically. Let $F = \{f_1,\ f_2,\ f_3, \ldots, f_n\}$ be the group of n data points in a dataset, $X = \{x_1,\ x_2,\ x_3, \ldots, x_p\}$ be the collection of p features, and $Y = \{y_1,\ y_2,\ y_3, \ldots, y_m\}$ be a group of m type label. The number $P(Cl_i)$ denotes the percentage of classes in F with the label P_i where $i = 1,\ 2,\ 3, ..., m$. Eq. (6) provides the dataset's entropy.

$$H(Y) = -\sum_{i=1}^{m} P(Cl_i)\ log_2\ P(Cl_i) \tag{6}$$

Each feature in the data classification system has an information gain (IG). Where $Q = \{Q_j^1,\ Q_j^2, , \ldots, Q_j^p\}$ represents the $p^{th}(p = 1, 2, \ldots, n)$ a data set feature. Regarding the characteristic, $Q = \{Q_j^1,\ Q_j^2, , \ldots, Q_j^p\}$ the corresponding conditional entropy is:

$$H(Y|Q_j) = -\sum_{p=1}^{k} P(Q_j^p) \sum_{i=1}^{m} P(Cl_i|Q_j^p) log_2 P(Cl_i|Q_j^p) \tag{7}$$

where, $P(Q_j^p)$ denotes the categorical variable Cl's prior probability and Q_j^p is the value of Q_j, an attribute with k different kinds of values. The $P(Cl_i|Q_j^p)$ represents the conditional probability of the variable Cl after the attribute Q_j is fixed. Consequently, the formula below states that the value of the information gained from the attribute Q_j is provided by the difference between $H(Y), H(Y|Q_j)$, which is given by Eq. (8).

$$IG(Q_j) = H(Y) - H(Y|Q_j) \tag{8}$$

An increased IG often denotes the importance of the feature for the categorization.

F. IG-CESSO for Optimal Feature Subset Selection

The CESSO algorithm finds the optimal feature subset with the fewest features and highest classification accuracy, which is what feature selection seeks to do for a given dataset. It is utilized for giving each component a binary string to represent whether an attribute is picked; for example, the feature is chosen when the binary value is 1, whereas a binary 0 denotes that it is not.

Each of these two indications uniquely impacts the classifier's classification performance. In this instance, just merge them into a single weighted indicator and use the same fitness function as in Eq. (9).

$$F_i = \delta_1 * A_{clasifier} + \delta_2 * (1 - \frac{N}{T}) \tag{9}$$

where T stands for the total number of characteristics and N for the number of attributes that were selected; here, δ_1 and δ_2 have values of 1 and 0.001, respectively[27]. Eq. (10) may be used to determine the classification accuracy derived from the HCRNN classifier.

$$A_{classifier} = \frac{K_c}{K_c + K_i} \tag{10}$$

In this example, the numbers K_i and K_c stand for the cases that were classified wrongly and properly, respectively. With the help of the fitness value, it is possible to ensure that the chosen features have a limited amount of features while yet providing the highest possible classification accuracy[27]. The following section provides a detailed description of the CESSO algorithm with basic Salp Swarm Optimization (SSO)

1) Salp Swarm Optimization: The Salp Swarm Optimization (SSO) imitates the behavior of salps, a kind of planktonic tunicate belonging to the Salpidae family with a barrel-like structure. Additionally, their tissues resemble those of jellyfish, and they move similarly to jellyfish, and a large portion of their weight is made up of water. Pumping water through their jello-like bodies causes them to move by contracting, which alters their locations. Salps in the seas engage in a swarm activity known as the salp chain, which may aid salps in better mobility by employing rapid, coordinated shifts and feeding. Based on this behavior, they created a mathematical model of the salp chains and evaluated it using optimization issues. The leader and the followers are the two groups that are initially separated out in SSO [28]. The leading Salp in a chain is referred to as the leader, while the trailing Salps are referred to as the followers. In n-dimensions, where n stands for the variables in the issue, and n denotes the search space, the salps' location is established. These salps look for food sources since they can identify the swarm's goal. The position should be updated often. Thus, the salp leader updates the position using Eq. (11) below-

$$a_j^1 = \begin{cases} S_j + K_1((U_j - L_j) * K_2 + L_j), & K_3 <= 0 \\ S_j - K_1((U_j - L_j) * K_2 + L_j), & K_3 > 0 \end{cases} \quad (11)$$

where, a_j^1 is the position of the leader within j^{th} dimension, where the food source in this dimension is S_j, the upper and the lower bounds are U_j and L_j , respectively. The K_2 and K_3 are generated randomly in the range [0, 1] to maintain the search space. Additionally, the parameter K_1 is a crucial coefficient in this method since it helps to balance the exploration and exploitation phases [29]. It is derived as follows:

$$K_1 = 2e^{-(\frac{v}{v_{max}})^2} \quad (12)$$

where, the letters v and v_{max} stand for the current iteration and the maximum number of iterations, respectively. Using Eq. (13), the SSA begins updating the followers' position after changing the leader's position.:

$$a_j^i = \frac{1}{2}(a_j^i + a_j^{i-1}) \quad (13)$$

where a_j^i is the i^{th} position of the follower within j^{th} dimension and $i > 1$.[28]

However, the SSO has the limitation of falling into the local optima problem. Therefore, the proposed work has introduced a new SSO variant based on the Chimp Optimization Algorithm, i.e., the Chaotic Map. Therefore, chaotic enriched SSO provides high-level efficiency in optimization performance. The following shows a detailed description of the CESSO algorithm.

2) CESSO-based on COA Algorithm: Salp Swarm Optimization is improved by incorporating the strategy of the COA. The COA enhances the performance of the SSO due to the incorporation of the chaotic map strategy. The COA is mainly focusing on hunting its prey based on determining the prey and the chimp. In the CESSO algorithm, this step is reformulated by considering the current best solution, and the next solution is respectively represented as a_j^i and $a_j^(i-1)$. The mathematical derivation of the Chaotic map based on the COA [30] is shown in Eq. (14) and Eq. (15).

$$d = |K.a_j^i - C_v.a_j^{i-1}| \quad (14)$$

$$a_j^i(t+1) = a_j^i - p.d \quad (15)$$

where t denotes the current iteration, p, C_v, and c are the coefficient vectors, $a_j^{i-1}$is the position vector of the targeted solution, and a_j^i is the position vector of a current best solution. The coefficient vectors p, C_v, and K are determined using Eq. (16).

$$\begin{cases} p = 2.f.r_1 - f \\ K = 2.r_2 \\ C_v = Chaoticvalue \end{cases} \quad (16)$$

In the aforementioned equation, the variable f has a specific range, i.e., drastically minimizing from 2.5 to 0 in every iteration process [30]. Therefore, exploitation and exploration have maintained the equilibrium. On the other hand, the random vectors are set between the range of 0 and 1. At last, the chaotic vector is illustrated as m, i.e., a strategy of the COA. This strategy is executed based on executing several chaotic map processes. This strategy can enrich the optimization performance, and the chaotic enriched chimp-based salp optimization can provide higher performance than the traditional SSO and the COA. This hybrid model, Chaotic Enriched SSO (CESSO), can provide better performance.

The proposed work uses the CESSO algorithm to improve the feature selection, especially on Information Gain (IG). The IG-CESSO provides a high-level feature that helps to improve the classification performance. Apart from feature selection, the CESSO algorithm is utilized for improving the prediction performance; therefore, the parameters of the RNN are optimized using CESSO, which is discussed in section IV.G.2.

G. Prediction of Zero-day Attack

The optimized hybrid classifier that makes the zero-day attack detection will be trained using the selected optimal features. The hybrid classifier will be a combination of optimized RNN and CNN. To enhance the prediction accuracy of the zero-day attack detection model, the activation function of RNN is fine-tuned via a new CESSO.

1) CNN: The suggested model's structure is discussed in this part, along with specifics on its underlying motives. The proposed model has developed based on a set of layers, including the Input layer, Convolutional Layer, Max Pooling, and Recurrent Layer or Output Layer. Here, the CNN architecture has the fully connected layer that is replaced by Recursive Neural Networks (RNN); therefore, the proposed model has concatenated the Convolutional and recurrent layers for detecting zero-day attacks. These layers are discussed in their respective sections as follows.

In the architecture of the HCRNN for Zero Day attack detection, the CNN executes the layer-wise feature extraction using convolution layers and pooling layers [31]. The shortcomings of conventional machine learning methods, which need to extract data features manually, are overcome by CNN because it does not call for preprocessing or reconstruction of the original data. Furthermore, the model's complexity is substantially reduced by its weight-sharing characteristics.

The CNN's structure is depicted in Fig. 2, and the kernel function transforms the input data, which must be in two-dimensional form. The hidden layer comprises several convolution layers and pooling layers, where the input of one layer becomes the output of the following layer. There are not many restrictions on the input data using this structure, and it can accurately retrieve the hierarchical expression of the input data.

a. Convolutional Layer The convolution layer—which includes several feature matrices—is used to extract the local characteristics of the input data. Each feature matrix is capable of parallel processing since it can be thought of as a plane, which can significantly reduce the number of free parameters. Additionally, because convolution kernels vary depending on the plane, it is possible to display the extracted features completely. The following Eq. (17) and Eq. (18) determine the size of the output feature map following the convolution procedure:

$$h_o = [\frac{h_i - f_i + p_c}{s_i} + 1] \tag{17}$$

$$w_o = [\frac{w_i - f_i + p_c}{s_i} + 1] \tag{18}$$

Where the resulting feature map's height is h_o, w_o is the feature map's width at the output side, the input image's height is represented as h_i, w_i is the representation of input image's width, the size of the filter is denoted as f_i, the convolution operation's padding is represented as p_c, and the convolution' stride is denoted as s_i.

b. Max Pooling

The typical pooling technique known as max pooling (MP) operates on the feature map side and picks the largest value from a small n×n patch of the input feature map. The output feature map is then created by MP by combining the chosen values. The following Eq. (19) and Eq. (20) may be used to determine the output feature map size following the pooling procedure:

$$h_o = [\frac{h_i - f_i}{s_i}] \tag{19}$$

$$w_o = [\frac{w_i - f_i}{s_i}] \tag{20}$$

where, f_i is the size of the pooling area, the output feature map's height is indicated by the symbol h_o, The output feature map's width is indicated by the prefixw_o, the width of the input feature map is indicated by the letter w, the height of the input feature map's is indicated by the symbol h_i, and the width of the input feature map is denoted as w_i.

2) Recurrent Layer For Prediction: In the HCRNN architecture, the RNN layer joints the CNN by replacing the fully connected layer for prediction. The reason behind the selection of the RNN layer is for selecting temporal features, whereas CNN is for dealing with spatial features. However, the RNN has a drawback in that phrase parsing can be complex and slow. Interestingly, different parse trees may exist for the same text. Additionally, labeling the training data for recursive neural networks takes more time and effort than building recurrent neural networks. It takes more time and effort to manually break down a statement into smaller parts than it does to give it a label. To overcome this drawback, the parameters of the RNN are optimized using the CESSO algorithm.

The trees are symbolized by the letter T, and N samples are indicating the leaves of a tree. The neural network receives concatenated data containing a tree's possible child nodes. A combined score and a parent feature are produced by the neural network, which are the two features combined after being learned by a nonlinear mapping. A parent is created mathematically by applying a non-linear projection on the concatenated of the child characteristics, that is

$$PA_{x,y} = \rho[W_i(ch_x + ch_y) + B_p] \tag{21}$$

Where $(ch_x + ch_y)$ is the result of concatenating the child features x and y, W_i is the neural network's weight matrix, B_p is its bias parameter, ρ is a function that may be continually changed, and $PA_(x, y)$ is the last parent attribute [32]. The sign for this work was the logistic sigmoid since its derivative may be stated in terms of the primary equation. The neural network's learned merging score is used to determine whether samples are suitable for merging on the tree, as well as if they are neighboring when appropriate. To produce a binary tree, combining just the samples with the top score is permitted. The parent feature's score is calculated using Eq. (22)

$$PA_s = W^k_{score} PA_{x,y} \tag{22}$$

While the tree's score can be increased K, it is possible to determine the parameters W, B_p, and W^K_{score} of the neural network,

$$\underset{k \in N}{\operatorname{argmax}}\, PA_s(RNN(\theta, x, k)) \tag{23}$$

where θ is the representation of neural network parameters to calculate a score PA_s, and T are any potential trees produced by the aforementioned merging procedure. Finally, the RNN provides the outcome with respect to zero-day attacks. Moreover, the prediction performance is further improved by using the CESSO algorithm. The CESSO algorithm utilizes the

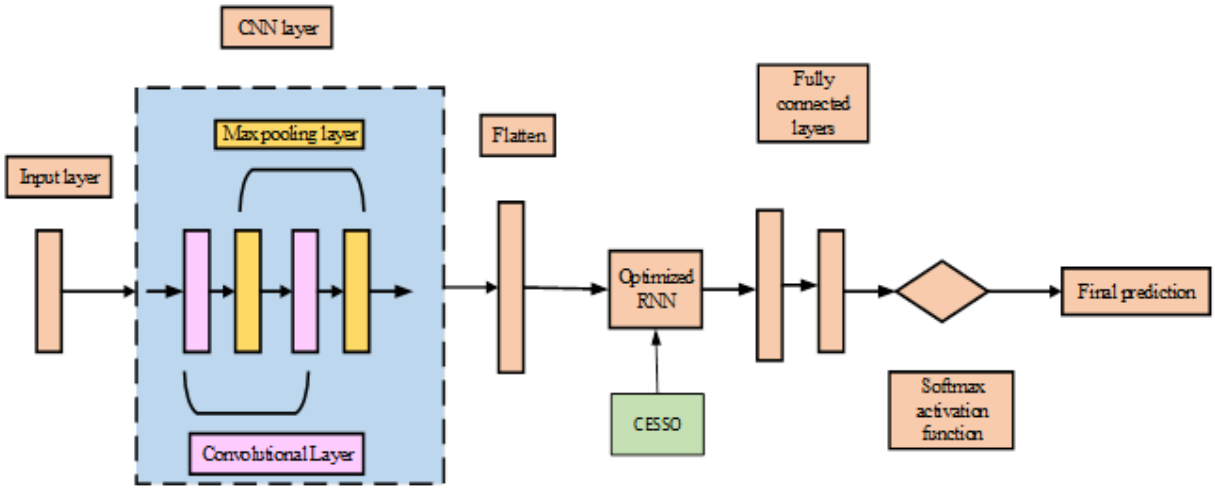


Fig. 2. Hybrid CNN with RNN architecture.

parameters of the RNN such as BATCH SIZE, MINI-BATCH, MOMENTUM, and LEARNING RATE. The prediction performance is further improved by this strategy. In this process, the CESSO algorithm evaluated the prediction performance of the RNN using the RMSE algorithm Eq. (24)

$$RMSE = \frac{1}{n}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2 \tag{24}$$

The n Number of samples, y_i Original value, y_i cap: Predicted outcome.

In the process, the RNN generates the value that is evaluated through the RMSE value. The minimal RMSE is considered the best solution. In each outcome, the parameter values are changed and updated with new solutions. Finally, the prediction performance of the RNN will be reached higher results. The overall flow of the proposed model is shown in Fig. 3.

V. Result and Discussion

The performance of the proposed strategy is compared to that of well-known methods such as CESSO-RNN, RNN, CNN, Bidirectional Long Short-Term Memory (Bi-LSTM), and Long Short-Term Memory (LSTM). The proposed CESSO-HCRNN model is compared with all existing models and performance for this is tabulated. The performance of the proposed is better than the other existing models because of improved IG using CESSO used in feature selection. Optimal RNN combined with CNN resultant the HCRNN also produces more accurate results. In this part, the performance measurements are also covered.

A. Performance Metrics

A number of matrices are used to measure the performance, including sensitivity, specificity, accuracy, recall, F-score, NPV, MCC, FPR, and FNR.

Sensitivity

The sensitivity value is obtained by just dividing the total positives by the proportion of true positive predictions

$$Sensitivity = \frac{TP}{TP + FN} \tag{25}$$

Specificity

Specificity is calculated by dividing the number of accurately anticipated negative outcomes by the total number of negatives

$$Specificity = \frac{TN}{TN + FP} \tag{26}$$

Accuracy The accuracy is the ratio of correctly classified data to all of the data in the log. The precision is described below-

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{27}$$

Precision By employing the entire number of samples used in the classification process, precision is the representation of the total number of genuine samples that are appropriately taken into consideration during the classification process.

$$Precision = \frac{TP}{TP + FP} \tag{28}$$

Recall

Recall rate is a measure of how many genuine samples overall are considered when categorizing data using all samples from the same categories from the training data.

$$Recall = \frac{TP}{TP + FN} \tag{29}$$

F- Measure

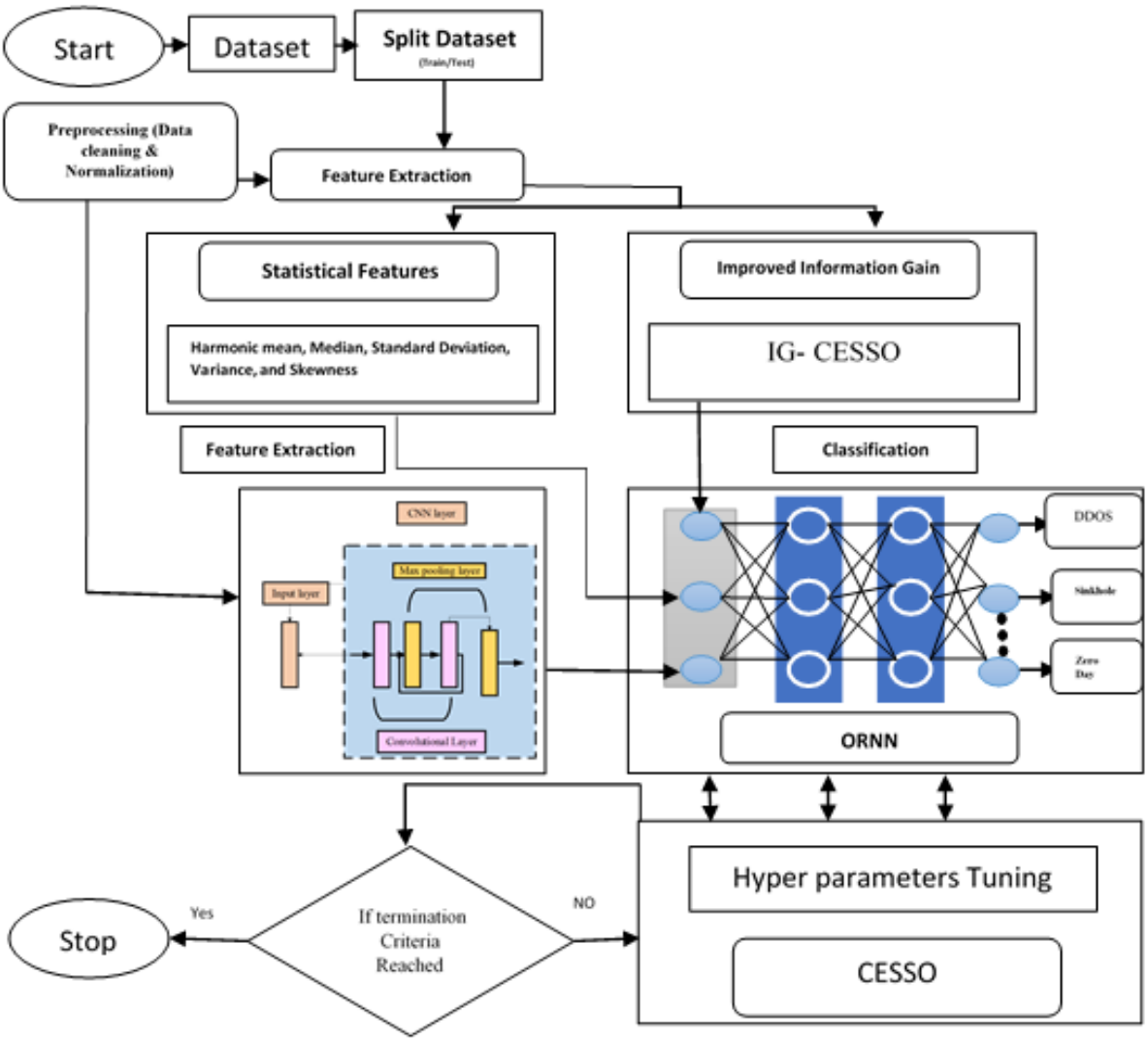


Fig. 3. Overall architecture of the proposed CESSO-HCRNN model.

The definition of the F-score is the harmonic mean of recall rate and accuracy.

$$F_{measure} = \frac{2precision * Recall}{Precision + Recall} \tag{30}$$

Negative Prediction Value (NPV)

NPV describes the effectiveness of a diagnostic test or other quantitative metrics.

$$NPV = \frac{TN}{TN + FN} \tag{31}$$

Matthews correlation coefficient (MCC) Below is a representation of the two-by-two binary variable association measure known as MCC.

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FN)(TN + FP)(TN + FN)(TP + FP)}} \tag{32}$$

False Positive Ratio (FPR) The number of negative occurrences divided by the number of negative events that were incorrectly classified as positive yields the false positive rate (false positives).

$$FPR = \frac{FP}{FP + TN} \tag{33}$$

False Negative Ratio (FNR) The likelihood that an actual positive may be overlooked by the test is known as the false-negative rate, sometimes referred to as the "miss rate".

$$FNR = \frac{FN}{FN + TP} \tag{34}$$

B. Comparative analysis for the performance metrics

1) CSE-CIC-IDS2018 Dataset: Using the CSE-CIC-IDS2018 Dataset, the performance metrics for the proposed CESSO-HCRNN approach were computed and contrasted with

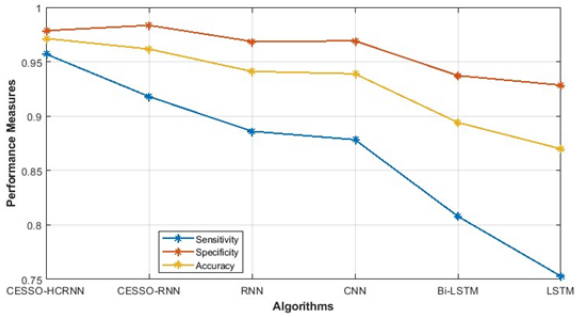


Fig. 4. Comparison of sensitivity, specificity, and accuracy of CSE-CIC-IDS2018 dataset for the proposed CESSO-HCRNN and existing techniques.

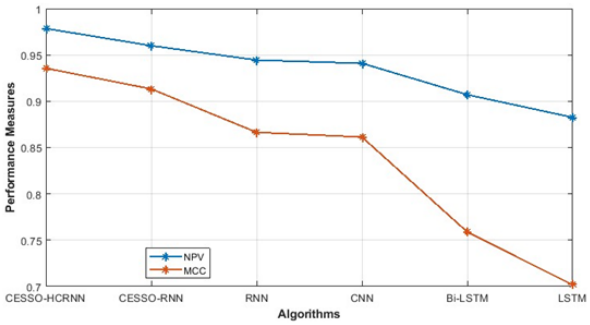


Fig. 6. Comparison of the NPV and MCC of CSE-CIC-IDS2018 dataset for the proposed CESSO-HCRNN and existing techniques.

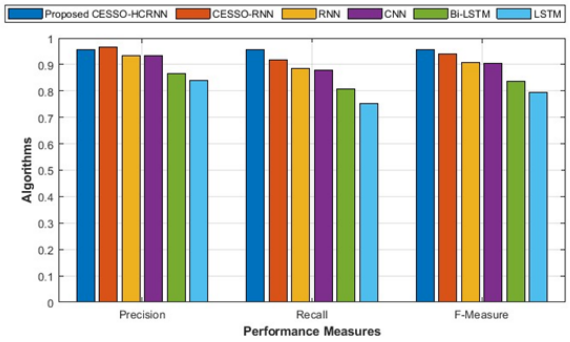


Fig. 5. Comparison of the precision, recall, and F- Measure of CSE-CIC-IDS2018 dataset for the proposed CESSO-HCRNN and existing techniques.

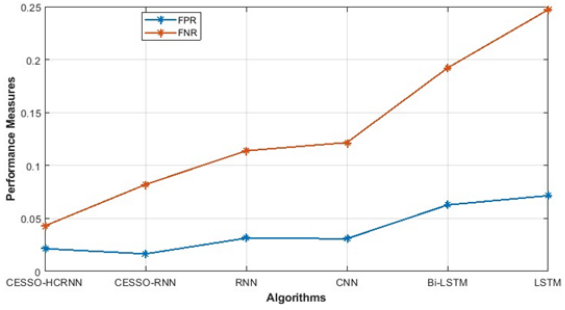


Fig. 7. Comparison of FPR and FNR of CSE-CIC-IDS2018 dataset for the proposed CESSO-HCRNN and existing techniques.

those of CESSO-RNN, RNN, CNN, Bi-LSTM, and LSTM. The suggested CESSO-HCRNN strategy's efficacy is evaluated here using the other metrics, whilst the proposed model's efficacy is evaluated here using measures like accuracy, sensitivity, precision, recall, and F-Measure. Table I compares the CSE-CIC-IDS2018 Dataset's performance metrics.

The metrics in the table are analyzed and their values are combined. Comparing the suggested CESSO-HCRNN model against older methods reveals that it is quite accurate. Graphs are used to show the difference in performance. The suggested and current models are contrasted using the graph in Fig. 4 in terms of metrics like sensitivity, specificity, and accuracy.

Fig. 4 displays performance metrics, such as sensitivity, specificity, and accuracy, for both the proposed and current approaches. Sensitivity values are 0.9571, 0.9180, 0.8861, 0.8784, 0.8079, and 0.7529 for the proposed CESSO-RNN, RNN, CNN, Bi-LSTM, and LSTM. The accuracy values are 0.9714, 0.9617, 0.9411, 0.9389, 0.8941, and 0.8699, respectively, while the specificity values are 0.9785, 0.9835, 0.9686, 0.9692, 0.9373, and 0.9285. The proposed model performs better in terms of sensitivity, specificity, and accuracy than the widely used methods.

Precision, recall, and F-Measure performance characteristics for the proposed CESSO-HCRNN and current techniques are shown in Fig. 5. Precision values for the proposed CESSO-RNN, RNN, CNN, Bi-LSTM, and LSTM are 0.9571, 0.9653, 0.9339, 0.9344, 0.8656, and 0.8403 respectively. Recall values for the proposed CESSO-HCRNN and current approaches are 0.9571, 0.9180, 0.8861, 0.8784, 0.8079, and 0.7529, respectively. The F-measure values are 0.9571, 0.9410, 0.9093, 0.9055, 0.8358, and 0.7942. The suggested CESSO-HCRNN model provides higher precision, recall, and F- Measure than the current approaches.

Fig. 6 illustrates the performance metrics, including NPV and MCC, for the proposed and existing techniques. The proposed CESSO-HCRNN and current models like CESSO-RNN, RNN, CNN, Bi-LSTM, and LSTM have NPV values of 0.9785, 0.9600, 0.9445, 0.9071, and 0.8826, respectively. The MCC values are 0.9356, 0.9133, 0.8664, 0.8614, 0.7588, and 0.7018 as well.

Fig. 7 visually represents the FPR and FNR performance indicators for both the proposed CESSO-HCRNN and the existing methods. The CESSO-HCRNN model has a lower FPR and FNR when compared to the approaches currently in use. The suggested CESSO- HCRNN and current approaches like CESSO-RNN, RNN, CNN, Bi-LSTM, and LSTM have the FPR and FNR are 0.0215, 0.0165, 0.0314, 0.0308, 0.0627, and 0.0715; and 0.0429, 0.0820, 0.1139, 0.1216, 0.1921, and 0.2471.

2) NSL- KDD Dataset: Performance metrics for the proposed CESSO-HCRNN approach are compared to those for

TABLE I. COMPARISON OF PERFORMANCE METRICS OF CSE-CIC-IDS2018 DATASET FOR THE PROPOSED CESSO-HCRNN AND EXISTING TECHNIQUES

Techniques	Sensitivity	Specificity	Accuracy	Precision	Recall	F-Measure	NPV	MCC	FPR	FNR
Proposed CESSO - HCRNN	0.9571	0.9785	0.9714	0.9571	0.9571	0.9571	0.9785	0.9356	0.0215	0.0429
CESSO-RNN	0.9180	0.9835	0.9617	0.9653	0.9180	0.9410	0.9600	0.9133	0.0165	0.0820
RNN	0.8861	0.9686	0.9411	0.9339	0.8861	0.9093	0.9445	0.8664	0.0314	0.1139
CNN	0.8784	0.9692	0.9389	0.9344	0.8784	0.9055	0.9410	0.8614	0.0308	0.1216
Bi-LSTM	0.8079	0.9373	0.8941	0.8656	0.8079	0.8358	0.9071	0.7588	0.0627	0.1921
LSTM	0.7529	0.9285	0.8699	0.8403	0.7529	0.7942	0.8826	0.7018	0.0715	0.2471

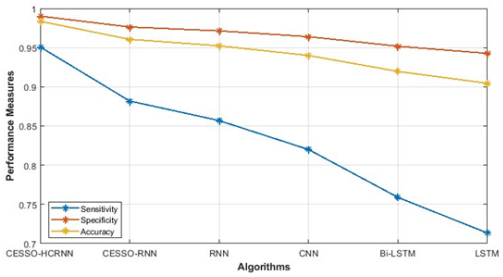


Fig. 8. Comparison of the sensitivity, specificity, and accuracy of NSL-KDD dataset for the proposed CESSO-HCRNN and existing techniques.

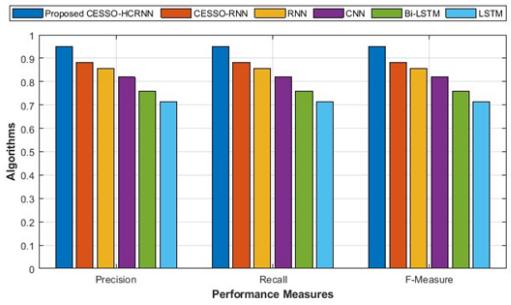


Fig. 9. Comparison of precision, recall, and F- Measure of NSL-KDD dataset for the proposed CESSO-HCRNN and existing techniques.

existing techniques like CESSO-RNN, RNN, CNN, Bi-LSTM, and LSTM. In this case, the loss values are evaluated using the FNR and FPR error measures, and the effectiveness of the suggested strategy and the chosen dataset for zero-day attack prediction are determined using the other metrics. A comparison of the performance metrics is shown in Table II.

From the table, the metrics values are calculated and contrasted. The proposed model is quite accurate when compared to previous approaches. To show the performance contrast, graphs are utilized. Utilizing the NSL-KDD dataset, the CESSO method is used to improve the prediction of zero-day attacks.

For both the proposed and current methodologies, Fig. 8 shows the graphical depiction of performance parameters including sensitivity, specificity, and accuracy. In comparison to previous methods, the CESSO-HCRNN model has good accuracy, sensitivity, and specificity.

Fig. 9 displays performance characteristics such as precision, recall, and F-measure graphically for both the proposed and existing techniques. When compared to earlier approaches, the CESSO-HCRNN model offers high precision, recall, and F- Measure.

Fig. 10 illustrates the performance metrics, including NPV and MCC, for the proposed and existing techniques. The CESSO-HCRNN model has a higher NPV and MCC when compared to the methods currently in use.

Fig. 11 visually represents the FPR and FNR performance metrics for the proposed and current techniques. The CESSO-HCRNN model exhibits lower FPR and FNR when compared to currently employed approaches, according to the comparison.

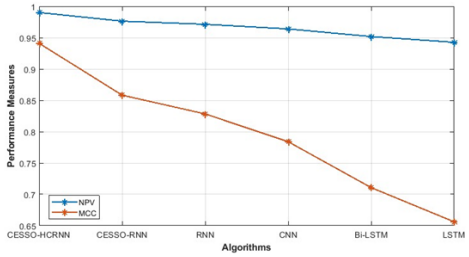


Fig. 10. Comparison of the NPV and MCC of NSL-KDD dataset for the proposed CESSO-HCRNN and existing techniques.

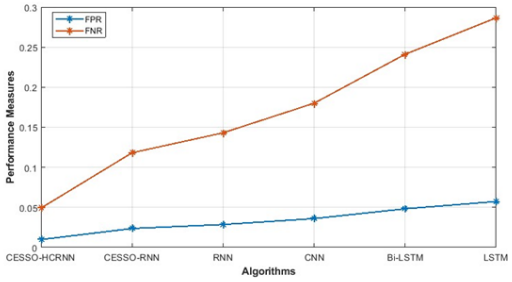


Fig. 11. Comparison of FPR and FNR of NSL-KDD dataset for the proposed CESSO-HCRNN and existing techniques.

TABLE II. COMPARISON OF PERFORMANCE METRICS OF NSL-KDD DATASET FOR THE PROPOSED CESSO-HCRNN AND EXISTING TECHNIQUES

Techniques	Sensitivity	Specificity	Accuracy	Precision	Recall	F-Measure	NPV	MCC	FPR	FNR
Proposed CESSO - HCRNN	0.9509	0.9902	0.9836	0.9509	0.9509	0.9509	0.9902	0.9410	0.0098	0.0491
CESSO-RNN	0.8819	0.9764	0.9606	0.8819	0.8819	0.8819	0.9764	0.8582	0.0236	0.1181
RNN	0.8569	0.9714	0.9523	0.8569	0.8569	0.8569	0.9714	0.8283	0.0286	0.1431
CNN	0.8200	0.9640	0.9400	0.8200	0.8200	0.8200	0.9640	0.7840	0.0360	0.1800
Bi-LSTM	0.7589	0.9518	0.9196	0.7589	0.7589	0.7589	0.9518	0.7107	0.0482	0.2411
LSTM	0.7134	0.9427	0.9045	0.7134	0.7134	0.7134	0.9427	0.6560	0.0573	0.2866

VI. CONCLUSION

As a zero-day attack is a random assault that cannot be anticipated, the zero-day attack has recently taken on highly dangerous consequences. The zero-day threat takes the use of a software vulnerability to access the system or do significant harm, and system developers have no time to fix this vulnerability to reduce the threat. The input data is received from the two datasets, the missing values are removed, and the data normalization is performed in the pre-processed section. A unique hybrid feature selection method that is based on the CESSO and Information Gain(IG) are implemented. The CESSO is also used to improve the Recursive Neural Network (RNN) performance to produce an optimized RNN. The improved RNN is hybrid with the CNN to produce the HCRNN, which is used to predict the zero-day attack. The results of the proposed CESSO-HCRNN are compared with the existing models with reduced error values.

In two test cases, the performance of the proposed methodology is examined. The NSL-KDD dataset is used as the initial test case for the proposed system. The DoS module of the CSE-CIC-IDS2018 is utilized as the source dataset in the second test scenario, where the framework is employed to identify zero-day attacks on the NSL-KDD dataset. In this case, both domains have unique feature spaces and probability distributions. Although there is significant variation between the source and target domains, the experimental findings show that the suggested strategy is effective in identifying zero-day attacks with no tagged occurrences.

REFERENCES

[1] P. Anand, Y. Singh, and A. Selwal, "Learning-based techniques for assessing zero-day attacks and vulnerabilities in iot," *Recent Innovations in Computing: Proceedings of ICRIC 2021, Volume 1*, pp. 497–504, 2022.

[2] M. Nkongolo, J. P. Van Deventer, S. M. Kasongo, S. R. Zahra, and J. Kipongo, "A cloud based optimization method for zero-day threats detection using genetic algorithm and ensemble learning," *Electronics*, vol. 11, no. 11, p. 1749, 2022.

[3] R. Kumar and G. Subbiah, "Zero-day malware detection and effective malware analysis using shapley ensemble boosting and bagging approach," *Sensors*, vol. 22, no. 7, p. 2798, 2022.

[4] R. Vinayakumar, M. Alazab, K. Soman, P. Poornachandran, and S. Venkatraman, "Robust intelligent malware detection using deep learning," *IEEE Access*, vol. 7, pp. 46 717–46 738, 2019.

[5] A. Mihoub, O. B. Fredj, O. Cheikhrouhou, A. Derhab, and M. Krichen, "Denial of service attack detection and mitigation for internet of things using looking-back-enabled machine learning techniques," *Computers & Electrical Engineering*, vol. 98, p. 107716, 2022.

[6] B. I. Hairab, M. S. Elsayed, A. D. Jurcut, and M. A. Azer, "Anomaly detection based on cnn and regularization techniques against zero-day attacks in iot networks," *IEEE Access*, vol. 10, pp. 98 427–98 440, 2022.

[7] A. Abeshu and N. Chilamkurti, "Deep learning: The frontier for distributed attack detection in fog-to-things computing," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 169–175, 2018.

[8] A. Fatima, S. Kumar, and M. K. Dutta, "Host-server-based malware detection system for android platforms using machine learning," in *Advances in Computational Intelligence and Communication Technology: Proceedings of CICT 2019*. Springer, 2021, pp. 195–205.

[9] I. A. Khan, N. Moustafa, D. Pi, W. Haider, B. Li, and A. Jolfaei, "An enhanced multi-stage deep learning framework for detecting malicious activities from autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 25 469–25 478, 2021.

[10] M. Ali, A. Siddique, A. Hussain, F. Hassan, A. Ijaz, and A. Mehmood, "A sustainable framework for preventing iot systems from zero day ddos attacks by machine learning," *Int. J. Emerg. Technol*, vol. 12, pp. 116–121, 2021.

[11] W. Haider, N. Moustafa, M. Keshk, A. Fernandez, K.-K. R. Choo, and A. Wahab, "Fgmc-hads: Fuzzy gaussian mixture-based correntropy models for detecting zero-day attacks from linux systems," *Computers & Security*, vol. 96, p. 101906, 2020.

[12] V. Sharma, K. Lee, S. Kwon, J. Kim, H. Park, K. Yim, and S.-Y. Lee, "A consensus framework for reliability and mitigation of zero-day attacks in iot," *Security and Communication Networks*, vol. 2017, 2017.

[13] S. Venkatraman and M. Alazab, "Use of data visualisation for zero-day malware detection," *Security and Communication Networks*, vol. 2018, pp. 1–13, 2018.

[14] F. Alhaidari, N. A. Shaib, M. Alsafi, H. Alharbi, M. Alawami, R. Aljindan, A.-u. Rahman, and R. Zagrouba, "Zevigilante: Detecting zero-day malware using machine learning and sandboxing analysis techniques," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[15] W.-S. Choi, S.-Y. Lee, and S.-G. Choi, "Implementation and design of a zero-day intrusion detection and response system for responding to network security blind spots," *Mobile Information Systems*, vol. 2022, 2022.

[16] X. Sun, J. Dai, P. Liu, A. Singhal, and J. Yen, "Using bayesian networks for probabilistic identification of zero-day attack paths," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 10, pp. 2506–2521, 2018.

[17] Y. Afek, A. Bremler-Barr, and S. L. Feibish, "Zero-day signature extraction for high-volume attacks," *IEEE/ACM Transactions on Networking*, vol. 27, no. 2, pp. 691–706, 2019.

[18] T. Zoppi, A. Ceccarelli, and A. Bondavalli, "Unsupervised algorithms to detect zero-day attacks: Strategy and application," *Ieee Access*, vol. 9, pp. 90 603–90 615, 2021.

[19] I. Mbona and J. H. Eloff, "Detecting zero-day intrusion attacks using semi-supervised machine learning approaches," *IEEE Access*, vol. 10, pp. 69 822–69 838, 2022.

[20] S. I. Popoola, R. Ande, B. Adebisi, G. Gui, M. Hammoudeh, and O. Jogunola, "Federated deep learning for zero-day botnet attack detection in iot-edge devices," *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3930–3944, 2021.

[21] R. Bar and C. Hajaj, "Simcse for encrypted traffic detection and zero-day attack detection," *IEEE Access*, vol. 10, pp. 56 952–56 960, 2022.

[22] M. Zhang, L. Wang, S. Jajodia, A. Singhal, and M. Albanese, "Network diversity: a security metric for evaluating the resilience of networks against zero-day attacks," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 5, pp. 1071–1086, 2016.

[23] V. Kumar and D. Sinha, "A robust intelligent zero-day cyber-attack detection technique," *Complex & Intelligent Systems*, vol. 7, no. 5, pp. 2211–2234, 2021.

[24] A. Blaise, M. Bouet, V. Conan, and S. Secci, "Detection of zero-day attacks: An unsupervised port-based approach," *Computer Networks*, vol. 180, p. 107391, 2020.

[25] "Data set1 collected from:," https://www.kaggle.com/datasets/solarmainframe/ids-intrusion-csv" , dated on 29/11/2022.

[26] "Data set2 collected from:," "https://www.kaggle.com/datasets/hassan06/nslkdd" , dated on 29/11/2022.

[27] G. Zhang, J. Hou, J. Wang, C. Yan, and J. Luo, "Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 12, pp. 288–301, 2020.

[28] S. Kassaymeh, S. Abdullah, M. A. Al-Betar, and M. Alweshah, "Salp swarm optimizer for modeling the software fault prediction problem," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 3365–3378, 2022.

[29] N. Singh, S. Singh, and E. H. Houssein, "Hybridizing salp swarm algorithm with particle swarm optimization algorithm for recent optimization functions," *Evolutionary Intelligence*, pp. 1–34, 2022.

[30] M. Khishe and M. R. Mosavi, "Chimp optimization algorithm," *Expert systems with applications*, vol. 149, p. 113338, 2020.

[31] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Bencherif, and M. S. Hossain, "Multilevel weighted feature fusion using convolutional neural networks for eeg motor imagery classification," *Ieee Access*, vol. 7, pp. 18 940–18 950, 2019.

[32] J. Ma, W. Gao, S. Joty, and K.-F. Wong, "An attention-based rumor detection model with tree-structured recursive neural networks," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 4, pp. 1–28, 2020.

Triggered Screen Restriction: Gamification Framework

Majed Hariri[1], Richard Stone[2]
HCI Department, Iowa State University, Iowa State University, Ames, USA[1]
Industrial and Manufacturing Systems Engineering Department, Iowa State University, Ames, USA[2]

***Abstract*—The prevalence of sedentary lifestyles is increasingly becoming a significant public health concern, with numerous health risks ranging from obesity to heart disease. Several gamified interventions have been employed to counter sedentary behavior by promoting physical activity. However, the existing approaches have yielded mixed results, making it crucial to explore new methodologies. While existing approaches have utilized gamification elements to encourage activity, they often need a comprehensive blend of psychological elements and advanced technology to drive a meaningful behavioral alteration. This paper introduces the Triggered Screen Restriction (TSR) framework, an interdisciplinary approach integrating behavioral psychology, gamification, and screen-time restriction technologies. The TSR framework aims to elevate gamified physical activity by leveraging the psychological Fear of Missing Out phenomenon, encouraging users to meet specific activity goals to unlock social media applications. The TSR framework presents a promising avenue for future research. The proposed framework's unique approach is designed to motivate users to be more physically active. The proposed framework fills a literature gap in the current implementation of the gamified physical intervention. Further studies are needed to empirically validate the framework's effectiveness and potential to contribute to the gamification ecosystem.**

Keywords*—Gamification; physical activity; sedentary behavior; Triggered Screen Restriction (TSR) framework

I. Introduction

Engaging in regular physical activity is not just a lifestyle choice; it is a cornerstone for a high-quality life. Physical activity positively impacts not only your physical health but also enhances your mental well-being, enriches your social interactions, and boosts your self-confidence [1]. Despite the wide-ranging benefits, alarming data shows a global drop in levels of physical activity [2]. One factor amplifying this decline is the widespread use of smartphones. While these devices have made life easier in many ways, they have also unintentionally encouraged a lifestyle that involves minimal movement [3]. The issue is especially severe in countries with both high smartphone usage and alarming rates of obesity, such as Saudi Arabia [4]. For instance, in Saudi Arabia, a staggering 41.5% of young men are not engaged in any form of regular physical activity, contributing to an obesity rate that is higher than the global average [5], [6]. This trend is not just a national crisis but a global one, necessitating swift and effective solutions to promote active lifestyles [4]. Despite the continuous development, most of the existing gamified physical interventions still have untapped potential. Many of the current methods using gamification to encourage physical activity overly rely on positive rewards [7]. These constant rewards can eventually wear out a person's motivation, making these strategies unsustainable [8].

Gamified physical interventions are often perceived by users as either excessively complex or insufficiently supportive, leading to decreased levels of interaction [9]. Even the most popular fitness interventions available on Apple or Google stores are often too narrow in focus [10]. The proposed framework, aiming to leverage the Fear of Missing Out, addresses these limitations by motivating users to engage in physical activity using gamified intervention to access their social media applications. By leveraging insights from current research and integrating novel activity tracking with motivational game-like elements, the framework might offer a comprehensive solution in gamified physical interventions. The paper seeks to serve as a foundational framework for future gamified physical interventions. The following sections will explore the limitations of current frameworks and the potential of the novel approaches proposed in this framework. The remainder of this paper is organized as follows:

- Objective - Presents the main objective, which is to propose a theoretical framework that might enhance the application of gamification in encouraging physical activity.
- Literature Review - Examines existing gamification strategies and their application in diverse domains, particularly focusing on their role in enhancing physical activity.
- Previous Frameworks - Examines and evaluates popular gamification frameworks, including Octalysis, MDA, SGD, and FRAGGLE.
- Introducing the TSR Framework - Introduces the TSR framework, which outlines its interdisciplinary components for promoting physical activity through gamification.
- Components of the TSR Framework - Explores the four key pillars of the TSR framework: screen time restriction, notification triggers, computer vision model, and reward engine. Describes how these aspects work together to help encourage physical activity through gamified intervention.
- Conclusion - Summarizes the study, emphasizing the TSR framework's potential as a gamified framework to promote physical activity. The conclusion highlights the necessity for empirical validation and proposes future research directions.

II. Objective

The main objective is to propose a theoretical framework that might enhance the application of gamification in encouraging physical activity. This enhancement could be achieved by exploring innovative intersections between behavioral psychology and current technological approaches. The paper seeks to identify novel tools and methods that might increase the effectiveness of gamified interventions, with a particular focus on the potential application of the fear of missing out phenomenon.

The significance lies in the potential impact of gamified physical interventions. By offering a novel perspective on gamification, this paper desires to address the limitations of motivating physical activity in various settings. Its insights are especially relevant in societies where traditional methods to promote physical activity have had limited success, highlighting the need for more innovative and engaging approaches.

This paper's contribution is laying down a theoretical groundwork for a new approach to gamification, focusing on behavioral and technological aspects rather than empirical data. The exploration of the fear of missing out phenomenon as a motivational tool represents a fresh perspective in gamification research. The paper seeks to inspire further academic inquiry, discussion, and development. Ultimately, the proposal in this paper might open up new possibilities for research and application, leading to the creation of more effective and engaging gamified physical interventions for promoting physical activity in the future.

III. Literature Review

The idea of gamification, which incorporates game design elements into non-gaming contexts, offers an innovative method to foster behavioral change. By including elements such as points, rewards, and challenges, gamification has been observed to elevate user engagement and motivation [11].

Further, within the context of gamification, components like badges, progress indicators, and stages hold a pivotal role. These elements, far from being just aesthetic enhancements, act as central motivational tools. Their presence invigorates users, encouraging sustained interaction and consistent effort towards reaching defined objectives [12].

Moreover, the utilization of gamification strategies has been examined across various sectors, including educational institutions, professional environments, and health-related centers. The efficacy in enhancing the user experience by making products and services more engaging has received substantial empirical support [13], [14]. For instance, a study that employed a randomized design explored the impact of a gamified intervention strategy. This strategy was further enhanced by adding elements of social support and financial incentives framed as potential losses. The study demonstrated a moderate yet promising increase in levels of physical activity among veterans who are struggling with weight issues, including obesity [15].

Another study focused on the efficacy of immediate financial rewards given out on a variable schedule. This study found that such financial incentives could significantly encourage people to engage more with mobile health applications [16]. Several studies further corroborate the sustainability of such interventions, especially when participants select their own goals [17], [18].

Despite these positive findings, it's critical to acknowledge that the effectiveness of gamification is not a one-size-fits-all solution. A study that used a randomized design with three different groups showed that although all participants lost a significant amount of weight, the groups that were exposed to gamified intervention did not outperform the control group in a statistically significant manner [8]. Moreover, a separate study aimed at exploring the role of personalized goal-setting in gamified mobile health interventions reported an initial increase in user engagement and performance, but this positive trend appeared to diminish over time [19]. These divergent findings clearly indicate an urgent need for more nuanced investigations to further refine and possibly explain the variability in outcomes associated with gamification in physical health interventions.

When examining the specific techniques and methodologies that are part of gamification, a rich tapestry of strategies reveals itself. One of these strategies involves the real-time monitoring of user metrics and the provision of immediate feedback, which has been shown to elevate the likelihood of successfully encouraging physical activity [20]. Similarly, allowing users to set their own physical activity goals introduces a competitive spirit that invigorates the user's motivation to be physically active [7]. Tangible systems of rewards, often implemented through the use of badges and points, provide compelling reasons for users to not only meet their activity goals but also to exceed them [21]. Creating a sense of community and social interaction is another essential feature of gamified fitness interventions, often achieved through leaderboards that enable users to compare their progress and celebrate their achievements [22], [23]. Additional strategies include sending out notifications and text messages as reminders, which act as consistent nudges to help users stay aligned with their fitness objectives [24], [25].

Psychological theories also play a significant role in how gamified interventions are designed. For instance, Self-Determination Theory is commonly used to ensure that the needs for competence, autonomy, and social connection are adequately addressed, thereby serving as a continual source of motivation [26], [27], [28]. Balancing intrinsic motivations, such as the innate enjoyment derived from an activity, and extrinsic motivations, such as rewards, is critical [26], [27], [28]. Another significant psychological theory that has been applied in gamified intervention is Flow Theory, which suggests that the most engaging and motivating experiences occur when there's a balance between the challenge at hand and the individual's skill level [29], [30]. While the Fear of Missing Out has been criticized for encouraging potentially addictive behaviors [31], recent research suggests that it can also have a positive impact, particularly for individuals who might otherwise be disengaged [32]. These psychological theories, although promising, require further study to confirm their effectiveness in gamified physical interventions.

Conclusively, a systematic review encompassing an examination of 1680 health applications available on the Apple

and Google App stores, including (Nike+ Running, Zombies, Run!, Strava Run, MyFit Fitness, Fitbit, and RunKeeper - GPS Track Run Walk) [10]. Astoundingly, the research found that a mere 4% of these applications employed gamification elements to enhance user engagement and promote healthier behaviors [10]. This relatively low number represents a monumental opportunity for advancement and innovation. Most of the existing gamified applications prioritize self-monitoring components and combine them with goal-setting features to maintain user engagement [10].

The current implementation of gamified interventions mainly relies on positive feedback and personal determination to encourage users. But this approach has limits. There are still unexplored ways to use technology and psychology to keep people interested and involved in physical activities. A promising avenue in employing on-device computer vision models that can detect and report user activities, offering a more immersive and interactive experience while preserving user privacy [33]. Using gamified intervention could boost users' interaction with physical activity applications and significantly increase user engagement and effectiveness.

Furthermore, there's a need to evolve beyond merely using positive reinforcement. The emphasis should shift towards a balanced approach, incorporating both positive and negative reinforcement mechanisms to create a more captivating, or even addictive, user experience [34]. By doing so, there is an incredible opportunity to develop novel frameworks that are not only engaging but also effective in battling sedentary lifestyles and encouraging physical activity among various population groups. The following subsection will focus on previous frameworks to provide a more comprehensive understanding of gamification's frameworks and their current state.

A. Previous Frameworks

1) Octalysis Framework: The Octalysis framework emerged from the realization of the need for a tool to devise strategies and evaluate the implementation of gamification [35]. The Octalysis framework identified eight distinct core drives that propel individuals to engage in certain activities. The Octalysis framework, visually represented as an octagon, encapsulates these core drives at each of its corners (see Fig. 1). The Octalysis framework emphasizes the importance of identifying whether core drives lean towards extrinsic or intrinsic motivation.

The Octalysis framework considers the core drives as essential components, which are defined in the following manner:

- Epic Meaning and Calling: This drive encapsulates the desire to be part of something larger than oneself or to pursue a higher purpose.
- Development and Accomplishment: This is the drive to improve, overcome challenges, and achieve goals.
- Ownership and Possession: This is the drive to own or control resources and protect one's investments.
- Scarcity and Impatience: This drive is about the desire to obtain rare or exclusive items or act before an opportunity passes.

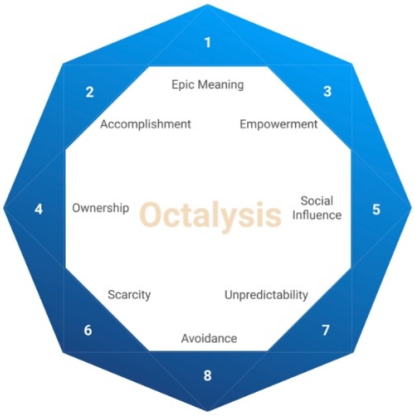


Fig. 1. [35] Octalysis framework.

- Loss and Avoidance: This is the drive to prevent loss, negative outcomes, or maintain one's status quo.
- Unpredictability and Curiosity: This drive encapsulates the desire for novelty and unexpected outcomes.
- Social Influence and Relatedness: This drive encompasses the need to connect with others and belong to a group.
- Empowerment of Creativity and Feedback: This drive covers the desire to express oneself, experiment, and receive feedback.

The Octalysis framework divides core drives into two impactful categories: Black Hat Gamification and White Hat Gamification. White Hat drives, found at the top, motivate positively, inspiring creativity, control, and purpose. These include Epic Meaning, Development, and Empowerment of Creativity.

In contrast, the lower drives, or Black Hat, are linked to negative motivations like urgency or addiction. They encompass Scarcity, Unpredictability, and Loss. The side drives, Ownership, and Social Influence can swing either way based on context. For effective gamification, the Octalysis framework recommends balancing both Black and White Hat techniques. White Hat fosters loyalty but may not prompt immediate reactions. Black Hat encourages immediate action but might cause burnout. An ideal gamification design balances both aiming for sustained motivation and a rewarding experience.

2) MDA Framework: The Mechanics, Dynamics, and Aesthetics (MDA) framework provides a comprehensive approach to game design and analysis, aiming to bridge different areas like game development, game criticism, and technical research in the gaming industry [36]. The MDA framework categorizes games into three main elements: Mechanics, Dynamics, and Aesthetics (see Fig. 2). These components represent various aspects of game design and player involvement [36].

- Mechanics: The core elements of a game that include data representation and algorithms used to support the game's framework.

- Dynamics: The real-time mechanics of responding to player inputs, which evolve during gameplay.
- Aesthetics: The emotional reactions players have while interacting with a game system, emphasizing the experiential aspect of games.

The MDA framework has been found to be effective in increasing user engagement in various platforms, including donation-based crowdfunding. Applying the MDA framework has resulted in higher levels of interaction from users, indicating the potential for increased funding for charitable initiatives [37].Furthermore, the MDA framework has been used as an educational tool to enhance learning experiences and comprehension of mathematical concepts in elementary students [38].

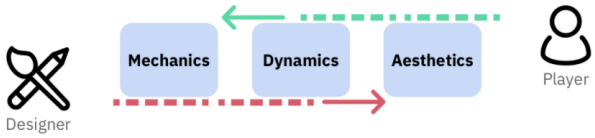


Fig. 2. [36] The MDA framework.

3) Sustainable Gamification Design Framework: The Sustainable Gamification Design (SGD) framework is a systematic approach to creating gamified systems with a focus on sustainability in user engagement, environmental impact, and social responsibility. The SGD framework emphasizes ethical and human-centered design principles [39]. The framework is structured around four key stages:

- Discover: This initial phase involves comprehensively understanding the setting and the individuals or groups that the system will impact.
- Reframe: In this stage, designers evaluate the gathered information to spot opportunities and develop potential solutions.
- Envision: Decision-making is key in this phase, as designers choose the most fitting solution for the system.
- Create: This final step sees the design and implementation of the gamified system, bringing the concept to fruition.

The integration of values and ethical considerations is central to the SGD framework, ensuring that the gamified systems produced are not only captivating but also responsible and considerate of broader impacts. The framework guides designers in creating systems that are beneficial for users (see Fig. 3).

4) FRAGGLE Framework: The FRAGGLE framework is an agile methodology tailored for enhancing learning experiences through gamification. The framework is designed to align gamified activities with educational goals, content, and assessment criteria, ensuring that game elements support the intended learning outcomes [40]. The framework consists of four phases:

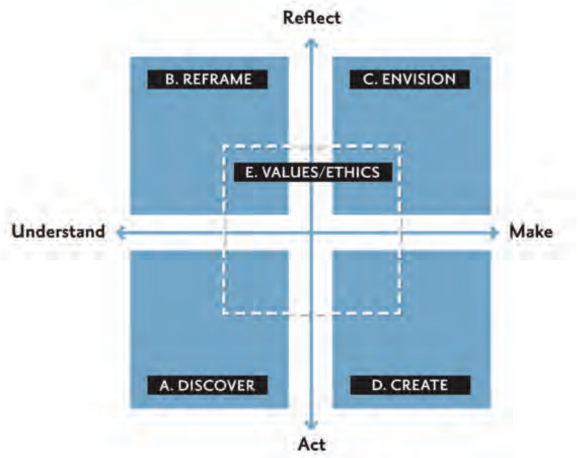


Fig. 3. [39] The SGD framework.

- Declaration: Identifying problems, user stories, and acceptance tests to define the project's scope and requirements.
- Creation: Designing engaging game elements like players, mechanics, stages, actions, and triggers that meet user needs.
- Execution: Implementing and deploying the gamified learning experience to deliver an MVP for user feedback.
- Learning: Measuring and analyzing the gamification's impact on learning outcomes and user satisfaction, with a focus on continuous improvement.

The framework emphasizes learning experiences rather than complete gamified systems. The framework-structured approach is helpful for creating engaging and educationally effective gamified activities. The step-by-step process of the framework, from conception to evaluation, facilitates the agile development of gamified learning experiences that align with educational goals and respond to learner feedback (see Fig. 4).

B. Overview and Comparison of the Models

Gamification has gained popularity as a strategy to increase engagement in different fields. Several frameworks guide how to incorporate game elements into non-game settings. The MDA framework provides a way to understand the relationship between game mechanics, dynamics, and aesthetics. However, the MDA framework has been criticized for focusing too much on mechanics and not considering other aspects like user experience and narrative elements in games [41].

The Octalysis framework offers valuable insights into the various factors that drive people's engagement in activities. However, the Octalysis framework doesn't provide a structured design process, and the generic approach may not cater to the diverse motivations and backgrounds of all users.

Similarly, the FRAGGLE framework is agile and learner-centered, designed to align educational activities with gamification elements efficiently. However, the FRAGGLE framework does not address real-world challenges, such as

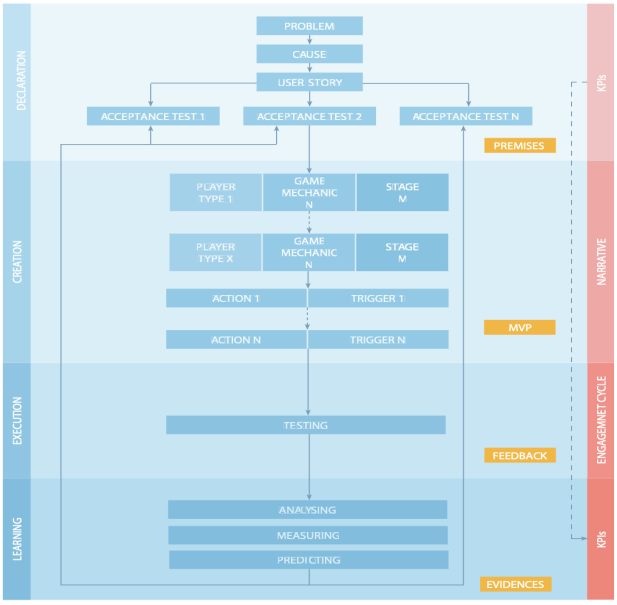


Fig. 4. [40] The FRAGGLE framework.

technical issues or navigating organizational culture, that can arise in the implementation of these systems. These frameworks provide valuable insight into gamification, but they also show the need for a more comprehensive model that can address some of their limitations.

As interest increases in using gamification to encourage physical activity, there is an exciting chance to develop a new framework. The novel framework could be customized to encourage physical activity by considering the specific needs and motivations of various population groups. The novel framework has the potential to lead to greater and more widespread adoption of healthy habits by harnessing gamification's ability to create a positive impact on users (see Table I).

TABLE I. Overview of Gamification Frameworks

Framework	Target	Implementation
Octalysis [35]	General	Transforming into a game-like process
MDA [36]	General	Connects game design to gamified development
SGD [39]	Design	Conceptual framework focused on design
FRAGGLE [40]	Education	Uses gamified agile to enrich learning
The Proposed Framework (TSR)	Physical Activity	Interdisciplinary approach for gamified physical intervention

IV. Introducing the TSR Framework: A Multidisciplinary Approach

The need for a novel approach to address the growing issue of physical inactivity is evident [34]. While existing gamified physical interventions offer a range of features to promote physical activity, they often fail to achieve lasting user engagement. Many such interventions rely too much on continual positive reinforcement, and their heavy reliance on willpower can deplete the user's motivation over time [7], [8]. Additionally, current approaches frequently overlook crucial aspects that could drive engagement [9].

Against this backdrop, the TSR framework is introduced as a conceptual solution to these pressing concerns. Central to the TSR framework's aim is using the Fear of Missing Out phenomenon to inspire a meaningful change by using gamified physical activity intervention, combining the latest technology with fundamental elements of behavioral psychology.

In contrast to existing gamified interventions, the TSR framework employs a balanced system of both rewards and restrictions to encourage increased physical activity. At the core of the TSR framework are four integral components: Screen Time Restriction, Notification Triggers, Computer Vision Model, and Reward Engine. These pillars serve specialized functions that collectively offer a well-rounded user experience:

- Screen Time Restriction: Restricts access to distracting social media applications unless specific physical activity goals are met, thus leveraging the Fear Of Missing Out phenomenon.
- Notification Triggers: Customizable alerts remind users of their activity goals and offer motivation at opportune moments.
- Computer Vision Model: Detects the user's physical activity in real-time, providing instant feedback while ensuring data privacy.
- Reward Engine: Offers tangible rewards like points and unlocks varying levels of exercise challenges, making the whole gamified experience more engaging.

The proposed framework builds upon existing research, which supports the efficacy of these elements. For example, goal-setting strategies have been proven to increase physical activity, improve well-being, and lower health risk factors like Body Mass Index (BMI) [42]. Similarly, studies show that the balanced use of gamification can indeed boost a person's intrinsic drive to exercise [21]. To provide a comprehensive understanding of the TSR framework's capabilities and how it differentiates itself from existing interventions, the following sections will dive into the technical and psychological aspects of each foundational pillar in detail (see Fig. 5).

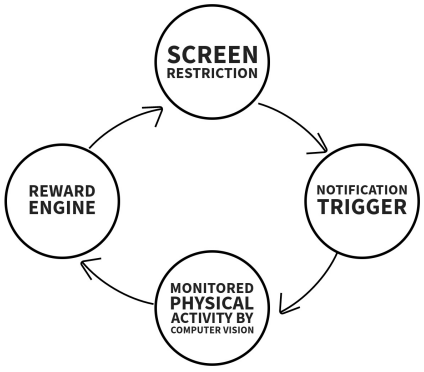


Fig. 5. The Proposed TSR's Workflow.

A. Screen Time Restriction: Tackling Physical Inactivity Through Behavioral Incentives and Technological Limits

The first foundational pillar of the TSR framework is Screen Time Restriction, a feature designed to balance screen-based activities with gamified physical activity. The mechanics of Screen Time Restriction are straightforward yet effective. Once the user crosses a predetermined time limit on social media or other distracting applications, the framework triggers a lockout mechanism. This restriction can be lifted only when the user achieves certain physical activity goals, which are recorded and verified in real-time by the framework's other components like the Computer Vision Model and Notification Triggers. This approach turns a usually passive screen time experience into an active pursuit of physical milestones.

Besides merely cutting off access to applications, the TSR framework utilizes the Fear Of Missing Out psychological phenomenon to make this strategy even more compelling. Users are enticed to achieve their physical activity goals not just for the sake of better health but also to regain access to their social circles online. This adds a rich social layer to the otherwise technology-focused Screen Time Restriction feature, elevating it beyond a simple tech-based solution.

To ensure user privacy, the framework operates under stringent privacy policies. The framework only accesses screen-time data in real-time and does not store any of this sensitive information, thereby adhering to top-level privacy standards. By combining technological control with behavioral psychology, the Screen Time Restriction component creates a multi-dimensional approach to encouraging physical activity. The Screen Time Restriction is not just about limiting screen time but about turning those limitations into a motivational force that encourages physical activity.

B. Notification Triggers: Steering User Attention Through Timely Reminders

The second pillar of the TSR framework is Notification Triggers. The Notification Triggers component is designed to provide real-time engagement through a system of push notifications. These reminders serve as nudges that propel users toward physical activity, filling the spaces in their day with opportune moments for exercise. The operational backbone of Notification Triggers is its interoperable system architecture. By leveraging Firebase Cloud Messaging (FCM) for Android and Apple Push Notification Service (APNs) for iOS, the framework ensures that notifications reach users irrespective of their choice of operating system. This universal approach guarantees that all users have equal opportunity to benefit from the framework, regardless of their device preference. The framework provides a range of pre-set messages, which can be as simple as a reminder.

While the content of these messages is standardized, the timing, frequency, and types of notifications can be personalized according to each user's needs and lifestyle. This degree of customization fosters a more personal connection between the user and the framework, increasing the likelihood of sustained engagement. Although the Notification Triggers start as external cues, the ultimate goal is to transition users from needing these reminders to developing intrinsic motivation for physical activity. This shift aligns with the objectives of the TSR framework, combining screen time restriction and notification triggers to promote gamified physical activity. By harmoniously integrating the notification triggers with the Screen Time Restriction component, the TSR framework might create a continuous loop of motivation and action, making strides toward more engagement in physical activity (see Fig. 6).

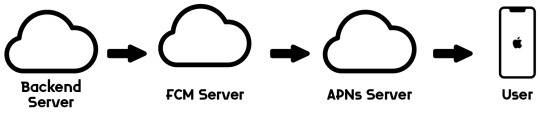


Fig. 6. The proposed notification workflow.

C. Computer Vision Model: Real-time Tracking and Feedback for Optimized Physical Activity

Building on the synergy of Screen Time Restriction and Notification Triggers, the next cornerstone of the TSR framework introduces an advanced technological interface—the Computer Vision Model. The Computer Vision component combines a machine learning model with the camera capabilities of mobile devices to offer real-time assessment and feedback on a gamified physical activity. Employing the power of the computer vision model, the component can identify and quantify a wide array of exercises, such as jumping jacks, in real-time. As users execute these exercises, the model counts the number of repetitions and assesses them, thus providing immediate, accurate metrics to inform more effective workout sessions.

What distinguishes the computer vision approach from traditional methods of tracking physical activities is its commitment to user privacy. While the traditional method relies on the usage of GPS data or accelerometer, the proposed framework performs all data processing directly on the device, ensuring that user data remains secure and private. This eliminates the need for data transfer and limits storage requirements, thus reducing the risk of unauthorized access and data breaches. The provision of instant feedback creates an environment of positive reinforcement. The instant feedback strengthens the user's engagement, as the immediate data allows for immediate adjustments, maximizing the effectiveness of the gamified workout [33]. The computer vision approach thus might enhance the interactive experience, making the TSR framework not just a novel approach but an engaged companion in promoting gamified physical activity. Integrating this advanced Computer Vision Model might add another layer of interactivity and personalization to the TSR framework. It not only advances the framework's primary aim of promoting physical activity using gamification but does so while prioritizing user security and data privacy (see Fig. 7).

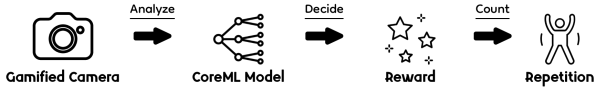


Fig. 7. The proposed computer vision workflow.

D. Reward Engine: An Approach to Sustaining Physical Activity and Reducing Screen Time

Capping off the TSR framework's multi-dimensional approach is the Reward Engine, a dynamic system designed to foster ongoing user engagement. Unlike traditional models, which often employ a one-size-fits-all strategy for physical activity, the Reward Engine employs a responsive mathematical model to tailor tasks and rewards according to the user's specific metrics and performance levels. Utilizing a two-variable formula, the system calculates task difficulty and reward value as follows:

$$Reward = valueReward \times difficultyFactor$$

$$DifficultyFactor = max(min(numberRepetition * valueReward, 0.9), 0.1)$$

Through the application of this equation, the reward engine dynamically adjusts both the challenge and rewards of each task based on real-time performance metrics. By doing so, the reward engine might ensure that activities are optimally engaging—without being overly difficult or too simplistic—creating an ideal balance that sustains user motivation over time. What sets the Reward Engine apart is its focus on providing a highly personalized experience. It offers a mix of intrinsic and extrinsic motivators, making physical activity not just a routine but a rewarding pursuit. This goes beyond merely distributing rewards and plays a crucial role in sustaining behavioral change. This aspect of the TSR framework is critical for the potential of encouraging users to integrate regular physical activities into their lives, potentially leading to a reduction in screen time and promoting more physical activity.

V. Conclusion

The TSR framework, as discussed in this paper, is a conceptual gamification framework awaiting actual deployment. The proposed TSR framework, emphasizing the Fear of Missing Out phenomenon, presents a potential comprehensive strategy to encourage physical activity through gamified activity goals linked to social media application access. The proposed framework has four main components: Screen Time Restriction, Notification Triggers, Computer Vision Model, and Reward Engine. The TSR's components initiatives aim to work together in order to bring a potentially significant change in our approach to gamification and physical activity.

While this paper outlines the blueprint of the TSR framework, it remains essential to mention that this is a concept framework—a proposal that is yet to be brought to life and measured against real-world scenarios. The importance of thoroughly evaluating the proposed framework in future research cannot be overstated. A well-considered plan for evaluation becomes a cornerstone for future research to turn this conceptual model into a tangible framework. This evaluation would leverage both numerical data and human experiences to provide a full-spectrum analysis of the framework's performance. For the numerical data, metrics such as user engagement, time spent on physical activities, and program adherence could be good starting points.

On the other hand, understanding human experiences could be achieved through qualitative methods, like interviews or surveys, to understand user satisfaction and program perception in depth. The duality of these methods provides a well-rounded look at the framework's effectiveness or areas requiring refinement. While this is still theoretical, the concept itself calls for a future academic inquiry that dives into its practical applicability.

The adaptability and modular design of the TSR framework serve as a launching pad for future research. The framework's capacity to be customized to fit a wide array of user preferences makes it a strong candidate for various real-world applications. Potential applications could range from educational settings targeting younger populations to workplace environments aiming to boost employee efficiency. Exploring the removal of boundaries between sectors could also be an interesting area of study.

To conclude, while the TSR framework remains a theoretical model at present, its potential applications and impact need further investigation. This paper does not claim to have achieved these outcomes but seeks to set the academic and practical communities a task: rigorously test, refine, and eventually implement the TSR framework.

Acknowledgment

The study efforts of Majed Hariri were supported by the Islamic University of Madinah.

Disclosure of Interest

The authors report there are no competing interests to declare

References

[1] M. H. Abdollahi, S. Gholami Torkesaluye, and F. Mohammad Hassan, "Effect of participating in physical activities on the quality of life," *Journal of Exercise and Health Science*, vol. 1, no. 1, pp. 51–60, 2021.

[2] S.-E. Kim, J.-W. Kim, and Y.-S. Jee, "Relationship between smartphone addiction and physical activity in chinese international students in korea," *Journal of behavioral addictions*, vol. 4, no. 3, pp. 200–205, 2015.

[3] K. J. Newzoo's, "Global mobile market report: Insights into the world's 3 billion smartphone users," *Newzoo.[(accessed on 3 January 2019)]*, 2018.

[4] J. A. Olson, D. A. Sandra, É. S. Colucci, A. Al Bikaii, D. Chmoulevitch, J. Nahas, A. Raz, and S. P. Veissière, "Smartphone addiction is increasing across the world: A meta-analysis of 24 countries," *Computers in Human Behavior*, vol. 129, p. 107138, 2022.

[5] V. Salem, N. AlHusseini, H. I. Abdul Razack, A. Naoum, O. T. Sims, and S. A. Alqahtani, "Prevalence, risk factors, and interventions for obesity in saudi arabia: a systematic review," *Obesity Reviews*, vol. 23, no. 7, p. e13448, 2022.

[6] J. Z. AlTamimi, R. I. Alagal, N. M. AlKehayez, N. M. Alshwaiyat, H. A. Al-Jamal, and N. A. AlFaris, "Physical activity levels of a multi-ethnic population of young men living in saudi arabia and factors associated with physical inactivity," *Frontiers in Public Health*, vol. 9, p. 734968, 2022.

[7] M. A. Harris, "Maintenance of behaviour change following a community-wide gamification based physical activity intervention," *Preventive medicine reports*, vol. 13, pp. 37–40, 2019.

[8] G. W. Kurtzman, S. C. Day, D. S. Small, M. Lynch, J. Zhu, W. Wang, C. A. Rareshide, and M. S. Patel, "Social incentives and gamification to promote weight loss: the lose it randomized, controlled trial," *Journal of general internal medicine*, vol. 33, pp. 1669–1675, 2018.

[9] J.-H. Yu, G. C.-M. Ku, Y.-C. Lo, C.-H. Chen, and C.-H. Hsu, "Identifying the antecedents of university students' usage behaviour of fitness apps," *Sustainability*, vol. 13, no. 16, p. 9043, 2021.

[10] E. A. Edwards, J. Lumsden, C. Rivas, L. Steed, L. Edwards, A. Thiyagarajan, R. Sohanpal, H. Caton, C. Griffiths, M. Munafò *et al.*, "Gamification for health promotion: systematic review of behaviour change techniques in smartphone apps," *BMJ open*, vol. 6, no. 10, p. e012447, 2016.

[11] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: defining" gamification"," in *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, 2011, pp. 9–15.

[12] K. Huotari and J. Hamari, "A definition for gamification: anchoring gamification in the service marketing literature," *Electronic markets*, vol. 27, no. 1, pp. 21–31, 2017.

[13] G. Zichermann and C. Cunningham, *Gamification by design: Implementing game mechanics in web and mobile apps*. " O'Reilly Media, Inc.", 2011.

[14] J. Hamari, J. Koivisto, and H. Sarsa, "Does gamification work?–a literature review of empirical studies on gamification," in *2014 47th Hawaii international conference on system sciences*. Ieee, 2014, pp. 3025–3034.

[15] A. K. Agarwal, K. J. Waddell, D. S. Small, C. Evans, T. O. Harrington, R. Djaraher, A. L. Oon, and M. S. Patel, "Effect of gamification with and without financial incentives to increase physical activity among veterans classified as having obesity or overweight: a randomized clinical trial," *JAMA Network Open*, vol. 4, no. 7, pp. e2 116 256–e2 116 256, 2021.

[16] R. Nuijten, P. Van Gorp, A. Khanshan, P. Le Blanc, A. Kemperman, P. van den Berg, and M. Simons, "Health promotion through monetary incentives: evaluating the impact of different reinforcement schedules on engagement levels with a mhealth app," *Electronics*, vol. 10, no. 23, p. 2935, 2021.

[17] M. S. Patel, C. Bachireddy, D. S. Small, J. D. Harrison, T. O. Harrington, A. L. Oon, C. A. Rareshide, C. K. Snider, and K. G. Volpp, "Effect of goal-setting approaches within a gamification intervention to increase physical activity among economically disadvantaged adults at elevated risk for major adverse cardiovascular events: the engage randomized clinical trial," *JAMA cardiology*, vol. 6, no. 12, pp. 1387–1396, 2021.

[18] X. S. Chen, S. Changolkar, A. S. Navathe, K. A. Linn, G. Reh, G. Szwartz, D. Steier, S. Godby, M. Balachandran, J. D. Harrison *et al.*, "Association between behavioral phenotypes and response to a physical activity intervention using gamification and social incentives: Secondary analysis of the step up randomized clinical trial," *PLoS One*, vol. 15, no. 10, p. e0239288, 2020.

[19] R. Nuijten, P. Van Gorp, A. Khanshan, P. Le Blanc, P. van den Berg, A. Kemperman, M. Simons *et al.*, "Evaluating the impact of adaptive personalized goal setting on engagement levels of government staff with a gamified mhealth tool: results from a 2-month randomized controlled trial," *JMIR mHealth and uHealth*, vol. 10, no. 3, p. e28801, 2022.

[20] I. Cho, K. Kaplanidou, and S. Sato, "Gamified wearable fitness tracker for physical activity: a comprehensive literature review," *Sustainability*, vol. 13, no. 13, p. 7017, 2021.

[21] J. Xu, A. Lio, H. Dhaliwal, S. Andrei, S. Balakrishnan, U. Nagani, and S. Samadder, "Psychological interventions of virtual gamification within academic intrinsic motivation: A systematic review," *Journal of Affective Disorders*, vol. 293, pp. 444–465, 2021.

[22] A. N. Saleem, N. M. Noori, and F. Ozdamli, "Gamification applications in e-learning: A literature review," *Technology, Knowledge and Learning*, vol. 27, no. 1, pp. 139–159, 2022.

[23] M. Lister, "Gamification: The effect on student motivation and performance at the post-secondary level," *Issues and Trends in Educational Technology*, vol. 3, no. 2, 2015.

[24] S. Liu and J. F. Willoughby, "Do fitness apps need text reminders? an experiment testing goal-setting text message reminders to promote self-monitoring," *Journal of health communication*, vol. 23, no. 4, pp. 379–386, 2018.

[25] Z. Zhao, A. Arya, R. Orji, G. Chan *et al.*, "Effects of a personalized fitness recommender system using gamification and continuous player modeling: system design and long-term validation study," *JMIR serious games*, vol. 8, no. 4, p. e19968, 2020.

[26] L. Legault and M. Inzlicht, "Self-determination, self-regulation, and the brain: autonomy improves performance by enhancing neuroaffective responsiveness to self-regulation failure." *Journal of personality and social psychology*, vol. 105, no. 1, p. 123, 2013.

[27] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary educational psychology*, vol. 25, no. 1, pp. 54–67, 2000.

[28] J. S. Nevid, *Essentials of Psychology: Concepts and Applications*, fifth edition ed., 2018. [Online]. Available: https://books.google.com/books?id=OMhgzgEACAAJ

[29] W. Oliveira, A. M. Toda, P. T. Palomino, L. Shi, J. Vassileva, I. I. Bittencourt, and S. Isotani, "Does tailoring gamified educational systems matter? the impact on students' flow experience." in *HICSS*, 2020, pp. 1–10.

[30] M. Csikszentmihalyi, "Beyond boredom and anxiety: Experiencing flow in work and play.-san fransisco, ca: Jossey-bass," 1975.

[31] S. Balta, E. Emirtekin, K. Kircaburun, and M. D. Griffiths, "Neuroticism, trait fear of missing out, and phubbing: The mediating role of state fear of missing out and problematic instagram use," *International Journal of Mental Health and Addiction*, vol. 18, pp. 628–639, 2020.

[32] F. Nursodiq, T. R. Andayani, and M. Supratiwi, "When fear of missing out becomes a good thing," in *International Conference on Community Development (ICCD 2020)*. Atlantis Press, 2020, pp. 254–259.

[33] H.-Y. Kao and Y.-J. Lee, "Design and implement a mobile fitness application based on realtime image detection," in *2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*. IEEE, 2021, pp. 1–2.

[34] M. Hariri and R. Stone, "Gamification in physical activity: State-of-the-art," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 10, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.01410105

[35] Y.-k. Chou, *Actionable gamification: Beyond points, badges, and leaderboards*. Packt Publishing Ltd, 2019.

[36] R. Hunicke, M. LeBlanc, R. Zubek *et al.*, "Mda: A formal approach to game design and game research," in *Proceedings of the AAAI Workshop on Challenges in Game AI*, vol. 4, no. 1. San Jose, CA, 2004, p. 1722.

[37] H. Golrang and E. Safari, "Applying gamification design to a donation-based crowdfunding platform for improving user engagement," *Entertain. Comput.*, vol. 38, p. 100425, 2021.

[38] S. D. Putra and V. Yasin, "Mda framework approach for gamification-based elementary mathematics learning design," *International Journal of Engineering, Science and Information Technology*, 2021.

[39] M. Raftopoulos, "Towards gamification transparency: A conceptual framework for the development of responsible gamified enterprise systems," *Journal of Gaming & Virtual Worlds*, vol. 6, no. 2, pp. 159–178, 2014.

[40] A. Mora, P. Zaharias, C. González, and J. Arnedo-Moreno, "Fraggle: a framework for agile gamification of learning experiences," in *Games and Learning Alliance: 4th International Conference, GALA 2015, Rome, Italy, December 9-11, 2015, Revised Selected Papers 4*. Springer, 2016, pp. 530–539.

[41] W. Walk, D. Görlich, and M. Barrett, "Design, dynamics, experience (dde): an advancement of the mda framework for game design," *Game dynamics: Best practices in procedural and dynamic game content generation*, pp. 27–45, 2017.

[42] N. Jiryaee, Z. D. Siadat, A. Zamani, and R. Taleban, "Comparing of goal setting strategy with group education method to increase physical activity level: A randomized trial," *Journal of Research in Medical Sciences: The Official Journal of Isfahan University of Medical Sciences*, vol. 20, no. 10, p. 987, 2015.

A Particle Filter Based Visual Object Tracking: A Systematic Review of Current Trends and Research Challenges

Md Abdul Awal[1], Md Abu Rumman Refat[2], Feroza Naznin[3], Md Zahidul Islam[4]
Dept. of Information & Communication Technology, Islamic University, Kushtia-7003, Bangladesh[1,2,4]
Dept. of Computer Science and Engineering, Green University of Bangladesh, Dhaka, Bangladesh[2,3]

***Abstract*—Visual object tracking is a crucial research area in computer vision because it can simulate a dynamic environment with non-linear motions and multi-modal non-Gaussian noises. However, This paper presents an overview of the recent developments in particle filter-based visual object tracking algorithms and discusses the pros and cons of particle filters, respectively. There are presentations of many different methodologies and algorithms in the research literature. The majority of visual object tracking research at present is on particle filters. In addition, the most advanced technique for visual object tracking has also been developed by combining the convolutional neural network (CNN) and the particle filter. The advantage of particle filters is that they can handle nonlinear models and non-Gaussian advancements, sequentially concentrating on the areas of the state space with higher densities, primarily parallelization, and simplicity of implementation. Despite this, it offers a robust framework for visual object tracking because it incorporates uncertainty and outperforms other filters like the Kalman filter, Kernelized correlation filter, optical filter, mean shift filter, and extended Kalman filter in recognition tests. In contrast, this study provided information on various particle filter features and classifiers.**

***Keywords*—Particle filter; visual object tracking; On-Gaussian noises; Kalman filter; CNN**

I. Introduction

In the last decade, object tracking has been a difficult research challenge in computer vision for video sequences, with applications including video surveillance, human-computer interaction, augmented reality, and robotics. However, particle filters (PFs) have recently gained popularity in visual object tracking. The ability to resolve non-Gaussian and non-linear problems allowed them to establish a reliable tracking method.

Tracking visual objects involves object recognition, classification, and frame-by-frame tracking. Identifying moving objects in a video stream is the first essential step in tracking. However, in addition to deep learning and CNN, various techniques such as background reduction, statistical models, and specific temporal techniques are applied for object detection. In order to keep track of and analyze observed objects, it is necessary to classify them, such as people, vehicles, animals, debris, etc. The second tracking stage is the creation of temporal connections between detected objects from frame to frame. The segmented regions can be identified temporally using this method, which also generates coherent information about the objects in the observed region, including their trajectory with direction and speed. Typically, the output of the tracking phase is used to assist and improve higher-level activity analysis, object classification, and motion segmentation.

Consequently, deterministic or stochastic frameworks can be used to classify object-tracking techniques. In addition to stochastic approaches, the Bayesian framework for state estimation incorporates stochastic approaches. In the stochastic framework, the Kalman filter [1], the extended Kalman filter, and the unscented Kalman filter [2] were used to estimate linear and Gaussian states. However, the particle filter (PF) [3], the condensation filter [4], and the bootstrap filter were used to estimate nonlinear and non-Gaussian states. Deterministic frameworks, such as mean shift [5], fragment-based tracker, and multi-stage tracker [6], use objective search in each frame to increase the similarity between the objective and search space. These techniques typically rely on less spatial information regarding the object, which makes them more susceptible to occlusion and background clutter [7]. However, stochastic approaches were able to address the problems caused by identical backgrounds and occlusion as well as large variations of a pose by decreasing target sampling patches throughout the track [8], [9].

Particle filters, a stochastic method, surpassed other methods for tracking objects in complex environments such as occlusion, background clutter, and illumination. These are used to estimate system states in state-space models; the system's various states are then tracked over time. By employing sequential Monte Carlo sampling, which uses a collection of samples known as particles to perform numerical approximation, particle filters can also solve the estimation problem. Additionally, particle filters demonstrate their effectiveness in overcoming various challenges associated with object tracking by performing exceptionally well with nonlinear and non-Gaussian estimation problems [4].

This paper's main contributions are summarized as follows:

(a) We present a comprehensive survey of particle filter-based visual object tracking techniques from different perspectives.
(b) We addressed the large-scale benchmark datasets for various visual object tracking approaches.
(c) We also summarize the state-of-the-art visual object tracking research over the past decade, including its advantages, limitations, and future directions.
(d) Finally, we offer some insightful observations and conclusions regarding the tracking of visual objects.

The remaining sections of the paper are organized as follows: The theoretical explanation of particle filters is briefly introduced in **Section II**. Visual object tracking with particle filter is discussed in **Section III**. However, in **Section IV**, benchmark datasets are detailed for visual object tracking. In **Section V**, the comparison techniques of visual object tracking using various PF. Finally, an extensive conclusion is outline in **Section VI**.

II. Particle Filter

Particle filter is a Monte Carlo and recursive Bayesian estimation-based filtering algorithm [10]. Particle filters exhibit superior performance compared to conventional methods such as the Kalman filter when applied to non-linear or non-Gaussian conditions. In addition, Particle filter is widely used in surveillance cameras, robotics, and navigation to track multiple objects using visual, geometric, and motion features, including color, texture, shape, outlines [13].

A particle filter is a Monte Carlo sampling approach for creating a recursive Bayesian filter [11], [12]. Particle filtering is a sampling technique that begins with a population of particles, each of which assigns no value to any variables. The goal of particle filter is to represent the posterior density with a set of random particles and weights, and then compute estimates the state variables' posterior density given the observation variables using these samples and weights. The particle filter is intended for use in a hidden Markov Model with both hidden and observable variables. The particle filter, sometimes known as the condensation filter, is a low-quality filter [15], [16].

The fundamental concept is that particle density distribution occurs when particles are sampled randomly. It is the most general Bayesian strategy since it has no restrictions on the state vector when dealing with nonlinear and non-Gaussian problems. The following is a description of how particle filters function. Based on a measure of probability, the state space is split into several parts, and each part is filled with particles. The higher the likelihood, the higher the particle concentration, according to the state equation, a particle system changes over time with the FPK equation determining the evolving pdf. By randomly selecting states from the state space, we produce a large number of particles that reflect the evolving pdf, because the point-mass histogram can approximate the pdf.

Particle filter technology's adaptability is a result of its dominance in nonlinear and non-Gaussian systems. The multi-modal processing abilities of the particle filter are another factor in the particle filter's widespread use. Particle filtering has been applied to numerous fields around the world. Fig. 1 below illustrates the core concept of the particle filtering (PF) technique.

Let, consider a system whose state changes over time, C_t= g(C_{t-1}, X_t), where C_t represents the current state of the system at time t. The state transition model g determines whether or not the system is $Markovian$. C_t is therefore dependent on the preceding state C_{t-1} as well as X_t is the system (process) dynamics, this enables the system to evolve over time. Imagine that part of the system is being observed using a set of noisy sensors S_t = z(C_t, Y_t). Where z describes the relationship between the sensor observations S_t, and current system state C_t with sensor noise Y_t of the

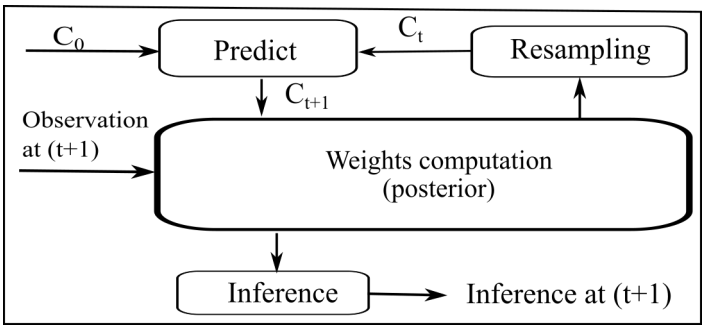


Fig. 1. The Workflow of the particle filter algorithm.

observation model. Hence, X_t and Y_t's unpredictability is believed to be known and captured using $pdfs$ [14].

The particle filter algorithm consists of three major steps including selection, prediction, and measurement.

In first, we construct a new particle set in the selection step by picking the particles from the previous particle set with the highest posterior probability. By passing an object via a dynamic model, the prediction step tries to forecast how its state will change. The dynamic model has two functions: firstly, it disperses the state density through stochastic diffusion, and secondly, it causes the state density to drift in a deterministic fashion. Each particle is updated according to the state model during prediction, which includes the insertion of random noise to simulate the effect of the noise on the state. The weight of each particle is re-evaluated based on the new data in the measurement stage, and we estimate likelihood probability. In summary, the following is a description of the particle filter:

(a) The following are the system and measurement equations:

$$A_{k+1} = g_k(A_k, W_k)\ and\ B_k = h_k(A_k, V_k) \tag{1}$$

Where the state variables vector A_k at time t are determined by a function g, with W_k and V_k being separate white noise processes with known $pdf's$. A function h connects the observations B_k with A_k.

(b) Let consider, the initial state of pdf $P(A_0)$ is known, produce N initial particles at random on the surface based on the pdf $P(A_0)$, where $A_{0,1} + (j = 1, ..., N)$ is the symbol for these particles. The user selects the parameter N as a compromise between computing effort and estimating precision.

(c) For k = 1, 2,..., do the following:

(1) Apply the known pdf with the known process equation of the process noise to the time propagation step to obtain the a priori particles a_k:

$$A^-_{k,j} = g_{k-1}(A^+_{k-1,j}, W^j_{k-1})(j = 1, \ldots N) \tag{2}$$

where each W^j_{k-1} noise vector is created at random using the known pdf of W_{k-1}.

(2) Calculate each particle's relative likelihood q_j, $A_{k,j}$ and conditional on B_k measurement. This is accomplished by calculating the pdf $p(B_K/A^-_{k,j})$ using the non-linear measurement equation and the measurement noise pdf.

(3) In the preceding step, scale the relative likelihoods as follows:

$$q_j = \frac{q_j}{\sum_{i=1}^{N}(q_i)} \quad (3)$$

Now add up all of the likelihood numbers is one.

(4) Using the relative likelihoods q_j, generate a collection of a posterior particles $A^+_{k,j}$. It is also called the re-sampling stage.

(5) Now that we have a collection of particles $A^+_{k,j}$ that are dispersed in accordance with the pdf of $p(A_k/B_k)$, any required statistical measure of the pdf can be calculated. However, calculating the mean and covariance is typically of significant concern.

III. Visual Object Tracking with Particle Filter

Visual tracking is a significant issue with many applications in surveillance, behaviour analysis, and human-computer interaction areas. It has a growing number of uses [17]. There are two classifications of visual tracking methods, deterministic and stochastic tracking. Mean their most familiar representatives are shift and particle filter, respectively [18].

However, particle filtering is the process of combining particles at a single location. A specific point into a single particle, resulting in weight for each particle to indicate the number of particles. It was created by combining various elements. Which eliminates the requirement for exact computations without skewing the results distribution of probabilities. As a result, the cost will be lower. In contrast, to limit the number of samples that must be processed, it is necessary to incur a computational charge that we investigate [19]. The basic architecture of particle filter-based visual object tracking is depicted in Fig. 2.

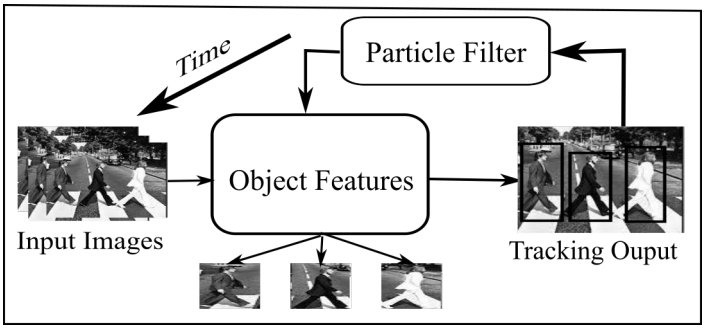


Fig. 2. Block diagram of the particle filter based visual object tracking.

Sequential importance sampling(SIS) is the most critical stage of the PF. At first, a set of particles is created, and then the crucial weights are assigned. Calculated in order to find the state's estimated worth. The majority of the consequences of SIS, after multiple iterations, a large number of particles are ignored, leaving only a small number of particles, possibly even a single particle [20].

SIS, unlike other methods, is not based on the Markov chain. One problem with SIS is that it is far too simple to get to a point where only a few (or one) have non-zero weights and almost all of the particles have weights of zero. These actual weights may differ significantly, leading to an inaccurate estimate. The SIS filter is affected by the so-called degeneracy problem. The SIS method is applicable to non-Bayesian computations as well, like figuring out the probability function in a case of missing data [21].

The following are the different versions of the particle filter [21], [22]:

(a) Local Linearization Particle Filter

(b) Sequential Importance Re-sampling (SIR) Filter

(c) Rejection Particle Filter

(d) Boosted Particle Filter

(e) Rao-Blackwellization

(f) Kernel Smoothing and Regularization

(g) Mixture Kalman Filter

(h) Interacting Multiple Model (IMM) Particle Filters

(i) MCMC Particle Filter

(j) Mixture Particle Filter.

(k) Auxiliary Particle Filter

(l) Unscented Particle Filter

Local Linearisation Particle Filter: The LLPF employs banks of Kalman filters with independent variants to generate credible densities using the most recent measurement.

Sequential Importance Re-sampling Filter: A generalization of particle filtering, SIR called Bootstrap Filter incorporates the re-sampling steps into the sequential importance sampling process.

Bootstrap and jackknife approaches serve as inspiration for sampling importance re-sampling (SIR). Description of a collection of computationally costly procedures built on sampling from observable data, the term "bootstrap technique" is used. The significance distribution is an approximation of the posterior distribution of states given the values from the previous stage, and its accuracy determines the effectiveness of the SIR algorithm. A SIR filter facilitates the examination of importance weights, importance density, substantial modularity, and generality [23].

SIR has the drawback of causing the particle swarm to "impoverish", as the name suggests. When the dynamic model's noise is tiny, many of the particle set's values will have the same value, which is what we mean by the term "the same". The second problem is that a large ensemble size is required to represent much regions in phase space (outliers occur), if the posterior distribution has long tails.

However, important Sampling techniques are used in both SIS and SIR filters. Re-sampling is always performed in the SIR filter (often between two critical sampling stages). In contrast, significance weights are calculated sequentially in the

SSIS filter and only need to be re sampled when necessary. This makes the SIS filter more computationally efficient.

Rejection Particle Filter: When we have a sufficient upper limit on the underlying distribution or density, the rejection particle becomes convenient. The rejection particle filter works better than the SIR filter when the distribution of proposals is preferable. The problem with using a rejection particle filter is that each time step takes a different amount of time to calculate.

Boosted particle Filter: There are two major issues that must be addressed in a traditional particle filter: setting up the particle filter and making sure the importance proposition is set up correctly. The BPF solves both problems by combining them with the AdaBoost detection algorithm [29].

Rao-Blackwellization: Since the process noise is formally zero in pure recursive estimation, the performance of a primary SIR-based particle filter is expected to be sub-optimal. Rao-Blackwellization is one method for increasing SIR efficiency. The fundamental idea behind this filter is to reduce the number of particles required to obtain the same level of accuracy as a standard PF filter.

Kernel Smoothing and Regularization: The problem of insufficient samples was fixed by these filters through the use of a ad hoc method called jittering. To smooth the posterior using a Gaussian kernel, Gaussian noise with a small amount is added to each re-sampled particle at each time step.

Mixture Kalman Filter: The MKF, which simulates a conditional Gaussian linear dynamic model, can be produced by Rao-Blackwellization and marginalization on a particle filter. The MKF represents the target distribution as a random mixture of normal distributions [27]. Primary benefit of MKF is the marginalization operation, which improves productivity.

Interacting Multiple Model Particle Filter: To forecast target behavior, the IMM comprises a range of models and adaptively chooses the models that are best for the current time step. It says that one of M modes best describes how a target behaves when moving with constant velocity, stopped, or accelerating. There is a probability attached to each mode [24].

Markov Chain Monte Carlo Particle Filter: In the MCMC sampling framework, samples are generated by a Markov chain that is ergodic, homogeneous, and reversible, and that has an invariant distribution. A faster rate of convergence than the particle filter is achieved by using this filter. Traditional re-sampling is replaced with MCMC sampling to prevent diversity loss and provide a re-sampling framework appropriate for pipeline processing [26].

Mixture Particle Filter: The MPF idea uses EKF/UKF as an approximation of a Gaussian proposal, and is comparable to the concepts of partitioned and stratified sampling.

Auxiliary Particle Filter: Pitt and Shephard proposed the APF in 1999 to address the tail observation density deficiencies of SIR [25]. As a result, calculating the APF requires more time. The tails of the distribution have low density because of sampling inefficiency and the unreliability of empirical prediction, downsides lack robustness against outliers. For the aforementioned issue, an auxiliary variable was added to the particle filter, resulting in a basic but more versatile and reputable framework with better performance than the SIR filter, although execution is not guaranteed.

Unscented Particle Filter: A particle filter with a UKF significance distribution constitutes the UPF. The proposal distribution in a conventional particle filter, such as CONDENSATION, is transition prior, but UPF was recently proposed to overcome the difficulties of wasting a large number of particles in the low likelihood region [28].

IV. Benchmark Dataset for Visual Object Tracking

A significant number of benchmark datasets have been generated for various visual object tracking applications. Most of them comprise video sequences with short-term (ST) visual object tracking, although long-term (LT) object tracking has been developed in recent years as illustrated in the Table I.

CAVIAR (CAVIAR2003) [39]: It includes two datasets containing several video sequences for various settings, such as persons walking alone or meeting with others. The first dataset contains films from $INRIA$ Labs' entry lobby, whereas the second is from a shopping mall hallway. Walking, browsing, relaxing, slumping or fainting, leaving luggage behind, people/groups walking together and splitting up, and two people arguing are among the scenarios covered in the first set. Each video sequence's ground truth is delivered in XML format and includes information such as bounding box locations and sizes, head and foot positions, and so on. Occlusion, appearance shifting, appearance, and disappearance are all issues addressed in this dataset.

Tracking and Surveillance Performance Evaluation (PETS2009 and PETS2012) [30],[38]: PETS-2009 and PETS-2012 are two of the most recent editions of PETS. PETS2017 has the most recent tracking video sequences, while PETS2012 datasets are identical to PETS2009, mainly aimed at surveillance applications. PETS 2009-S2 provides a dataset for people tracking in three crowd types: sparse, medium, and dense, with L1, L2, and L3 difficulty levels. Walking, running, and multiple flow merging are among the activities covered in the datasets. The goal is to monitor every person in the sequence, with the results being given as a 2D bounding box location for each person. There are difficulties in the dataset, such as occlusion, lighting changes, two people with the same appearance, etc.

Benchmark for online tracking (OTB-13 and OTB-15) [41]: OTB-13 has just 50 video sequences. The current version, OTB-15, is an expansion of OTB-13. The benchmark includes 50 and 100 video sequences, each with features. The datasets take into account eleven different attributes. Lighting, scale, occlusion, deformation, motion blur, fast motion, in-plane rotation, out-of-plane rotation, out-of-view, background clutters, and low resolution are some of them. The tracker's accuracy and location error are the two measures used to evaluate trackers. For assessing and comparing various tracker findings, success plots and precision plots are employed.

LITIV (LITIV) [31]: The dataset consists of people's heads being tracked in various situations (occlusions, many distractors, etc.). Ground facts containing the centers, widths,

and heights of tracked object bounding boxes are delivered with video sequences.

Benchmark for Multiple Object Tracking (MOT) [40]: The MOT challenge was initially released in 2015. It comprises annotated datasets, metrics for evaluating tracking methods, and a unified framework for multi-object tracking. MOT15, MOT16, MOT17, MOT20, and many other video sequences are available on the website. The CLEAR metrics(measures used in the classification of events, activities, and relationships workshops) and the collection of track quality measurements are utilized for evaluation [43]. The video sequences of ETH Central, TUD Stadtmitte, and TUD crossing can be found in MOT15.

Visual Object Tracking Challenge (VOT) [42]: The first tracker challenge for ST tracking was VOT2013. Every year since then, a challenge has been held. This paper makes use of VOT2016 [37], VOT2017 [36], and VOT2018 [35] datasets. It evaluates using two metrics. The first is accuracy, which is determined by calculating the overlap ratio between the tracker and ground truth bounding boxes, and the second is robustness, which is determined by tracking failure frequency. These measures have been replaced by the predicted average overlap measure [44] since VOT2015. The current version of VOT2020 includes video sequences for ST and LT tracking as well as an assessment platform.

A UAV Tracking Benchmark and Simulator (UAV123 and UAV20L) [32]: It features video sequences shot from an aerial perspective, with a subset of these sequences dedicated to long−term aerial tracking. Bounding boxes have been added to all sequences and twelve characteristics are used to evaluate trackers.

Single item tracking on a large scale (LaSOT) [34]: LaSOT is a long-term tracking benchmark for large-scale single object tracking. It includes video sequences in which the target vanishes and reappears. Annotations are supplied for each frame in the sequences. Lighting variation, full occlusion, partial occlusion, deformation, motion blur, fast motion, scale variation, camera motion, rotation, background clutter, poor resolution, viewpoint shift, out-of-view, and aspect ratio change are among the 14 qualities assigned to each sequence. It uses three measures for evaluation: accuracy, normalized precision, and success plots [33].

V. Comparison of Particle Filter System for Visual Object Tracking

Object tracking is a critical issue with the expansion of computer vision applications such as video surveillance, human-computer interaction, human behavior analysis, and so on. The following Table II provides a comparison of various algorithms for tracking visual objects using particle filters. Different versions of particle filters are employed with several features to obtain better tracking results.

TABLE I. Benchmark Dataset for Visual Object Tracking in Recent Years

Dataset	Type	Frame Per Seconds	Number of Attributes used for Evaluation	Number of Videos
CAVIAR (2003)	ST	25	–	44
PETS (2009, 2012)	ST	–	–	3(2009), 3(2012)
OTB (2013, 2015)	ST	30	11	50(2013), 100(2015)
LITIV (2014)	ST	15	–	4
MOT (2015, 2016, 2017, 2020)	ST	30(MOT15), 30(MOT16), 30(MOT17) 25(MOT20)	–	22(2015), 14(2016), 42(2017), 8(2020)
VOT (2013-2020)	ST, LT	30	–	16ST(2013), 25ST(2014), 60ST(2015-2017), 60ST & 35 LT(2018), 60ST & 50LT(2019 -2020)
UAV (2016)	ST, LT	30	12	123 ST & 20 LT
LaSOT (2019)	LT	30	14	1400

Note: ST means short-term tracking video, – means not mentioned and LT means long-term tracking video.

VI. Conclusion

This article reviewed particle filter-based approaches for visual object tracking and provided a succinct overview of related subjects. Due to its many practical applications, visual object tracking is one of the most researched computer vision topics. Numerous techniques have been developed for visual object tracking. Particle filter-based visual object tracking has been the subject of extensive research due to its capacity to deal with the complex dynamic environments of real life, random motions, many objects, and non-Gaussian sensor noises. In addition, the convolutional neural network and particle filter have been combined to create the most sophisticated method for visual object tracking. Several characteristics and classifiers that are frequently used with particle filters are provided. It was also found that combining multiple particle algorithms led to successful tracking outcomes and that each method has a unique advantage when considering the different visual object tracking circumstances.

Acknowledgment

We are very much thankful to Computer Vision and Intelligent Interfacing Lab (CVIIL) at Islamic University, Kushtia-7003, Bangladesh for providing all kinds of logistic and other support.

References

[1] Weng, S., Kuo, C. & Tu, S. Video object tracking using adaptive Kalman filter. *Journal Of Visual Communication And Image Representation*. **17**, 1190-1208 (2006)

[2] Ristic, B., Arulampalam, S. & Gordon, N. Beyond the Kalman filter: Particle filters for tracking applications. (Artech house,2003)

[3] Arulampalam, M., Maskell, S., Gordon, N. & Clapp, T. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions On Signal Processing*. **50**, 174-188 (2002)

[4] Comaniciu, D., Ramesh, V. & Meer, P. Kernel-based object tracking. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **25**, 564-577 (2003)

TABLE II. SUMMARIZING OF VISUAL OBJECT TRACKING USING VARIOUS PF BASED METHODS

SL No.	Paper Title	Year	Contribution	Remarks
1.	"Particle filter-based video object tracking using feature fusion in template partitions." [45]	2023	The research provides a novel approach to feature fusion that aims to improve the target and scene models in order to effectively partition the tracking template and capture the object's local attributes. The proposed strategy combines two features, namely the local binary pattern (LBP) and the mean RGB color features, to account for challenges such as shadows, dynamic background entities, adverse weather conditions, and variations in illumination.	The effectiveness of the suggested feature fusion strategy may exhibit variation depending on the various attributes of the tracked entities and the complex nature of the background scenes. However, the absence of an exhaustive examination of the trade-off between precision and computational efficacy may pose limitations in terms of execution in real-time and efficiency.
2.	"Particle Filter Based on the Harris Hawks Optimization Algorithm for Underwater Visual Tracking." [46]	2023	This study introduced a novel tracking algorithm, Harris-Hawks-optimized particle filters (HHOPF), to address the problem of low tracking accuracy in conventional particle filters caused by sample impoverishment. Using the Harris hawks algorithm, the proposed algorithm generates a nonlinear escape energy that effectively balances the exploration and exploitation processes. Consequently, the computational speed of tracking underwater targets is significantly increased.	The paper lacks a comprehensive analysis of the potential effects of diverse environmental conditions or scenarios on the algorithm's effectiveness. It does not address the difficulties or constraints associated with the use of the Harris hawks optimization technique for underwater visual tracking. In addition, the study lacks an exhaustive assessment of various cutting-edge tracking techniques, which may limit the practical significance of the findings.
3.	"Self-scale estimation of the tracking window merged with an adaptive particle filter tracker." [47]	2023	This research presents a new methodology for single-object tracking that integrates the particle filter algorithm and kernel distribution. This approach enables the tracking window to be updated based on changes in the object's scale, resulting in improved precision in localizing the object within a designated area. The proposed method addresses the difficulty of adapting tracking windows in real-time to account for variations in the object's appearance.	The work lacks a comprehensive examination of the computing complexity associated with the suggested approach. This aspect holds significance for real-time tracking applications that aim to tackle the difficulties of simultaneously tracking several objects.
4.	"A scale adaptive generative target tracking method based on modified particle filter." [48]	2023	This study presents a novel approach to enhance the accuracy and sample diversity of target tracking. The proposed method, named Quantum Particle Swarm Optimization (QPSO) and Adaptive Genetic Algorithm (QAPF), integrates an advanced particle filter (PF) algorithm with the mean shift technique. The objective is to optimize the robustness and intelligence of the traditional PF algorithm across various tracking conditions. By improving the position estimation of particles, the QAPF algorithm aims to enhance the overall performance of target tracking.	This study does not address the computational complexity or resource demands of the QAPF method. Additionally, the comparison of the QAPF algorithm with several state-of-the-art tracking techniques is limited in scope.
5.	"Robust Bayesian particle filter for space object tracking under severe uncertainty." [49]	2022	This paper proposes a particle filtering-based tracking system based on the concept of feature fusion in a complicated real-world environment with shadows, dynamic objects, adverse weather, and changing lighting. The tracking template comprises the local binary pattern (LBP) and the mean RGB color characteristics combined within a probabilistic framework.	Occlusion and concealment issues in dynamic scenes are among the limitations of the proposed system.
6.	"Multi-Feature Single Target Robust Tracking Fused with Particle Filter." [50]	2022	This work presents a credible particle filter with the correlation filtering framework-based multi-feature algorithm for tracking single targets in complex scenes. During feature extraction, depth features and manually extracted target object features are combined to train a tracking filter.	The proposed algorithm's high computational cost for utilizing depth features renders it unsuitable for real-time applications, necessitating additional research to improve its speed.
7.	"Minimax Monte Carlo object tracking." [51]	2022	The authors propose a new particle filter-based approach based on a minimax estimator combined with sequential Monte Carlo filtering for visual object tracking using a minimax strategy.	Only a few algorithms are used to compare performance.
8.	"An object tracking algorithm based on adaptive particle filtering and deep correlation multi-model." [52]	2022	The proposed object tracking algorithm solves the significant number of particles problem by utilizing a deep conventional correlation multi-model on adaptive particle filtering.	The proposed algorithm's performance decreases when the tracking object moves quickly.
9.	"Deep convolutional correlation iterative particle filter for visual tracking." [53]	2022	This paper proposes a novel iterative particle filter-based visual tracking framework based on the combination of a correlation filter and a deep convolutional neural network. In contrast, the iterative particle filter is used to select the correct particles and converge on the target position.	Compared to the other algorithms in the proposed work, the processing rate is lower.
10.	"Occluded object tracking using object-background prototypes and particle filter." [54]	2021	This article introduces an object-background prototypes-based discriminative model to address the shortage of valuable observations and the particle filter's efficient motion. A discriminative model is built using background and object knowledge and attempts to distinguish between the object, background, and any occluded parts of the object.	This algorithm performs well with occluded challenges in complex environments but is limited to other pose changes, illumination, and scale variations settings.
11.	"Multi-object tracking by mutual supervision of CNN and particle filter." [55]	2021	The authors proposed the mutual supervision of the trained CNN detector to identify the target bounding box in a traffic scene and the PF tracker to generate the multi-object tracking trajectory based on this identification.	The proposed algorithm can track rigid targets in controlled environments, like vehicles, but not overall targets like pedestrians.
12.	"Target tracking using a mean-shift occlusion aware particle filter." [56]	2021	This research paper aims to overcome the challenge of occlusion in visual tracking by including the mean-shift approach in a probabilistic filtering framework. This integration allows for a reduced particle set while still effectively maintaining the state probability density function of the model. The proposed approach demonstrates superior performance compared to state-of-the-art tracking algorithms on three benchmark datasets.	The research lacks a comprehensive examination of the computational expenses related to reducing the number of particles for tracking purposes and fails to offer precise quantitative comparisons with methods that employ larger quantities of particles.
13.	"Robust model adaption for color-based particle filter tracking with contextual information." [57]	2021	This study introduces a novel algorithm designed to improve the durability of color-based particle filters. Utilizing environmental data, the proposed algorithm dynamically updates both the scale of the tracker and the reference appearance model. It addresses the challenges posed by abrupt and significant changes in the appearance of the target during object tracking in video sequences.	The work lacks a comparative analysis of the proposed algorithm to a variety of cutting-edge trackers to provide a greater understanding of the algorithm's strengths and weaknesses. Further analysis that does not fully reflect the complexity and variability of real-world tracking scenarios could shed light on the robustness and generalizability of the proposed methodology.

SL No.	Paper Title	Year	Contribution	Remarks
14.	"Visual Vehicle Tracking via Deep Learning and Particle Filter." [58]	2021	The research presented a novel and efficient method for tracking multiple vehicles by combining the particle filter technique with the deep learning approach called You Only Look Once (YOLO). The authors perform an evaluation of a pragmatic tracking approach that employs the particle filter and the Bhattacharyya kernel to provide a feasible resolution for real-time tracking of multiple vehicles in different situations.	The study does not include a comprehensive examination of the computational complexity and lacks a comparison of the proposed approach with other well-recognized vehicle tracking techniques. As a result, the authors failed to address shortcomings or difficulties associated with using the YOLO detector for vehicle detection, such as how well it functions in different lighting conditions or in occlusion-prone situations.
15.	"Tracking and grasping of moving target based on accelerated geometric particle filter on the colored image." [59]	2021	This study presents a novel approach for tracking and retrieval of objects in unpredictable motion by employing a geometric particle filter tracker. The reduction in computational expenditure is accomplished by employing edge detection and morphological dilation methodologies. Furthermore, the tracking algorithm integrates HSV image features rather than grayscale image features, thereby improving its capacity to adjust to changes in illumination conditions.	A novel method for tracking and grasping moving objects is presented in this paper. Nevertheless, it does not analyze the handling of complex background scenarios or object occlusion, which are common challenges in tracking and grasping tasks in the real world.
16.	"A genetic optimization re-sampling based particle filtering algorithm for indoor target tracking." [60]	2021	This study presents a novel resampling technique intended to address the issue of particle erosion commonly observed with conventional resampling methods. The distribution of resampled particles is optimized by employing five operators, namely selection, roughening, classification, crossover, and mutation. These operators are employed to decrease particle depletion and improve particle positioning precision.	The study focuses primarily on indoor target tracking via wireless sensor networks. Uncertain is the extent to which the proposed algorithm would function effectively in outdoor settings or with various sensor types.
17.	"Particle filter and entropy-based measure for tracking of video objects." [61]	2021	The authors present a novel approach based on the particle filter for tracking objects in outdoor and indoor video sequences, using an entropy-based time motion searching model for selecting the particles to detect objects.	The proposed approach includes a performance comparison.
18.	"Deep convolutional likelihood particle filter for visual tracking." [62]	2021	This paper proposes a novel particle filter for visual tracking based on convolutional-correlation particle filters that estimate likelihood distributions from correlation response maps.	The weakness of the proposed method is that it does not fully explain all parts of the algorithm.
19.	"Infrared target tracking based on improved particle filtering." [63]	2021	To solve the infrared image particle degradation problem, this paper presents a combination of genetic algorithms and extended Kalman filter-based particle filter tracking with improved performance for infrared target tracking.	Utilizing a genetic algorithm and an extended Kalman filter increases the computational complexity of the proposed particle filter.
20.	"Visual object tracking via iterative ant particle filtering." [64]	2020	The author provides a particle filter-based tracking method with discriminative model-based object background prototypes for identifying objects and occluded portions of objects. The discriminative model uses prior knowledge of object and background classes to differentiate between three categories: object, background, and occluded object parts.	This method overlooked other relevant variables such as lighting variance, pose change, and scale change.
21.	"Tracking objects based on multiple particle filters for multipart combined moving directions information." [65]	2020	This study introduces the mutual supervision of the PF tracker and CNN detector to create the multi-object tracking trajectory, whereas a trained CNN is used to identify the bounding box of the detected target in a traffic scenario.	The main limitation of this study occurs under controlled conditions, such as tracking people from a traffic scene.
22.	"Occlusion robust object tracking with modified particle filter framework." [66]	2020	The authors present a particle filter that combines the mean-shift approach into a probabilistic filtering framework while using fewer particles to maintain many modes' state probability density function. The reference target and candidate location correlation coefficient, however, detect occlusion.	The proposed MSOAPF technique includes occlusion and fast motion challenges but fails in extensive situations.
23.	"A hybrid algorithm based on particle filter and genetic algorithm for target tracking." [67]	2020	This paper proposes a crew search optimization-based robust particle filter re-sampling method for overcoming low-performing particles identified as outliers by incorporating multi-cue extracted for each evaluated particle.	This approach is computationally demanding as an outlier detection is used to categorize unimportant particles.
24.	"Robust object tracking with crow search optimized multi-cue particle filter." [68]	2020	In this study, an observer marks an object at the beginning of a video sequence to reduce the image size by selecting particles with the highest weights to evolve a genetic algorithm known as the Reduced Particle Filter Genetic Algorithm during the re-sampling phase of the particle filter.	This algorithm's performance is not calculated and compared conclusively with other state-of-the-art algorithms.
25.	"Intelligent visual object tracking with particle filter based on Modified Grey Wolf Optimizer." [69]	2019	In this paper, the author proposes a visual object tracking system based on the Modified Grey Wolf Optimizer (MGWO) to address the issue of many particles in the particle filter (PF). Before re-sampling, the PF particles were optimized using a new variant of Grey Wolf Optimizer known as Modified Grey Wolf Optimizer (MGWO).	The proposed MGWO-based PF is not fully explained and is only compared to a few algorithms.
26.	"Multiple pedestrian tracking by combining particle filter and network flow model." [70]	2019	This paper suggested a novel particle filter tracking system for multiple pedestrians based on network flow models. To compensate for the problem of long-term occlusion, the proposed model combines local and global data association strategies.	The weakness of the proposed method is that both the computational complexity and the number of particles increase as the number of objects increases.
27.	"Multi-feature fusion in particle filter framework for visual tracking." [71]	2019	The authors applied the color distribution, and KAZE features with the particle filter (PF) to track a target in a video sequence of complex environments. The proposed multi-feature fusion-based PF utilized the Bhattacharya coefficient for the color distribution model as a similarity metric. At the same time, KAZE features utilized the Nearest Neighbor Distance ratio for matching feature points.	In PF experiments, only color and KAZE characteristics are employed. A set of features that accurately represents the target in one environment may not do so in another.
28.	"Particle filter re-detection for visual tracking via correlation filters." [72]	2019	This paper proposes a novel particle filter re-detection tracker with a correlation filters framework (CFPFT) to solve the problem of accurate object localization.	Long-term occlusions may still cause the tracking method to fail, so a strategy that re-detects the target and reuses the correlation filter tracker was to be implemented in real-world applications.
29.	"Robust object tracking via integration of particle filtering with deep detection." [73]	2019	This paper proposes a video object tracker with variable-rate color particle filtering that incorporates two innovations. First, a deep region proposal network was used to select the bounding box based on the dynamic prediction of particle filtering and a fusion integrating the particle filter and a deep object detector to enhance object tracking accuracy.	Implementation of the proposed method is slower than the conventional particle filter method, with only minor improvements.

SL No.	Paper Title	Year	Contribution	Remarks
30.	"Visual tracking based on adaptive interacting multiple model particle filter by fusing multiples cues." [74]	2018	This paper presents a robust tracking system based on the fusion of a particle filter (PF) with interacting multiple models (IMM) to overcome various severe challenges, such as camera motion, fast motion, background clutter, or target appearance changes.	The performance of the proposed algorithm decreases while the tracked object moves quickly.
31.	"Chaotic particle filter for visual object tracking." [75]	2018	The authors introduced a chaotic particle filter (PF) based on global motion to enhance the performance of PFs. In two steps of chaos theory, estimation of the object's movement is used to identify its position over frames, followed by the color-based particle filter technique to find the object in the local region.	The chaotic PF method outperforms the conventional PF method and occasionally the tracking by detection techniques, but only color features are used in all studies.
32.	"Robust object tracking via part-based correlation particle filter." [76]	2018	This paper proposes a part-based correlation particle filter method for reliable visual tracking. In a particle filter architecture, correlation filters manage target parts and accurately represent the target's appearance by employing overlapping local parts of varying sizes and locations.	The main downside of the method is that it works better when tracking nonrigid objects with changing appearances, but it disregards other relevant difficulties.
33.	"A box particle filter method for tracking multiple extended objects." [77]	2018	This paper proposed a box particle filter-based extended multiple object tracking technique to address the challenging problem of multiple object tracking caused by data association. This research also highlights how interval-based techniques can manage data association concerns effectively while minimizing computational complexity.	The validation of the proposed approach is limited to laser rangefinder sensor data in real time and needs to be expanded and validated in more generic application settings.
34.	"Multi-task correlation particle filter for robust object tracking." [78]	2017	This study presented a particle filter based on a multitask correlation named MCPF for effective visual object tracking. In contrast, correlation filters jointly learn the interdependencies across numerous features from the multitask correlation filter (MCF).	Using a multitask correlation filter increases the computational cost of the proposed MCPF.
35.	"Particle Filter Object Tracking Based on Color Histogram and Gabor Filter Magnitude." [79]	2017	The authors present a novel particle filter-based tracking method that uses the combination of Gabor filter features and the color histogram of the detected object, obtained by background subtraction, to determine the likelihood of the observation and a state space model based on the Box-Muller transformation.	The proposed method's main drawback is that the number of particles increases when the object's color histogram increases, causing computational complexity.
36.	"Correlation particle filter for visual tracking." [80]	2017	The authors propose a robust correlation particle filter (CPF) for visual tracking by combining the strengths of a correlation filter and a particle filter.	The performance metric of the algorithm is not computed comprehensively.
37.	"Deep convolutional particle filter for visual tracking." [81]	2017	The integration of a particle filter and deep CNN has been proposed in which the particle filter predicts the target position using motion mode at each frame, and the HCFT-CNN adjusts the target position using the surrounding particles of the anticipated position. Finally, the CNN correlation map was used to calculate the current frame's particle weight and target position.	The presented paper contains no accurate experimental results.
38.	"Top-down visual attention integrated particle filter for robust object tracking." [82]	2016	This study blends frequency analysis into a particle filter-based computational of top-down visual attention to address the challenges of rapid motion and long-term obstruction.	The tracker's performance suffers slightly if the object appearance changes and if descriptors calculated directly from feature maps are used instead of color histograms, making it more coherent and faster.
39.	"Adaptive Cell-Size HoG Based Object Tracking with Particle Filter." [83]	2016	The authors proposed an Adaptive cell-size HoG (acHoG)-based Particle Filter Tracking (PFT) algorithm to handle the repeating feature extraction problem by applying Adaptive cell size to HoG.	The main drawback of these approaches is that they reduce tracker performance due to failure detection.
40.	"Object-tracking based on particle filter using particle swarm optimization with density estimation." [84]	2016	The authors analyze the problems of particle filter tracking degeneracy and impoverishment degradation in the Bayesian particle filter tracking framework. Particle Swarm Optimization (PSO) is employed as a sampling mechanism to address these two issues.	The proposed solution did not handle other critical challenges such as backdrop adaption, scene tracking in a cluttered environment, and object interactions.
41.	"Real-time and model-free object tracking using particle filter with joint color-spatial descriptor." [85]	2015	The authors incorporate the Joint Color-Spatial Descriptor (JCSD) and particle filtering to develop a multiple model-free object tracking technique that evaluates the hypothesis step within a particle filtering framework.	However, the method's accuracy decreases with low computational power and must be implemented with GPU for high accuracy and real-time applications.
42.	"Single object tracking via a robust combination of particle filter and sparse representation." [86]	2015	This author presented a particle filter-based tracking system that uses a particle filter and reversed sparse representation (RC-PFRSR) comprehensive combination to reduce drifting and increase tracking robustness.	In this study, the proposed system tracks only a single visual object and not multiple visual objects.
43.	"Particle filter with occlusion handling for visual tracking." [87]	2015	The authors proposed a particle filter system with a patch-based appearance model for occlusion handling, which includes two main components: color and motion vector, to address the challenge of visual tracking with three critical components including feature extraction, particle weighting, and occlusion handling.	The key drawback of the strategy is that it only worked for occluded environments as opposed to other constraints in visual object tracking.
44.	"Intelligent video target tracking using an evolutionary particle filter based upon improved cuckoo search." [88]	2014	This paper aims to incorporate an evolutionary particle filter with an improved cuckoo search algorithm to overcome the sample impoverishment problem of a generic particle filter in real-time video object tracking. Regarding tracking performance, the proposed algorithm outperforms the PSO particle filter and generic particle filter based on an improved cuckoo particle filter.	The proposed system is adequate for tracking a single object, not multiple objects, or in ambiguous or uncertain environments. Furthermore, this tracking framework necessitates additional cues to handle dynamic object shapes.
45.	"Using local saliency for object tracking with particle filters." [89]	2014	Instead of employing the saliency map of the entire image, the authors proposed a tracking algorithm that estimates the saliency map by extracting salient regions based on visual attention from the particle areas and the target area of each particle.	The proposed algorithm resolves particle filter divergence for targets with similar characteristics. However, tracking errors are inevitable when target and background colors are similar.
46.	"Abrupt motion tracking using a visual saliency embedded particle filter." [90]	2014	This article illustrates how to solve the problem of abrupt motion by combining a particle filter tracking system with an improved visual saliency model. In addition, it is possible to recover a lost tracking object by detecting its region in salient regions.	In this proposed work, many tracking aspects under challenging conditions, such as when the target object is visually similar to the background or when its appearance abruptly changes, are not considered.

SL No.	Paper Title	Year	Contribution	Remarks
47.	"Object tracking from image sequences using adaptive models in the fuzzy particle filter." [91]	2013	This paper addresses issues caused by unexpected events. In the proposed work, a fuzzy particle filter's process and observation noises are considered fuzzy variables. The fuzzy particle filter is equipped with two adaptive models that enhance tracking performance: adaptive AR and appearance mixture. The adaptive AR calculation is adopted as the state transition model, and recursive AR determines the optimal parameters over time to enhance state prediction. Additionally, the observation model, the adaptive appearance mixture model, consists of three Gaussian components (W, S, and E).	This method did not consider object illumination or occlusion when tracking multiple objects.
48.	"A combined Color-Correlation Visual Model for Object Tracking using Particle Filters." [92]	2013	The proposed work utilizes a particle filter to track semi-rigid video sequence objects. In addition, the MOSSE correlation filter and color histogram combination create a robust feature descriptor of the object that outperforms for tracking fast-moving objects in a complex environment in real-time applications with a small number of particles.	This paper's limitation is that the required computing power for the tracking system depends not only on the size of the patches but also on the number of particles. In addition, the system must investigate sampling inefficiencies and maximize particle utilization.
49.	"Efficient visual tracking using particle filter with incremental likelihood calculation." [93]	2012	This paper proposes an improved real-time particle filter in which the weight of each particle is computed using incremental likelihood. However, the proposed particle filter-based tracking method performs surprisingly well on hardware platforms with limited resources.	The proposed tracking method's effectiveness diminishes as the tracked object's speed increases. Therefore, it is only effective when an object moves at a reasonable speed; this is the trade-off between the proposed approach's accuracy and the object's speed.
50.	"A compact association of particle filtering and kernel-based object tracking." [94]	2012	The main goal of this paper is to combine kernel-based object tracking (KBOT) and particle filtering (PF) to develop a more reliable visual tracking method (PF). In addition, the question of what types of particles are appropriate for using KBOT to refine their position states for more accuracy is also addressed, and a two-stage solution is proposed.	There are still some issues that demand more research. The disadvantage of the proposed algorithm's use of color histograms as object descriptors for a fair comparison is that it is unsuitable for dealing with varying lighting conditions.
51.	"Hierarchical Kalman-particle filter with adaptation to motion changes for object tracking." [95]	2011	The authors examined the combination of the particle filter and subspace representation and successfully applied it to tracking algorithms, such as the Eigen-tracking technique. Their combination has shown improved performance in terms of accuracy and robustness despite requiring a relatively small number of particles.	Although the proposed method works with greater accuracy and precision, it is computationally expensive. Moreover, in the proposed technique, the particle filter handles only nonlinear motion locally, while the Kalman filter deals with linear motion globally.
52.	"Robust visual object tracking using multi-mode anisotropic mean shift and particle filters." [96]	2011	This study presents a novel tracking technique based on multi-mode anisotropic mean shift and particle filters. The method performs online learning of reference objects, is more resistant to objects' dynamic shape and appearance, and requires a small number of particles.	The proposed system necessarily requires accelerated computation and empirical parameter selection.

[5] Isard, M. & Blake, A. Condensation—conditional density propagation for visual tracking. *International Journal Of Computer Vision*. **29**, 5-28 (1998)

[6] Walia, G., Raza, S., Gupta, A., Asthana, R. & Singh, K. A novel approach of multi-stage tracking for precise localization of target in video sequences. *Expert Systems With Applications*. **78** pp. 208-224 (2017)

[7] Wang, D., Lu, H., Xiao, Z. & Chen, Y. Fast and effective color-based object tracking by boosted color distribution. *Pattern Analysis And Applications*. **16**, 647-661 (2013)

[8] Saad, E., Bardawiny, E., Ali, H. & Shawky, N. MCMC Particle Filter Using New Data Association Technique with Viterbi Filtered Gate Method for Multi-Target Tracking in Heavy Clutter. *International Journal Of Advanced Computer Science And Applications*. **2** (2011)

[9] Islam, M., Oh, C. & Lee, C. Video based moving object tracking by particle filter. *International Journal Of Signal Processing, Image Processing And Pattern*. **2** (2009)

[10] Nummiaro, K., Koller-Meier, E. & Van Gool, L. An adaptive color-based particle filter. *Image And Vision Computing*. **21**, 99-110 (2003)

[11] Kitagawa, G. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal Of Computational And Graphical Statistics*. **5**, 1-25 (1996)

[12] Vo, B., Vo, B. & Cantoni, A. Bayesian filtering with random finite set observations. *IEEE Transactions On Signal Processing*. **56**, 1313-1326 (2008)

[13] Nummiaro, K., Koller-Meier, E., Van Gool, L. & Others A color-based particle filter. *First International Workshop On Generative-Model-Based Vision*. **2002** pp. 01 (2002)

[14] Forte, D. & Srivastava, A. Resource-aware architectures for particle filter based visual target tracking. *2011 International Green Computing Conference And Workshops*. pp. 1-6 (2011)

[15] Islam, M., Oh, C. & Lee, C. Real time moving object tracking by particle filter. *International Symposium On Computer Science And Its Applications*. pp. 347-352 (2008)

[16] Ly, Q., Nguyen, T. & Others A New Framework of Moving Object Tracking based on Object Detection-Tracking with Removal of Moving Features. *International Journal Of Advanced Computer Science And Applications*. **11** (2020)

[17] Liu, H. & Sun, F. Efficient visual tracking using particle filter with incremental likelihood calculation. *Information Sciences*. **195** pp. 141-153 (2012)

[18] Zhang, B., Tian, W. & Jin, Z. Joint tracking algorithm using particle filter and mean shift with target model updating. *Chinese Optics Letters*. **4**, 569-572 (2006)

[19] Rao, G. & Satyanarayana, C. Visual object target tracking using particle filter: a survey. *International Journal Of Image, Graphics And Signal Processing*. **5**, 1250 (2013)

[20] Wei, Q., Xiong, Z., Li, C., Ouyang, Y. & Sheng, H. A robust approach for multiple vehicles tracking using layered particle filter. *AEU-international Journal Of Electronics And Communications*. **65**, 609-618 (2011)

[21] Chen, Z. & Others Bayesian filtering: From Kalman filters to particle filters, and beyond. *Statistics*. **182**, 1-69 (2003)

[22] Ristic, B., Arulampalam, S. & Gordon, N. Beyond the Kalman filter: Particle filters for tracking applications. (Artech house,2003)

[23] Simon, E., Ros, L., Hijazi, H., Fang, J., Gaillot, D. & Berbineau, M. Joint carrier frequency offset and fast time-varying channel estimation for MIMO-OFDM systems. *IEEE Transactions On Vehicular Technology*. **60**, 955-965 (2011)

[24] Kreucher, C., Hero, A. & Kastella, K. Multiple model particle filtering for multitarget tracking. *Proceedings Of The Twelfth Annual Workshop On Adaptive Sensor Array Processing*. **770** (2004)

[25] Pitt, M. & Shephard, N. Filtering via simulation: Auxiliary particle filters. *Journal Of The American Statistical Association*. **94**, 590-599 (1999)

[26] Wang, D., Zhang, Q. & Morris, J. Distributed Markov Chain Monte Carlo kernel based particle filtering for object tracking. *Multimedia Tools And Applications*. **56**, 303-314 (2012)

[27] Chen, R. & Liu, J. Mixture kalman filters. *Journal Of The Royal Statistical Society: Series B (Statistical Methodology)*. **62**, 493-508 (2000)

[28] Rui, Y. & Chen, Y. Better proposal distributions: Object tracking using unscented particle filter. *Proceedings Of The 2001 IEEE Computer Society Conference On Computer Vision And Pattern Recognition. CVPR 2001*. **2** pp. II-II (2001)

[29] Verma, R. & Verma, A. Particle Filter Based Visual Tracking: A Review.

[30] Ferryman, J. & Shahrokni, A. Pets2009: Dataset and challenge. *2009 Twelfth IEEE International Workshop On Performance Evaluation Of Tracking And Surveillance*. pp. 1-6 (2009)

[31] Verma, R. & Verma, A. Particle Filter Based Visual Tracking: A Review.

[32] Mueller, M., Smith, N. & Ghanem, B. A benchmark and simulator for uav tracking. *European Conference On Computer Vision*. pp. 445-461 (2016)

[33] Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C. & Ling, H. Lasot: A high-quality benchmark for large-scale single object tracking. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 5374-5383 (2019)

[34] Verma, R. & Verma, A. Particle Filter Based Visual Tracking: A Review.

[35] Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., ˇCehovin Zajc, L., Vojir, T., Bhat, G., Lukezic, A., Eldesokey, A. & Others The sixth visual object tracking vot2018 challenge results. *Proceedings Of The European Conference On Computer Vision (ECCV) Workshops*. pp. 0-0 (2018)

[36] Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Cehovin, L., Fernandez, G., Vojir, T., Hager, G., Nebehay, G. & Pflugfelder, R. The visual object tracking vot2015 challenge results. *Proceedings Of The IEEE International Conference On Computer Vision Workshops*. pp. 1-23 (2015)

[37] Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Cehovin, L., Fernández, G. & Vojir, T. Hager, and et al. The visual object tracking vot2016 challenge results. *ECCV Workshop*. **2**, 8 (2016)

[38] Ferryman, J. & Shahrokni, A. Pets2009: Dataset and challenge. *2009 Twelfth IEEE International Workshop On Performance Evaluation Of Tracking And Surveillance*. pp. 1-6 (2009)

[39] Negri, P. & Lotito, P. Pedestrian detection on CAVIAR dataset using a movement feature space. *XIII Argentine Symposium On Technology (AST 2012)(XLII JAIIO, La Plata, 27 Y 28 De Agosto De 2012)*. (2012)

[40] Milan, A., Leal-Taixé, L., Reid, I., Roth, S. & Schindler, K. MOT16: A benchmark for multi-object tracking. *ArXiv Preprint ArXiv:1603.00831*. (2016)

[41] Wu, Y., Lim, J. & Yang, M. Online object tracking: A benchmark. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 2411-2418 (2013)

[42] Maresca, M. & Petrosino, A. Clustering local motion estimates for robust and efficient object tracking. *European Conference On Computer Vision*. pp. 244-253 (2014)

[43] Bernardin, K. & Stiefelhagen, R. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal On Image And Video Processing*. **2008** pp. 1-10 (2008)

[44] Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Pflugfelder, R., Kamarainen, J., ˇCehovin Zajc, L., Drbohlav, O., Lukezic, A., Berg, A. & Others The seventh visual object tracking vot2019 challenge results. *Proceedings Of The IEEE/CVF International Conference On Computer Vision Workshops*. pp. 0-0 (2019)

[45] Panda, J. & Nanda, P. Particle filter-based video object tracking using feature fusion in template partitions. *The Visual Computer*. **39**, 2757-2779 (2023)

[46] Yang, J., Yao, Y. & Yang, D. Particle Filter Based on Harris Hawks Optimization Algorithm for Underwater Visual Tracking. *Journal Of Marine Science And Engineering*. **11**, 1456 (2023)

[47] Youssef, A., Aouatif, A., Bouchra, N. & Mohammed, N. Self-scale estimation of the tracking window merged with adaptive particle filter tracker. *International Journal Of Electrical And Computer Engineering*. **13**, 374 (2023)

[48] Yuqi, X., Yongjun, W. & Fan, Y. A scale adaptive generative target tracking method based on modified particle filter. *Multimedia Tools And Applications*. pp. 1-21 (2023)

[49] Greco, C. & Vasile, M. Robust Bayesian particle filter for space object tracking under severe uncertainty. *Journal Of Guidance, Control, And Dynamics*. **45**, 481-498 (2022)

[50] Liu, C., Ibrayim, M. & Hamdulla, A. Multi-Feature Single Target Robust Tracking Fused with Particle Filter. *Sensors*. **22**, 1879 (2022)

[51] Lim, J., Park, J. & Park, H. Minimax Monte Carlo object tracking. *The Visual Computer*. pp. 1-16 (2022)

[52] Duan, K. & Yu, Y. An object tracking algorithm based on adaptive particle filtering and deep correlation multi-model. *Third International Conference On Electronics And Communication; Network And Computer Technology (ECNCT 2021)*. **12167** pp. 656-664 (2022)

[53] Jalil Mozhdehi, R. & Medeiros, H. Deep convolutional correlation iterative particle filter for visual tracking. *Computer Vision And Image Understanding*. **222** pp. 103479 (2022)

[54] Mondal, A. Occluded object tracking using object-background prototypes and particle filter. *Applied Intelligence*. **51**, 5259-5279 (2021)

[55] Xia, Y., Qu, S., Goudos, S., Bai, Y. & Wan, S. Multi-object tracking by mutual supervision of CNN and particle filter. *Personal And Ubiquitous Computing*. **25**, 979-988 (2021)

[56] Bhat, P., Subudhi, B., Veerakumar, T., Di Caterina, G. & Soraghan, J. Target tracking using a mean-shift occlusion aware particle filter. *IEEE Sensors Journal*. **21**, 10112-10121 (2021)

[57] Xiao, J. & Oussalah, M. Robust model adaption for colour-based particle filter tracking with contextual information. *Journal Of Visual Communication And Image Representation*. **79** pp. 103270 (2021)

[58] Ait Abdelali, H., Bourja, O., Haouari, R., Derrouz, H., Zennayi, Y., Bourzex, F. & Oulad Haj Thami, R. Visual vehicle tracking via deep learning and particle filter. *Advances On Smart And Soft Computing*. pp. 517-526 (2021)

[59] Gong, Z., Qiu, C., Tao, B., Bai, H., Yin, Z. & Ding, H. Tracking and grasping of moving target based on accelerated geometric particle filter on colored image. *Science China Technological Sciences*. **64**, 755-766 (2021)

[60] Zhou, N., Lau, L., Bai, R. & Moore, T. A genetic optimization resampling based particle filtering algorithm for indoor target tracking. *Remote Sensing*. **13**, 132 (2021)

[61] Panda, J. & Nanda, P. Particle filter and entropy-based measure for tracking of video objects. *Green Technology For Smart City And Society*. pp. 339-354 (2021)

[62] Mozhdehi, R. & Medeiros, H. Deep convolutional likelihood particle filter for visual tracking. *Advances In Computer Vision And Computational Biology*. pp. 27-38 (2021)

[63] Hu, Z. & Su, Y. Infrared target tracking based on improved particle filtering. *International Journal Of Pattern Recognition And Artificial Intelligence*. **35**, 2154015 (2021)

[64] Wang, F., Wang, Y., He, J., Sun, F., Li, X. & Zhang, J. Visual object tracking via iterative ant particle filtering. *IET Image Processing*. **14**, 1636-1644 (2020)

[65] Ha, N., Shimizu, I. & Others Tracking objects based on multiple particle filters for multipart combined moving directions information. *Computational Intelligence And Neuroscience*. **2020** (2020)

[66] Gupta, S., Bhuyan, M. & Sasmal, P. Occlusion robust object tracking with modified particle filter framework. *2020 IEEE Applied Signal Processing Conference (ASPCON)*. pp. 257-261 (2020)

[67] Moghaddasi, S. & Faraji, N. A hybrid algorithm based on particle filter and genetic algorithm for target tracking. *Expert Systems With Applications*. **147** pp. 113188 (2020)

[68] Walia, G., Kumar, A., Saxena, A., Sharma, K. & Singh, K. Robust object tracking with crow search optimized multi-cue particle filter. *Pattern Analysis And Applications*. **23**, 1439-1455 (2020)

[69] Narayana, M., Nenavath, H., Chavan, S. & Rao, L. Intelligent visual object tracking with particle filter based on Modified Grey Wolf Optimizer. *Optik*. **193** pp. 162913 (2019)

[70] Cui, Y., Zhang, J., He, Z. & Hu, J. Multiple pedestrian tracking by combining particle filter and network flow model. *Neurocomputing*. **351** pp. 217-227 (2019)

[71] Bhat, P., Subudhi, B., Veerakumar, T., Laxmi, V. & Gaur, M. Multi-feature fusion in particle filter framework for visual tracking. *IEEE Sensors Journal*. **20**, 2405-2415 (2019)

[72] Yuan, D., Lu, X., Li, D., Liang, Y. & Zhang, X. Particle filter re-detection for visual tracking via correlation filters. *Multimedia Tools And Applications*. **78**, 14277-14301 (2019)

[73] Gurkan, F., Gunsel, B. & Ozer, C. Robust object tracking via integration of particle filtering with deep detection. *Digital Signal Processing*. **87** pp. 112-124 (2019)

[74] Dhassi, Y. & Aarab, A. Visual tracking based on adaptive interacting multiple model particle filter by fusing multiples cues. *Multimedia Tools And Applications*. **77**, 26259-26292 (2018)

[75] Firouznia, M., Faez, K., Amindavar, H. & Koupaei, J. Chaotic particle filter for visual object tracking. *Journal Of Visual Communication And Image Representation*. **53** pp. 1-12 (2018)

[76] Wang, N., Zhou, W. & Li, H. Robust object tracking via part-based correlation particle filter. *2018 IEEE International Conference On Multimedia And Expo (ICME)*. pp. 1-6 (2018)

[77] De Freitas, A., Mihaylova, L., Gning, A., Schikora, M., Ulmke, M., Angelova, D. & Koch, W. A box particle filter method for tracking multiple extended objects. *IEEE Transactions On Aerospace And Electronic Systems*. **55**, 1640-1655 (2018)

[78] Zhang, T., Xu, C. & Yang, M. Multi-task correlation particle filter for robust object tracking. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 4335-4343 (2017)

[79] Lahraichi, M., Housni, K. & Mbarki, S. Particle Filter Object Tracking Based on Color Histogram and Gabor Filter Magnitude. *Proceedings Of The 2nd International Conference On Big Data, Cloud And Applications*. pp. 1-5 (2017)

[80] Zhang, T., Liu, S., Xu, C., Liu, B. & Yang, M. Correlation particle filter for visual tracking. *IEEE Transactions On Image Processing*. **27**, 2676-2687 (2017)

[81] Mozhdehi, R. & Medeiros, H. Deep convolutional particle filter for visual tracking. *2017 IEEE International Conference On Image Processing (ICIP)*. pp. 3650-3654 (2017)

[82] Li, W., Wang, P. & Qiao, H. Top–down visual attention integrated particle filter for robust object tracking. *Signal Processing: Image Communication*. **43** pp. 28-41 (2016)

[83] Kim, J. & Kim, K. Adaptive Cell-Size HoG Based Object Tracking with Particle Filter. *Contemporary Engineering Sciences*. **9**, 539-545 (2016)

[84] Xia, G. & Ludwig, S. Object-tracking based on particle filter using particle swarm optimization with density estimation. *2016 IEEE Congress On Evolutionary Computation (Cec)*. pp. 4151-4158 (2016)

[85] Li, S., Koo, S. & Lee, D. Real-time and model-free object tracking using particle filter with joint color-spatial descriptor. *2015 IEEE/RSJ International Conference On Intelligent Robots And Systems (IROS)*. pp. 6079-6085 (2015)

[86] Yi, S., He, Z., You, X. & Cheung, Y. Single object tracking via robust combination of particle filter and sparse representation. *Signal Processing*. **110** pp. 178-187 (2015)

[87] Lin, S., Lin, J. & Chuang, C. Particle filter with occlusion handling for visual tracking. *IET Image Processing*. **9**, 959-968 (2015)

[88] Walia, G. & Kapoor, R. Intelligent video target tracking using an evolutionary particle filter based upon improved cuckoo search. *Expert Systems With Applications*. **41**, 6315-6326 (2014)

[89] Yuan, Y., Gao, C., Liu, Q., Wang, J. & Zhang, C. Using local saliency for object tracking with particle filters. *2014 IEEE International Conference On Signal Processing, Communications And Computing (ICSPCC)*. pp. 388-393 (2014)

[90] Su, Y., Zhao, Q., Zhao, L. & Gu, D. Abrupt motion tracking using a visual saliency embedded particle filter. *Pattern Recognition*. **47**, 1826-1834 (2014)

[91] Yoon, C., Cheon, M. & Park, M. Object tracking from image sequences using adaptive models in the fuzzy particle filter. *Information Sciences*. **253** pp. 74-99 (2013)

[92] Centir, M., Fragneto, P., Denaro, D., Rossi, B. & Marchisio, C. A combined color-correlation visual model for object tracking using particle filters. *2013 8th International Symposium On Image And Signal Processing And Analysis (ISPA)*. pp. 153-158 (2013)

[93] Liu, H. & Sun, F. Efficient visual tracking using particle filter with incremental likelihood calculation. *Information Sciences*. **195** pp. 141-153 (2012)

[94] Yao, A., Lin, X., Wang, G. & Yu, S. A compact association of particle filtering and kernel-based object tracking. *Pattern Recognition*. **45**, 2584-2597 (2012)

[95] Yin, S., Na, J., Choi, J. & Oh, S. Hierarchical Kalman-particle filter with adaptation to motion changes for object tracking. *Computer Vision And Image Understanding*. **115**, 885–900 (2011)

[96] Khan, Z., Gu, I. & Backhouse, A. Robust visual object tracking using multi-mode anisotropic mean shift and particle filters. *IEEE Transactions On Circuits And Systems For Video Technology*. **21**, 74-87 (2011)

Deep Speech Recognition System Based on AutoEncoder-GAN for Biometric Access Control

Oussama Mounnan[1]
LABSI Laboratory, Faculty of Sciences
Ibn Zohr University Agadir, Morocco
LIASD Laboratory, Saint-Denis, France

Otman Manad[2]
Umanis S.A Research & Innovation,
7 Rue Paul Vaillant Couturier
92300 Levallois-Perret, France

Abdelkrim El Mouatasim[3]
Department of Mathematics and Management
Faculty of Polydisciplinary Ouarzazate (FPO)
Ibn Zohr University, Ouarzazate, Morocco

Larbi Boubchir[4]
LIASD Laboratory
Department of Computer Science
Paris 8 University, 93526 Saint-denis, France

Boubaker Daachi[5]
LIASD Laboratory
Department of Computer Science
Paris 8 University, 93526 Saint-denis, France

***Abstract*—Speech recognition-based biometric access control systems are promising solutions that have resolved many issues related to security and convenience. Speech recognition, as a biometric modality, offers unique advantages such as user-friendliness and non-intrusiveness, etc. However, developing robust and accurate speaker identification and authentication systems pose challenges due to variations in speech patterns and environmental factors. Integrating deep learning techniques, especially AutoEncoder and Generative Adversarial Network models, has shown promising results in addressing these challenges. This article presents a novel approach based on the combination of two deep learning models, namely, AE and GAN for speech recognition-based biometric access control. In the model architecture, the AutoEncoder takes the MFCC coefficients as input, and the encoder converts the latter to the latent space, whereas the decoder reconstructs the data. Then, speech features extracted from the latent space are used in the GAN generator to generate additional speech data. The discriminator network has a dual role, serving as both a feature extractor and a classifier. The first extracts relevant features from generated samples, while the latter distinguishes between generated and authentic samples that come from AutoEncoder. This strategy outperforms DNN and LSTM models on VoxCeleb 2, LibriSpeech, and Aishell-1 datasets. The models are trained to minimize Mean Squared Error (MSE) for both the generator and discriminator, aiming at achieving highly realistic datasets and a robust, interpretable model. This approach addresses challenges in feature extraction, data augmentation, realistic biometric samples generation, data variability handling, and data generalization enhancement, providing therefore, a comprehensive solution.**

***Keywords*—Speaker identification; speech recognition; biometric access control; authentication; verification**

I. Introduction

Speech recognition systems [1] have become increasingly important in various domains, including biometric access control, where the identification and authentication of individuals based on their unique voice characteristics are crucial. These systems aim, securely, to use biological or behavioral characteristics to authenticate and authorize individuals for access to a physical location, a device, or a system. It relies on unique and measurable traits that are specific to an individual, making it difficult to forge or replicate. These characteristics can include physiological characteristics such as fingerprints, face features, iris patterns, and voiceprints, as well as behavioral characteristics such as typing patterns, gait, and signature dynamics as shown in Fig. 1. The main function of this

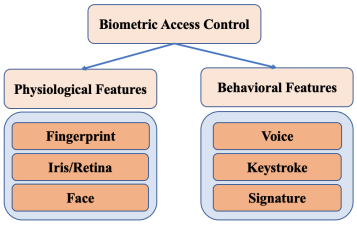


Fig. 1. Biometric categories.

paradigm is to collect biometric data, convert it into a digital template, and then compare this template to templates stored in a database. If the comparison results in a match, the individual is given access. Otherwise, access is blocked. Fig. 2 presents a system architecture based on speech recognition.

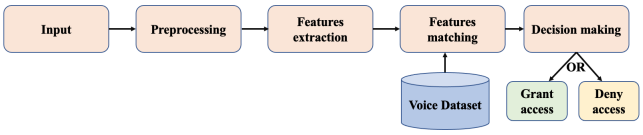


Fig. 2. Biometric access control architecture based on speech recognition.

Biometric access control systems based on speech recognition offer numerous advantages [2], such as universality, non-intrusiveness, high authentication security, and convenience i.e., bypassing the use of memorized passwords or access cards, audit trail (keep track and accountability), and faster processing than the traditional one. This concept is a powerful and convenient way to improve security and access management in a variety of areas, such as physical facilities, digital systems, and digital transactions... etc. However, to achieve reliable and robust performance, it is essential to develop accurate speaker identification and authentication mechanisms.

Deep learning models are characterized by their ability to learn sophisticated patterns and representations from data, providing a strong foundation for tackling the complexities of speech recognition [3]. Hence, providing powerful tools for improving the accuracy and effectiveness of its tasks. In this context, the use of AutoEncoder models for feature extraction [4] has shown promising results in identifying and authenticating speakers. The AutoEncoder model, a type of neural network architecture, has gained significant attention in recent years due to its capability to learn meaningful and compact representations of input data. AutoEncoders can be leveraged to extract discriminative features from raw speech signals. By training the AutoEncoder model on a large dataset of labeled speech samples, it can learn to encode the essential characteristics of a speaker's voice into a lower-dimensional representation, which facilitates efficient and accurate speaker identification and authentication processes. However, this model faces several challenges among them:

- Lack of Realism in Generated Samples: Because AutoEncoders concentrate on recreating the input data, they may produce generated samples that are excessively similar to the training data and are devoid of variation.
- Noisy or incomplete reconstructions: AutoEncoders may have trouble accurately reconstructing the input speech signals, mainly when there is noise or fluctuation.
- Limited Generalization of novel data: Because AutoEncoders tend to concentrate on recreating well-known patterns from the training set, they may have trouble in generalizing novel or unseen data.
- Incapability to Distinguish: AutoEncoders are typically unsupervised models concentrating on feature learning and reconstruction. The capacity to discriminate is essential for precise authentication in a biometric access control context.
- Limited data augmentation: AutoEncoders can be used for limited data augmentation by reconstructing and producing synthetic samples. The produced samples, however, may not represent the whole range of variability included in the training data.
- Adversarial Attacks: AutoEncoders can be vulnerable to adversarial attacks [5], which include making tiny and purposeful changes to input data in order to trick the model. In the case of speech recognition systems, this might include discreetly changing a voice recording to deceive the system into providing access to an unauthorized user.
- Inadequate Temporal Information: Traditional AutoEncoders struggle with sequential data, which is an issue in speech recognition, where the order of the input (i.e., the sequence of sounds or words) is important. Recurrent or convolutional AutoEncoders, for example, can alleviate this, although they are more sophisticated and computationally intensive.

Generative Adversarial Network (GAN) is a promising paradigm that consists of two main components: a generator and a discriminator. The generator produces synthetic data and the discriminator tries to differentiate between real and synthetic data. In the context of biometric access control using speech recognition [6], GANs can be applied to generate synthetic speech data to augment the training dataset [7], which can help address data scarcity issues, increase the diversity of the training data, and improve the robustness and generalization of the speech recognition system. Synthetic data generated by the GAN can be combined with real training data to create a more comprehensive and representative dataset for training speaker identification or authentication models, mitigating most AE model issues. It's important to note that GAN training can be challenging and may suffer from issues [8] such as training instability requiring careful tuning of hyperparameters and balancing the training dynamics between the generator and discriminator, lack of control over generated data, GANs typically generate data based on random noise input, resulting in limited control over specific characteristics of the generated speech samples, lack of feature extraction: GANs primarily focus on generating data and may not explicitly learn meaningful features from the input speech samples, and data augmentation: GANs are commonly used for data augmentation by generating synthetic samples. However, without the guidance of meaningful features, the generated samples may not effectively capture the desired variations and characteristics of real speech data. The combination of AutoEncoder (AE) and Generative Adversarial Networks (GAN) models, for biometric access control based on speech including speaker identification and authentication, resolves many problems and drawbacks related to feature extraction, data augmentation, and generating realistic biometric samples. To this end, the main goal is to contribute to the advancement of this research by leveraging the capabilities of deep learning through a novel approach that combines both models in a manner complementary to each other, providing control, stability, better feature learning, and enhanced data augmentation capabilities, leading to improved speech recognition performance.

The key contributions of this project can be outlined as follows:

- Introducing a novel approach rooted in deep learning models, specifically AutoEncoder and Generative Adversarial Network in biometric access control through speech recognition context. This integrated model enhances system performance, accuracy, robustness, and efficiency.
- Employing the AutoEncoder (AE) model as an unsupervised method for extracting meaningful and discriminative features, reducing dimensionality, and addressing storage and computational challenges associated with raw audio data analysis. Additionally, features are extracted from the latent space and utilized in the Generative Adversarial Network (GAN) to augment the training dataset, enhancing model generalization, mitigating overfitting, and alleviating data scarcity issues, resulting in the creation of high-quality, realistic biometric samples.
- Proposing a GAN model where the generator network produces synthetic speech data resembling that from the latent space representation, thereby expanding

the training dataset. This approach improves model generalization, reduces overfitting, and addresses data scarcity, leading to the generation of high-quality biometric samples. The discriminator in this proposal serves two roles: feature extraction and classification. The former extracts features from generated samples, capturing more informative and efficient features, while the latter distinguishes between generated samples from both models, enhancing overall system performance.

- Application of this approach to diverse datasets, including VoxCeleb 2, Aishell-1, and LibriSpeech, has yielded positive results when compared to outcomes from Deep Neural Network (DNN) and Long Short-Term Memory (LSTM) models.

The remainder of this article is organized as follows: Section II provides related works on speech recognition systems for biometric access control. Section III presents the proposed solution. Section IV presents the results and analysis of the experiments conducted, highlighting the performance gains achieved through the AutoEncoder-GAN-based approach. Section V presents a discussion. Finally, Section VI concludes the article with a summary of the findings and discusses potential directions for future research in this field.

II. Related Work

In the literature, there is a lot of research related to biometric access control based on speech recognition topics, including speaker identification and authentication, and speaker verification. This section presents an overview of some works and propositions published recently that achieved significant results.

Najim Dehak et al [9] proposed two speaker verification systems models, In which the first one is based on SVM, by using the cosine similarity, and the second one utilises directly the cosine similarity in the final phase which decides the final score. The experiments are done through three different methods in the variability space, which are within-class covariance normalization, linear discriminate analysis, and nuisance attribute projection. Their study conducted on the combination of LDA with WCCN has achieved good results compared to the other ones. The test was carried out using the NIST 2008 Speaker Recognition Evaluation dataset.

Yen Lei et al [10] presented a new approach based on deep learning speaker recognition using a phonetically that aims at improving speaker recognition performance by using an i-vector model [11] to represent the speech signal (extract the main features) and DNN model is used to replace the UDM-GMM [12] paradigm in order to train the model. The experiments proved that this approach has significantly improved the i-vector speaker recognition system.

Another research done by [13] has proposed d-vector instead of i-vector that aims at extracting hidden layers of a DNN as features. D-vector represents the averaged activations from the last hidden layer of DNN. Experiments of this approach have proved its efficiency in a small-footprint text-dependent speaker task. Generally, this scheme underperforms the predecessor based on i-vector-DNN.

Another research made by [14] has proposed a multi-task deep learning scheme based on the j-vector method that consists of extracting features from multitask DNN using probabilistic linear discriminant analysis (PLDA). This scheme has achieved good results than the predecessor models (i-vector, d-vector).

The Authors in [15] have proposed a new scheme based on deep neural network DNN to extract speech features called as x-vector. This latter represents the fixed-dimensional embeddings of variable-length traits. Furthermore, this research tackled also data augmentation by adding the noise and the reverb to the existing dataset to improve the efficiency of the model in the text-independent speaker tasks. Effectively, this approach has achieved better findings than the ones based on the i-vector and d-vector. Another research conducted by [16] has proposed a new end-to-end architecture based on neural networks, especially DNN and LSTM to speaker verification in the text-dependent context that aims at mapping the utterances to a score and joining them to optimize the representation of the speaker. In the same area, the authors of [17] have proposed another approach based on the end-to-end attention model. They use the CNN model to extract the noise-robust frame-level features that will become utterance-level speaker vectors using the attention model. This approach proves its effectiveness on Windows 10 “Hey Cortana”.

Another research carried out by [18] in the context of text independence has presented a new end-to-end approach based on the deep learning model to optimize the triple loss function using Residual Net block and measuring the similarity by Euclidean distance within trials. The findings show that this approach outperforms that based on conventional i-vector schemes, namely on short utterances.

In [19], the authors have proposed a new generalized End-to-end model based on LSTM. The training process has relied on the large number of utterances forming a batch. This scheme aims at optimizing the loss function through the training process in an efficient manner. The experiments show that this platform has achieved good results. N. Le et al [20] have proposed a new approach based on deep learning model, namely CNN. The main objective of this proposition is to optimize the deep speaker embedding through intra-class loss distance variance regularization compactness. The findings have proved that this approach accelerates the convergence of the training model, which enhances the model’s performance.

Another research carried out by the authors in [21] has presented an end-to-end optimized scheme based on deep convolutional features extractor combined with self-attentive and large-margin loss functions in the text-independent tasks context. They use a modular neural network instead probabilistic linear discriminant analysis (PLDA) classifier. This work made use of the experiments on VoxCeleb and NIST-SRE 2016 and has achieved an enhancement model than the others based on i-vectors.

The authors in [22] have suggested a novel approach for learning speaker embeddings based on a simulated model of GAN, especially the discriminator. This architecture aims to maximize mutual information, improving the model performance on the VoxCeleb corpus. Experiments show that this model outperforms the model based on i-vector and that based

on triples loss systems.

Many works are proposed to optimize the performance of speech recognition tasks and provide a robust system using deep learning model. Each research has focused on one aspect or more, such as data augmentation, features extraction, denoising and de-reverberation. The proposed solution has designed a new architecture based on deep learning models, namely AutoEncoder and Generative Adversarial Network in a complementary manner to improve the model performance by minimizing the loss function. The MFCC is used to extract features and the model AE to capture the meaningful speech representation and GAN is used to generate speech data from the latent space of the model AE.

III. Proposed Scheme

The proposed scheme is based on two models which are AE and GAN models as depicted in Fig. 3. At first, the speech inputs are collected, and their Mel-Frequency Cepstral Coefficient (MFCC) characteristics are extracted and used for training and tuning the model. Generally, The AE model comprises three components: Encoder, latent space, and Decoder. The encoder captures the main representation of the meaningful speaker speech features extracted from MFCC and produces the latent space, the latter will be used to reconstruct the input data. In this model, the Latent space will be extracted and used as input to the generator of the GAN model to generate more real data from it. The generated samples will be then used as input to the discriminator. This latter plays two roles, namely a features extractor and a classifier. At first, the discriminator extracts features from the generated samples and then feeds to the classification between that extracted and that comes from the AutoEncoder i.e. the decoder, to make a decision. This section presents more details of this model.

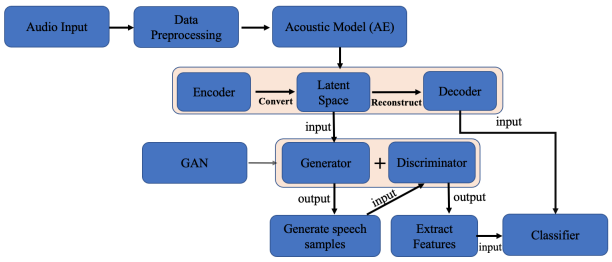


Fig. 3. AutoEncoder-Generative adversarial network model architecture.

A. Data Preprocessing

Generally, the preprocessing process [23] is crucial for preparing the data. This phase involves the capture and splitting of data into segments, feature extraction, noise removal, features normalization, and data loading, etc. Among the main steps that represent the backbone of the model namely in the context of the biometric access control based on speech, is feature extraction. To this end, the proposed architecture involves the adoption of the Mel-frequency Cepstral coefficients (MFCC).

1) Mel-Frequency Cepstral Coefficients: MFCC [24] is a technique that consists of extracting features from the signal. In the speech processing context, this method is widely used to capture the spectral features of sound well-suitable for various machine learning and deep learning tasks including speech recognition and speech analysis. Simply this technique is an amount of coefficients that represent the shape of the speech power spectrum signal. Fig. 4 represents the components of the MFCC. To calculate the coefficients of MFCC, some steps are crucial as depicted in the figure. After capturing the speech signal, the first step is breaking the signal into frames (windowing process) and then applying the Fast Fourier Transform (FFT) to determine the power spectrum of each frame. Following that mel-scale filter bank processing is performed on the power spectrum by the formula 1:

$$mel(f) = 2595 log_{10}(1 + \frac{f}{700}) \quad (1)$$

Where mel(f) represents the frequency on mels and f represents the frequency on Hz. The power spectrum is converted then by log domain and the Discrete Cosine Transform (DCT) is applied to get the coefficients of MFCC through the Eq. 2:

$$\hat{C}_n = \sum_{n=1}^{k} (log\hat{S}_k) cos[n(k - \frac{1}{2})\frac{\pi}{k}] \quad (2)$$

Where $k, \hat{S}_k, and \hat{C}_n$ represent, respectively the mel cepstrom coefficients numbers, the filter bank output and the MFCC coefficients.

Fig. 4. MFCC architecture.

B. AutoEncoder model

The AutoEncoder (AE) model is a sort of neural network architecture used for feature learning, dimensionality reduction, and data reconstruction. It is especially effective for extracting relevant representations from biometric data and may be used in a wide range of biometric modalities, in various applications, namely Feature learning, data denoising, data compression, Anomaly detection, Privacy preservation, biometric template protection...etc. An AutoEncoder's primary principle is to learn a compact and efficient representation of incoming data. It is made up of an encoder and a decoder as shown in Fig. 5. The encoder takes raw data as input and converts it to a lower-dimensional latent space representation. The fundamental traits and qualities of the data are captured by this latent space representation, and the decoder uses this later to attempt to recreate the original input data. The objective is to maintain the information required for reconstruction in the latent space. AE is an unsupervised model that aims at minimizing the loss function between the input data and the reconstructed data, capturing the most relevant representations.

The proposed solution incorporates the use of AutoEncoder to capture the relevant representation of the inputs from MFCC

coefficients, optimizing the speech processing system. The main objective of MFCC is extracting features and converting the input signal into coefficients that are retained as features which represent the main relevant features. The AutoEncoder takes these coefficients as input and converts them into latent space, reducing therefore, the dimensionality of the representation represented by the coefficients, and extracting the main relevant representation. The other network i.e. decoder network reconstructs the representation from that reduced (latent space). The main goal of this proposition is to get the most salient and compact representation from MFCC coefficients in a lower-dimensional space by training the AutoEncoder model.

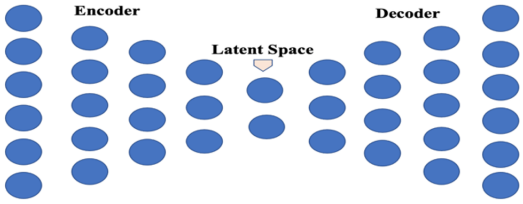


Fig. 5. AutoEncoder architecture.

C. GAN model

Generative Adversarial Network (GAN) is a generative model that is distinguished by two distinct networks, each with its unique set of attributes called Generator and Discriminator. The first seeks to produce realistic data from a specific class, while the second is used to determine whether the generated data is realistic or phony, as shown in Fig. 6: GAN is a deep-learning class used especially to produce synthetic data from the raw data input. In the scope of biometric access control, the generator takes a random noise as input and attempts to produce biometric data samples that mimic actual biometric data, while the objective of the Discriminator is to distinguish between the generated samples and the real ones, generally a binary classifier. The training procedure comprises a competition between the generator and the discriminator. As training advances, the generator improves at creating more realistic data, while the discriminator improves at differentiating between actual and phony data. This repeated procedure should result in high-quality synthetic data that is difficult to differentiate from genuine data. GANs may be used for a variety of reasons in the context of biometric access control, including data augmentation, Privacy-Preserving Research, Training Data Generation, Data Imputation, and Adversarial Attacks and Defense.

To this end, the proposed scheme extracts the latent space from the AutoEncoder model and uses it as input in the Generative Adversarial Network (GAN) model namely the Generator. This latter Generates more speech data from those reduced features, producing then data simulated to that of input. Whereas the discriminator in this architecture plays two roles, namely a features extractor and a classifier. At first, the discriminator takes the generated samples as input, extracts relevant features and then distinguishes them from that produced and trained by the AE model, especially the decoder.

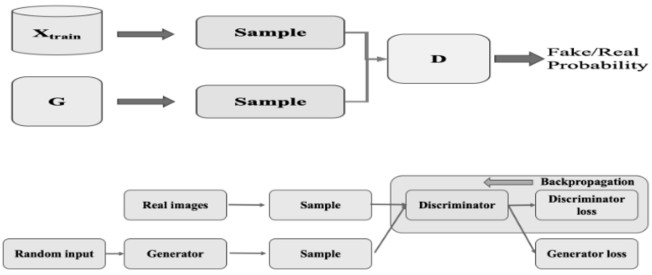


Fig. 6. GAN architecture.

IV. Experiments and Results Analysis

In this section, the experiments carried out by the laboratory team are presented, describing therefore, the datasets, the metrics and the implementation details of the proposed model, and finally the analysis of the result.

A. Datasets

In this model, three different datasets have been used, which are VoxCeleb 2, Aishell-1 And LibriSpeech.

VoxCeleb: It is an open-source dataset that is widely used in the experiments of speech processing tasks [25]. It contains videos interviews uploaded to YouTube. There are two types of VoxCelb datasets, VoxCeleb 1 and VoxCeleb 2. The first one has over 100,000 utterances For celebrities, whereas the VoxCeleb 2 has over a million utterances. In the proposed solution, the experiments have occurred on the VoxCeleb 2.

Aishell-1: This dataset is used also in the speech preprocessing tasks[26]. It is an open-source and freely accessible speech dataset that contains Mandarin speech captured with a high-fidelity microphone (44.1 kHz, 16-bit). The Aishell-1 dataset was created by downsampling the audio collected by the high-fidelity microphone to 16 kHz. A set of 400 speakers from various accent areas in China took part in the record capture.

LibriSpeech: is an open-source corpus, available in [27]. It contains 1000 hours of speech sampled at 16KHz and is generated from audiobooks in the LibriVox project. This dataset is used mainly in speech preprocessing including speech recognition and speaker identification. Table I represents the specification of the used datasets:

TABLE I. Datasets Specification

	Speakers	Utterances	Hours
VoxCeleb 2	6112	1 128 246	2442
Aishell 1	400	141 925	Over 170
LibriSpeech	2087	252 702	1000

B. Evaluation Metrics

Generally, a metric is a method used to evaluate a system's performance on a specific task. The main metric objective is measuring the quality of classifications or predictions carried out by a system or model. A loss or error function [28] is a function that determines how much the output or predicted

value departs from reality or actual value aiming at optimizing the model (either maximizing or minimizing issues). The Mean Squared Error (MSE) [29] is a loss function that measures the error between the observed and predicted values. The average of errors squared is calculated by this Eq. 3.

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n} \quad (3)$$

Where y_i represents the observed value, $\hat{y}_i$ represents the predicted value, and n is the observations number In this study, the MSE metric is used in the AutoEncoder model to evaluate its performance, and the Binary Cross Entropy (BCE) metric is also used in the GAN model that represents the difference between the predicted probability distribution and the reel one. On one hand, BCE is used to solve the binary classification issues, evaluating then, the model's performance. On the other hand, is used to quantify the training loss, minimizing therefore, the loss function of the model during training.

C. Implementation Configuration and Results Analysis

In the proposed scheme, the PyToch library, written in Python programming language, has been used for training networks based on deep learning models. This model has adopted Graphics Processing Unit (GPU), due to its efficiency in Neural Network processing. After capturing the MFCC coefficients from the speech, 13 coefficients, The latter are then fed to the AE model namely, the encoder that converts the inputs to latent space, reducing therefore the dimensionality and capturing essential speech features. These high-level features serve to reconstruct the speech data from the bottleneck layer and aim at generating outputs that closely resemble the original data input. The model training aims at minimizing the reconstruction loss between the original and reconstructed speech. In the training of the AutoEncoder, the chosen specifications include 8 dimensions as the latent dimension, 64 as the batch size, and 0.001 as the learning rate. Within the first part of the proposed architecture, the AE model is implemented with input dimensions set to 8. The encoder network consists of 128 units or neurons in the hidden layer, employing the Rectified Linear Units (ReLU) as an activation function. The use of ReLU introduces non-linearity, facilitating the model in learning complex relationships within the data. The learning rate and the network size are identified using different settings based on the try-and-error approach to choose the best configuration in terms of performance.

At the beginning of the training process, the weights are initialized at random and then gradually updated. To solve the model overfitting challenges, different methods are used such as the regularization of the parameters to promote lower values of weight, and adding dropout layers within the encoder and the decoder, furthermore, the loss function regularization has been adopted to promote certain desired behaviors in the latent space. The data mapping process is carried out from the 128-dimensional hidden representation to the latent space representation. The Adam optimizer has been deployed. The loss function is selected as the Mean Squared Error (MSE) as mentioned before.

In this proposition, the latent space features are extracted representing the high-level speech representation to feed it into the GAN model, namely the generator network. This latter takes the high-level representations (more relevant speech features) as input to generate more speech data in a manner that resembles real speech. The architecture of the generator is composed of three fully connected linear layers with ReLU activation functions between them. The Tanh activation function has been applied in the final layer to ensure that the generated values are bounded within the range [-1,1]. The other GAN network, i.e., the discriminator, plays two roles in this architecture, a features extractor and a classifier. At first, the discriminator takes the generated samples from the generator, tries to extract the relevant representations and then feeds them to the classifier to distinguish them from those that come from the decoder of the AE model. The structure of the discriminator is similar to that of the generator. It consists of three fully connected layers with ReLU activation functions. In the final layer, the sigmoid activation function has been applied, which produces values within the range [0,1] where 1 identifies the real data, and 0 identifies the fake ones. Both the generator and the discriminator are adversarial trained. i.e. competing against each other. This process helps us to refine the ability of the generator to generate more high-level quality speech data, and therefore, achieve a robust system based on the combination of two promising deep learning models, AutoEncoder and Generative Adversarial Network, especially in the speech recognition tasks. Fig. 7 represents AE-GAN model training process using three different datasets, with the loss versus training epochs to illustrate how well the model learns. The experiments incorporate different utterances from three different datasets, including VoxCeleb 2, LibriSpeech, and Aishell-1. These datasets are divided into three parts for each dataset, 80% for training, 10% for validation, and 10% for test.

Experimentation involved assessing the proposed deep AE-GAN model by utilizing the state-of-the-art models, namely the Deep Neural Network (DNN) model and Long Short Term Memory (LSTM) model, using the datasets mentioned above to describe the experiment findings. Table II lists the overall loss function of the models that are used in the test process during various research phases. As shown in the table, the AE-GAN model has a high score in training and validation in three different datasets, which are LibriSpeech, VoxCeleb 2, and Aishell 1, it has achieved respectively in training loss, 0.0574, 0.0876, and 0.0886, and in validation loss 0.0581, 0.0888, and 0.0889. Compared to the results of DNN and LSTM models, they have gotten in the training phase values ranging from 0.07 and 0.168, while in the validation phase, huge values ranging between 0.30 and 0.48, proving generally the overfitting of the models. The proposed scheme has proved its efficiency and outperformed the performance of DNN and LSTM models in three different datasets. Fig. 8 depicts the results of the experiments carried out over the datasets using the DNN and LSTM models.

V. Discussion

Deep Neural Networks (DNNs) and Long Short-Term Memory networks (LSTMs) are reference models in speech recognition-based biometric access control context, and have been widely used in many studies. DNNs have demonstrated their performance in learning hierarchical representations from raw audio data. Their ability to handle complex features with

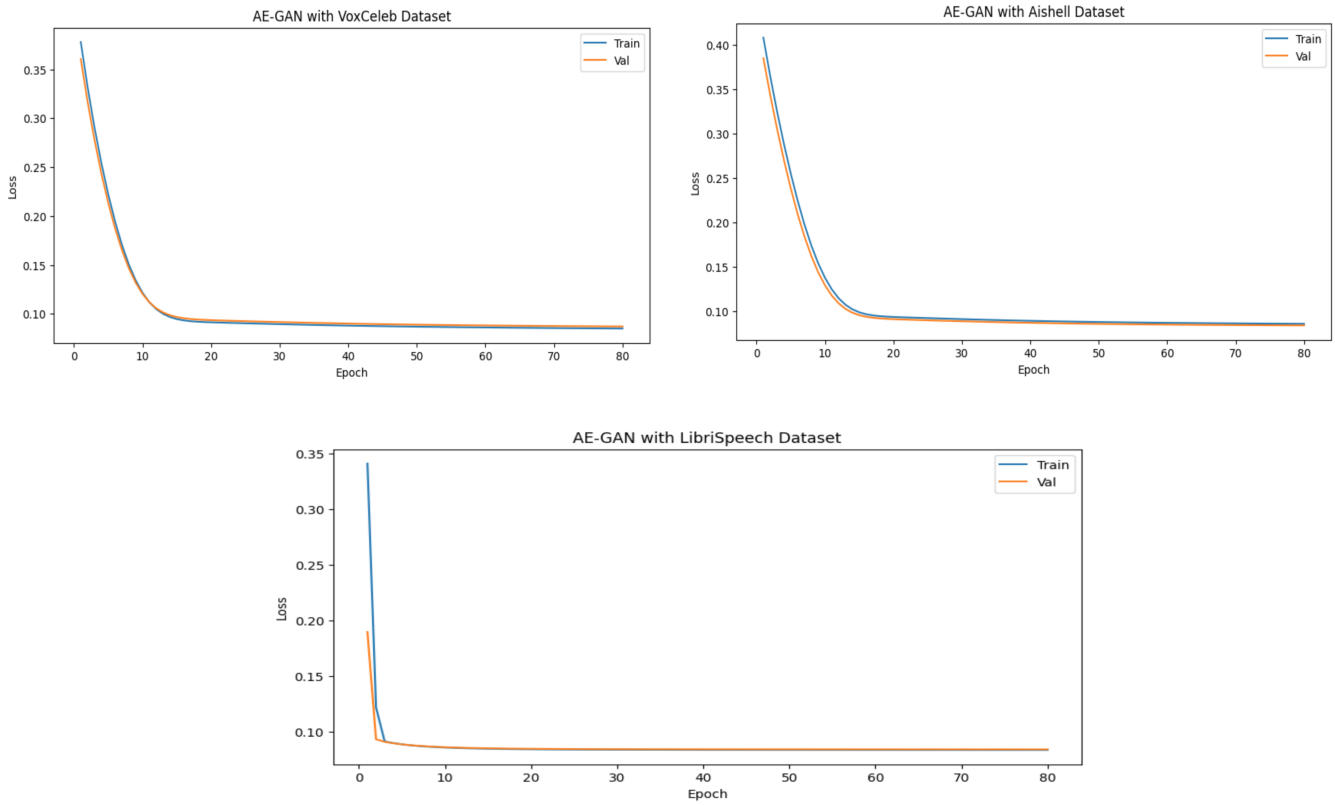


Fig. 7. AE-GAN model training loss curve vs epoch.

TABLE II. AVERAGE LOSS PER EPOCH FOR TRAINING AND VALIDATION

Dataset	Model	Loss function	
		Train	Validation
LibriSpeech	DNN	0.079	0.334
	LSTM	0.095	0.307
	AE-GAN	**0.0574**	**0.0581**
VoxCeleb 2	DNN	0.152	0.298
	LSTM	0.168	0.287
	AE-GAN	**0.0876**	**0.0888**
Aishell 1	DNN	0.084	0.389
	LSTM	0.079	0.486
	AE-GAN	**0.0886**	**0.0889**

deep increased network has contributed to their standing in the field. However, this model struggles with capturing long-range dependencies in sequential data, which is crucial for speech recognition tasks. Simultaneously, LSTMs are recognized for their effectiveness in modeling temporal dependencies within sequential data, making them well-suited for capturing long-term patterns in speech sequences. They have addressed the vanishing gradient issues that are inherent in traditional Recurrent Neural Networks (RNNs), making them more adept at learning from sequential data. But to capture complex patterns effectively, LSTMs require more data.

The AE-GAN's ability to leverage the latent space features extracted by the AutoEncoder to enhance the generative capabilities of the GAN is a potential advantage. In scenarios with limited labeled data, the AE-GAN's capacity for generating high-quality and realistic speech samples may prove advantageous. Additionally, its ability to address overfitting challenges through regularization techniques and dropout layers may contribute to superior performance in diverse speech recognition tasks. Although the proposed model offers several advantages, it may face challenges in scenarios where there is insufficient diversity in the training data, potentially leading to biased representations. If the dataset lacks sufficient variation in terms of speakers, accents, or speech characteristics, the model may struggle to generalize well to a broader range of real-world scenarios. Augmenting the dataset with more diverse samples could enhance the model's robustness. Additionally, the model's performance may be sensitive to hyperparameter settings, necessitating careful tuning. Implementing automated hyperparameter tuning methods or conducting a thorough sensitivity analysis may help identify robust configurations more efficiently. The computational complexity of the model, especially in training large-scale datasets, could pose limitations in terms of time and resource requirements. The computational costs related to the training deep learning models, including the proposed AE-GAN model, are a significant consideration.

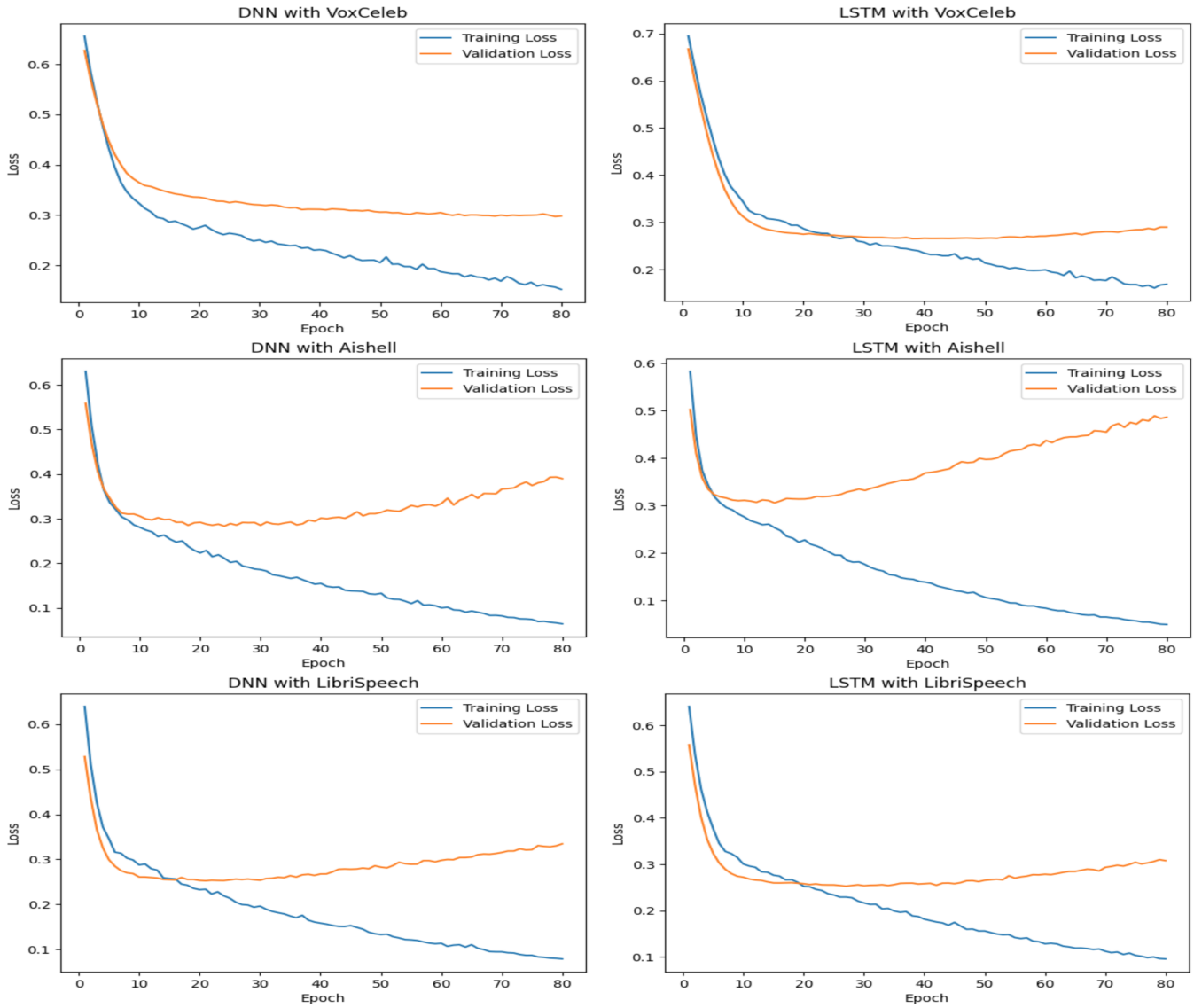


Fig. 8. The results of DNN and LSTM models with VoxCeleb, aishell, and libriSpeech datasets.

There are several solutions that contribute to mitigating these costs among them, efficient GPU utilization can be essential, optimizing the model architecture and exploring parallelization techniques can contribute to faster training times. Additionally, the use of transfer learning from pre-trained models can alleviate the need for extensive training on large datasets. Another potential solution is the exploration of model quantization techniques, reducing the precision of model weights to accelerate inference. Leveraging distributed training across multiple GPUs or utilizing cloud-based computing resources can further expedite the training process. Implementing early stopping and model checkpointing strategies can optimize training efficiency by preventing unnecessary iterations. Generally, a nuanced understanding of the AE-GAN model's strengths, a transparent acknowledgement of study limitations, and proactive strategies to address computational costs collectively contribute to a robust evaluation framework for advancing the field of speech recognition.

VI. Conclusion and Future Work

This paper has proposed a new approach based on speech recognition for speaker identification and authentication that is considered as the main and crucial task in the speech-based biometric access control scenario. The model has proved its efficiency and robustness based on the combination of AE and GAN models. The proposed model provides an optimized platform integrating the features learning and tackling the data augmentation and generalization issues, especially the speech dataset, and data imputation such as reconstructing degraded audio or denoise and tuning the hyperparameters of the models. This approach has been implemented on three different datasets: VoxCeleb2, LibriSpeech, and Aishell-1, and has achieved good results in terms of performance, compared to AE and GAN models. However, the proposed scheme is expensive in terms of time-consuming, especially in the training phase where there are two models AE and GAN. In

future endeavors, the focus will be on this aspect to optimize the proposed scheme.

REFERENCES

[1] S. Li, J. You, and X. Zhang, "Overview and Analysis of Speech Recognition," in 2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Aug. 2022, pp. 391–395. doi: 10.1109/AEECA55500.2022.9919050.

[2] V. S. S. S. Hari, A. K. Annavarapu, V. Shesamsetti, and S. Nalla, "Comprehensive Research on Speaker Recognition and its Challenges," in 2023 3rd International Conference on Smart Data Intelligence (ICSMDI), Trichy, India: IEEE, Mar. 2023, pp. 149–152. doi: 10.1109/ICSMDI57622.2023.00034.

[3] K. B. Bhangale and M. Kothandaraman, "Survey of Deep Learning Paradigms for Speech Processing," Wireless Pers Commun, vol. 125, no. 2, pp. 1913–1949, Jul. 2022, doi: 10.1007/s11277-022-09640-y.

[4] O. İrsoy and E. Alpaydın, "Unsupervised feature extraction with AutoEncoder trees," Neurocomputing, vol. 258, pp. 63–73, Oct. 2017, doi: 10.1016/j.neucom.2017.02.075.

[5] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning." arXiv, Jul. 06, 2018. doi: 10.48550/arXiv.1712.07107.

[6] A. Wali et al., "Generative Adversarial Networks for speech processing: A review," Computer Speech & Language, vol. 72, p. 101308, Mar. 2022, doi: 10.1016/j.csl.2021.101308.

[7] Y. Qian, H. Hu, and T. Tan, "Data augmentation using generative adversarial networks for robust speech recognition," Speech Communication, vol. 114, pp. 1–9, Nov. 2019, doi: 10.1016/j.specom.2019.08.006.

[8] D. Saxena and J. Cao. 2021. "Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions". ACM Comput. Surv. 54, 3, Article 63 (April 2022), 42 pages. https://doi.org/10.1145/3446374

[9] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, May 2011, doi: 10.1109/TASL.2010.2064307.

[10] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2014, pp. 1695–1699. doi: 10.1109/ICASSP.2014.6853887.

[11] N. S. Ibrahim and D. A. Ramli, "I-vector Extraction for Speaker Recognition Based on Dimensionality Reduction," Procedia Computer Science, vol. 126, pp. 1534–1540, 2018, doi: 10.1016/j.procs.2018.08.126.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, no. 1, pp. 19–41, Jan. 2000, doi: 10.1006/dspr.1999.0361.

[13] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2014, pp. 4052–4056. doi: 10.1109/ICASSP.2014.6854363.

[14] N. Chen, Y. Qian, and K. Yu, "Multi-task learning for text-dependent speaker verification," in Interspeech 2015, ISCA, Sep. 2015, pp. 185–189. doi: 10.21437/Interspeech.2015-81.

[15] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2018, pp. 5329–5333. doi: 10.1109/ICASSP.2018.8461375.

[16] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-End Text-Dependent Speaker Verification." arXiv, Sep. 27, 2015. doi: 10.48550/arXiv.1509.08062.

[17] S.-X. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-End Attention based Text-Dependent Speaker Verification." arXiv, Jan. 02, 2017. doi: 10.48550/arXiv.1701.00562.

[18] C. Zhang and K. Koishida, "End-to-End Text-Independent Speaker Verification with Triplet Loss on Short Utterances," in Interspeech 2017, ISCA, Aug. 2017, pp. 1487–1491. doi: 10.21437/Interspeech.2017-1608.

[19] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification." arXiv, Nov. 09, 2020. doi: 10.48550/arXiv.1710.10467.

[20] Le, N., Odobez, J.-M. (2018) Robust and Discriminative Speaker Embedding via Intra-Class Distance Variance Regularization. Proc. Interspeech 2018, 2257-2261, doi: 10.21437/Interspeech.2018-1685

[21] Bhattacharya, G., Alam, J., Kenny, P. (2019) Deep Speaker Recognition: Modular or Monolithic? Proc. Interspeech 2019, 1143-1147, doi: 10.21437/Interspeech.2019-3146

[22] M. Ravanelli and Y. Bengio, "Learning Speaker Representations with Mutual Information." arXiv, Apr. 05, 2019. doi: 10.48550/arXiv.1812.00271.

[23] M. Razavi et al., "Machine Learning, Deep Learning and Data Preprocessing Techniques for Detection, Prediction, and Monitoring of Stress and Stress-related Mental Disorders: A Scoping Review." arXiv, Aug. 08, 2023. doi: 10.48550/arXiv.2308.04616.

[24] V. Tiwari, "MFCC and its applications in speaker recognition," International journal on emerging technologies, vol. 1, no. 1, pp. 19–22, 2010.

[25] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in Interspeech 2017, Aug. 2017, pp. 2616–2620. doi: 10.21437/Interspeech.2017-950.

[26] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline." arXiv, Sep. 16, 2017. doi: 10.48550/arXiv.1709.05522.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2015, pp. 5206–5210. doi: 10.1109/ICASSP.2015.7178964.

[28] J. Terven, D. M. Cordova-Esparza, A. Ramirez-Pedraza, and E. A. Chavez-Urbiola, "Loss Functions and Metrics in Deep Learning." arXiv, Sep. 06, 2023. doi: 10.48550/arXiv.2307.02694.

[29] T. O. Hodson, T. M. Over, and S. S. Foks, "Mean Squared Error, Deconstructed," Journal of Advances in Modeling Earth Systems, vol. 13, no. 12, p. e2021MS002681, 2021, doi: 10.1029/2021MS002681.

Estimation of Hazardous Environments Through Speech and Ambient Noise Analysis

Andrea Veronica Porco
Dept. Information Engineering
University of the Ryukyus
Nishihara, Japan

Kang Dongshik
Dept. Information Engineering
University of the Ryukyus
Nishihara, Japan

***Abstract*—In recent years, significant attention has been directed towards the development of artificial empathy within the engineering academic community. Replicating artificial empathy necessitates the capability of agents to discern human emotions and comprehend environmental risks. Analyzing acoustic data in real environments offers a higher level of non-invasive privacy compared to video and camera data, limiting the agent's understanding to specific patterns. However, current studies are negatively affected by subjective inferences from real data, which can result in inaccurate predictions, leading to both false positives and negatives, especially when contextual data and human speech are involved. This paper work proposes the estimation of a dangerous environment in accordance with the emotional speech and additional ambient noises. In this approach we implement a variational autoencoder model in conjunction with a classifier for training the classification task. Additional regularization techniques are applied to bridge the gap between the original training data and the expected data. The classifier utilizes feature data generated by the variational autoencoder to extract class patterns and determine whether the environment is hazardous. Emotional speech is classified as angry, sad, or scared emotions, contributing to the classification of danger, while happy, calm, and neutral emotions are considered safe. Various ambient noise types, including gunfire and broken glass, are categorized as dangerous, while real-life indoor noises like cooking, eating, and movements are considered safe.**

***Keywords*—Dangerous environment detection; speech analysis; acoustic audio analysis; ambient noises; variational autoencoder model; empathetic systems**

I. Introduction

Ensuring the safety of individuals within indoor environments is a paramount concern, with implications spanning from residential spaces to critical infrastructure. The ability to accurately assess and respond to potential hazards is crucial for safeguarding lives and minimizing risks. In recent years, the pursuit of artificial empathy within the engineering domain has gained significant traction, aiming to imbue computational agents with the capacity to comprehend human emotions and navigate environmental dangers. An avenue of exploration in this pursuit involves the analysis of acoustic data, providing a non-intrusive means of understanding the surrounding environment. Unlike more invasive data sources like video and cameras, acoustic data analysis preserves privacy by focusing on discernible patterns, presenting a valuable approach for ensuring security in various settings.

The research presented in this paper addresses a critical facet of safety by proposing a method for estimating hazardous environments through the evaluation of emotional speech and ambient noises. This approach not only advances the field of artificial empathy but also holds substantial promise for real-world applications, particularly in indoor acoustic analysis and speech classification. The implications of this work extend beyond the academic realm, offering tangible benefits for society at large. The ability to accurately classify emotions and distinguish between safe and hazardous sounds has significant societal impact, enhancing security in public spaces, homes, and workplaces. Moreover, in the realm of engineering, the proposed method contributes to the refinement of hazard detection systems, with potential applications in areas such as smart home technologies, security surveillance, and other safety-critical environments. This research sets the stage for a more nuanced understanding of the acoustic environment, bridging the gap between subjective inferences and objective safety assessments, thereby paving the way for advancements in both theoretical understanding and practical implementation.

Analyzing a diverse range of events, including human speech and ambient sound, presents a formidable challenge for artificial agents. Consequently, accurately judging environmental characteristics becomes a complex endeavor. Moreover, this task necessitates numerous sensors, such as cameras for image and video processing, coupled or decoupled infrared sensors, and other costly apparatus, making the practical implementation of artificial home assistants exceedingly challenging to achieve. Addressing the challenges inherent in enhancing their practical implementation involves different tasks, such as managing real-time processes effectively, as evidenced in related papers [1]–[3].

Another challenge is to ensure accurate object localization such as in [4]–[7] where they proposed for example, a convolutional recurrent neural network for joint sound event localization and detection of multiple overlapping sound events in three-dimensional space. In particular the sound event localization and detection is extensively utilized by works based on robotics navigation and natural interaction with surroundings.

Background noise treatments and reduction have been extensible studied such as in [8]. The reduction of the noise comes to fulfil two different targets, the human hearing safety and the reduction of background noise to interpret another sounds or a clear speech. The source separation of overlapped sounds in acoustic event identification have studied in [9] to feat one of the pending challenges. While in [10]–[13], a study of event detection by ambient sounds analysis was performed, trying to give realism to the scenario through the addition of

diverse types of ambient sounds.

Among several challenges, to perform a precise and realistic danger classification and estimation is especially required, considering subjective human perspectives and the critical task of minimizing the false alarms [14]. Some studies were carried trying to estimating hazardous environment from ambient sounds with support vector machine models such as in [15]. However the issue continue active and open to date. The significance of false alarms cannot be overstated, owing to the inherent subjectivity found in real-life scenarios. Environmental sounds has been under-researched compared to standard speech and music, and its understanding tend to be subjective depending on the scenario and the listener [16].

Previous research endeavors have explored various ambient sounds as cited in [17], yet remarkably, the emotional states of individuals within the room have never been integrated into the equation, amplifying the complexity of the challenge at hand, and making it more realistic.

Furthermore, the usage of generative models have increased in the study of human emotions and context analysis due to the flexibility and versatility to represent and analize different types of data present all together in the same audio frame [18]–[24]. Generative models, particularly when integrated with classifiers such as variational autoencoders (VAEs), prove to be highly advantageous for classification tasks in the domain of speech and also with ambient sound. The combination of generative and discriminative capabilities allows for effective feature extraction and representation learning, enhancing the model's ability to discern patterns in complex audio data. The limitations of generative models in this context are primarily associated with tasks that demand perfect data reconstruction. Challenges arise when attempting to faithfully reproduce the intricate details of diverse audio signals, including variations in speech patterns, accents, and environmental sounds. However, in classification tasks, where the focus is on discerning relevant features rather than achieving precise data reconstruction, these limitations are mitigated. The flexibility and adaptability of generative models make them well-suited for classification applications, offering a powerful and efficient approach to audio analysis.

To the best of our knowledge, the detection of a dangerous environment was never judged by an emotional speech analysis in combination with ambient noises analysis with generative models, such as a variational autoencoder (VAE) model that learn the characteristics related with a subjective environment. Additionally, the proposed model make an adjustment of the difference among input data and expected data with phonetic and prosody features.

II. Proposed Approach

A. Proposed Model

The proposed model falls under the category of semi-supervised learning, which is a hybrid method combining labeled and unlabelled data. In this approach, the classifier learns from labeled examples and also utilizes information from unlabelled data to enhance its performance. Within our model, the VAE serves a dual purpose: it functions as an unsupervised autoencoder, learning a condensed representation of the data, and as a supervised classifier, predicting emotional classes. This dual role is possible because the model incorporates both the reconstruction loss (unsupervised) and the danger classification loss (supervised), allowing it to harness the advantages of both labeled and unlabelled data. In essence, our model is semi-supervised because it integrates labeled emotional class data along with unlabelled Mel spectrogram data during the training process, optimizing its performance.

The proposed classifier model utilizes the latent space generated by the VAE model. In our approach, these two models operate independently; first, the VAE is trained separately, and then the pre-trained encoder from the VAE, which has learned from unsupervised data, functions as a feature extractor in the classifier. This encoder transforms the data into a compact representation, which is then processed through a classifier. This classifier is specifically trained to predict danger labels based on these encoded features, which are both from labeled and unlabelled data. Consequently, our model is categorized as a semi-supervised learning approach incorporating both types of data during its training process.

Moreover, the regularization task in the extended VAE provides additional control over the decoded data. This control is achieved by utilizing pre-processed data and its representations within the VAE model, allowing for a more refined and controlled learning process [25]–[28].

The phonetic and prosody characteristics of each value derived from the input and decoded data, will be compared. The aim is to ensure that the phonetic and prosody attributes of the VAE representation closely match the pre-processed values from the input data during the extended VAE training process.

In this approach, the classifier relies on the trained encoder for its classification tasks. Consequently, the regularization process also impacts the final results of the classifier. Classifying the extended dataset containing dangerous patterns presents a challenge for our classifier. This challenge arises not only from the variations in volume and intonation between the actors' utterances but also from the inclusion of surrounded noises specific to the content of danger.

The prosodic control and regularization was previously observed in [29], [30]. The phonetic and prosodic features considered include spectral bandwidth, spectral contrast, and formants from 1 to 5 extracted from each input audio's Mel spectrogram. Our proposed classifier, coupled with the phonetic and prosodic regularized VAE method, encompasses multiple tasks. The subsequent subsections will delve into the specifics of each task involved in our approach.

The basic architecture of the hazardous environment classifier model can be observed in Fig. 1. The variational autoencoder is represented with an encoder, latent space and decoder structure, while the classifier consumes data from the features captured by variational autoencoder in the latent space.

B. Data Selection

In the process of data preprocessing, we obtained audio samples from the Ravdess datasets and custom data from distinct sources. Our primary objective was to construct a coherent emotional audio dataset and convert these audios into Mel spectrograms. Intermediate steps were taken to seamlessly

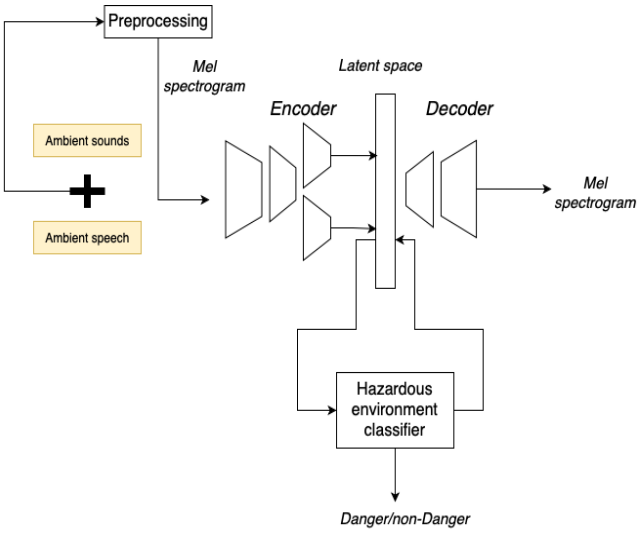


Fig. 1. Proposed architecture of the hazardous classifier model.

integrate these diverse data types, involving tasks such as aligning sampling rates, set at 44.1KHz.

Within the Ravdess dataset, audio clips exceeded three seconds, with approximately 2 seconds of null data. Consequently, trimming zero-data was imperative to obtain a high-quality signal for further processing. Normalization of voices to a standardized volume and noise reduction were performed, particularly essential for downloaded patterns that exhibited varying durations and significant zero data within different frames.

The generation of Mel spectrograms involved utilizing the short time Fourier transform (STFT) technique, followed by mapping the spectrogram to a Mel scale. This method, preconfigured with 128 Mel values, enabled a precise characterization of the audio data.

Emotion selection involved three emotions from the Ravdess dataset, excluding "Disgust" as it does not pertain to a dangerous or non-dangerous environment precisely. The chosen emotions (Neutral, Calm, Happy, Sad, Angry, Fearful) were carefully balanced to ensure equal representation in the dataset.

Regarding ambient noises, glass breaking and gun firing were chosen to represent dangerous environments, while cooking, eating noises, human steps, and opening/closing windows represented safe environments. The specific quantities used are detailed in the experiment section.

In our training approach, we focused on scenarios that involve neutral, calm, and happy emotions coupled with typical indoor noises like cooking, eating, and human movements. The decision to exclude scenarios where both dangerous and non-dangerous noises and speech coexist was deliberate. Training artificial neural networks with such mixed data might lead to the erroneous understanding that during routine, calm, or happy speech, potentially dangerous events like gunfire or breaking glass should be expected. This contradicts real-life situations where such consistency is infrequent and therefore was not incorporated into our training data to maintain the model's adherence to realistic scenarios.

C. Prosodic and Phonetic Regulariser Features's Description

Detecting hazardous environments using ambient sounds and speech poses a significant challenge, benefitting greatly from a multifaceted approach involving various audio features. In our research, we focus on employing spectral bandwidth, spectral contrast, and formants 1 to 5 for this purpose.

Formants in speech processing refer to the resonant frequencies of the vocal tract, manifesting as peaks in the sound spectrum. They play a vital role in speech production and perception, representing specific vocal tract configurations through their frequencies. Notably, the first few formants (such as F1, F2, F3, etc.) are pivotal in speech recognition, differentiating speech sounds based on their positions and transitions. Although they are not typically regarded as prosodic features, formants are instrumental in recognizing vowels and consonants, providing essential phonetic information in speech analysis [31].

Spectral contrast, another key feature, quantifies the amplitude disparity between peaks and valleys in the sound spectrum. This metric captures variations in spectral energy, indicating the sharpness or smoothness of transitions between different frequency bands. Drastic changes in spectral contrast signify specific events or objects in the environment. In danger detection scenarios, abrupt increases in spectral contrast can indicate events like glass breaking or gunshots. Monitoring these shifts enables the identification of unusual and potentially perilous situations [32].

Spectral bandwidth, the third feature under consideration, refers to the width of the frequency spectrum of a sound. It measures the dispersion of frequencies around the central frequency. Sounds with broader spectral bandwidths encompass a wider frequency range and are generally classified as broadband sounds. In contexts where danger needs to be identified, wide spectral bandwidths indicate loud and potentially hazardous noises, especially in otherwise quiet settings. For instance, while the sounds of everyday activities like cooking or eating are typically narrowband, noises such as gunfire or explosions produce broad-spectrum signals. Analyzing spectral bandwidth helps recognize the presence of such broad-spectrum events, aiding in the detection of potential threats [32].

By comprehensively analyzing formants, spectral contrast, and spectral bandwidth, an acoustic system can effectively differentiate between normal activities and events that might pose a danger within a specific environment. These distinctive features allow the system to identify specific sound patterns associated with perilous situations, making them invaluable tools in acoustic surveillance and safety systems.

D. Mathematical Definition of Proposed Model's Regularization

The additional phonetic-prosodic regularisation term $R(phopro(x), phopro(p_\theta(x|z)))$ was added to the well known evidence lower bound (ELBO) of VAE [33], in order to make each input datum x, to remain close to the corresponding decoded datum in the vector of danger $phopro(p_\theta(x|z))$ and $phopro(x)$, respectively.

The danger term to be added is as follows:

$$PhoProDiff\_R\_Loss = \alpha \cdot R(phopro(x), phopro(p_\theta(x|z))) \quad (1)$$

PhoProDiff stands for phonetic-prosodic difference regularization loss.

The prosodic regularised variational autoencoder loss function will finally be defined as follows.

$$L(\phi, \theta, x) = E_{q_\theta}(z|x)[log_{p_\theta}(x|z)] - D_{KL}(q_\phi(z|x)\|p(z)) + \alpha \cdot R(phopro(x), phopro(p_\theta(x|z)) \quad (2)$$

Letting α being $0 \leq \alpha \leq 1$. Assuming E as the expected value, D_{KL} as Kullback-Leibler Divergence, and R as regularisation.

The $R(phopro(x), p_\theta(x|z))$ term is defined as a mean squared error for each spectral feature. Phonetic-Prosodic regulariser over one spectral feature calculation can be defined as follows.

$$R(phopro(x), phopro(p_\theta(x|z))) = (phopro(x) - phopro(p_\theta(x|z))^2 \quad (3)$$

The combination of formants, spectral contrast, and spectral bandwidth extracted from the actual pre-processed Mel spectrogram data is represented as the phonetic and prosodic vector $phopro(x)$. This vector serves the purpose of enforcing regularization, ensuring alignment with the phonetic and prosodic attributes of the speech data.

The regularization based on phonetic and prosodic qualities is applied during the training process [34].

III. Experiments

A. Experiment Details

In our proposed methodology, we devised two distinct models: the Variational Auto-Encoder (VAE) and the danger classifier, each fulfilling specific roles. The proposed VAE operates as an unsupervised learning tool, generating a condensed data representation suitable for tasks like data generation and denoising. However, it is not optimized for direct danger classification.

Conversely, the danger classifier specialises in precisely this task, classifying danger based on the acquired features. It takes encoded features from the danger encoder and associates them with corresponding danger and non-danger classes. This separation allows for independent training processes and facilitates the exploration of various classifier architectures without impacting the proposed VAE. This design ensures the versatility of the learned representation from the proposed VAE for diverse downstream tasks, including danger classification.

Essentially, the proposed VAE learns a meaningful latent representation of input data, which the danger classifier utilizes for classification. This clear division of roles enhances modularity and adaptability in the overall learning process. Our danger classifier follows a supervised learning paradigm, categorizing input data into distinct danger and non-danger classes. Using an emotion dataset containing Mel spectrogram images and corresponding danger labels, the classifier learns to map these spectrograms to specific danger labels.

Our prosodic regularized variational auto-encoder model is trained with emotionally expressive speech audio. We have innovatively incorporated adjustments between speech and ambient noise sounds, introducing a novel approach. Notably, phonetic and prosodic adjustments have never been applied to this kind of input data within an adapted auto-encoder. Implementing our model under these conditions enables us to capture danger data more realistically, emphasizing the value of a generative model in extending real speech with authentic sounds often present in genuine danger environments.

Regarding ambient noises, we utilized a total of 40 audio clips. Ten audio clips were dedicated to glass breaking and gunfire, randomly interspersed with angry, fearful, and sad emotional audio clips. Similarly, there were ten audio clips, for cooking or eating noises and indoor movements (such as household steps), randomly distributed with happy, calm, and neutral emotional audio clips. The glass-breaking sounds varied, encompassing scenarios like breaking a window, objects falling and glass being thrown until breaking, as well as handling glasses, considering the potential harm to third parties in the room or the individual handling them. The gunfire sounds included various types of guns such as standard guns, pistols, and rifles. Additionally, we included gunfire from a distance sufficient to be heard from a room in a house.

For each emotion, we collected four neutral audio clips and eight audio clips from each of the other five emotions, per actor. Our dataset comprises a total of 24 actors, ensuring gender balance. In summary, from each actor, we utilized 40 audio clips, resulting in a total of 1056 audio clips used for training and testing. In our study, we assumed our data was initially separated, and we organized it placing speech at the beginning followed by danger noises, creating 1-second audio segments. While it is ideal for the data to be pre-separated, we consider this task accomplished within our work.

In our speech processing experiments, we utilized two categories of training data: acted speech and real daily conversation speech nuances. Acted speech, found in audiobooks, involves actors simulating emotions. In contrast, daily conversation speech nuances captures natural expressions from sources like YouTube talk-shows, street conversations, and shop dialogues. Both types of data were included in our study, restricted to indoor nuances and speech. Acted speech for testing purposes was sourced from the RAVDESS database [35], while daily conversation nuances data was collected from diverse real-world environment downloaded from Freesound public open datasets.

Our combined dataset merged the RAVDESS database, consisting of facial and vocal expressions in North American English from 24 gender-balanced actors, with custom data containing emotion- and ambient noise in present in the environment. The dataset we compiled included 1056 audio clips used for training and testing, recorded at 44.1kHz, from 12 actors, covering 6 selected emotions out of 8 available

emotions. Each emotion was associated with specific patterns, enhancing authenticity. Sentences from the database, such as "Kids are talking by the door" and "Dogs are sitting by the door," were utilized. The training and testing data were divided into 80 percent and 20 percent, respectively.

To maintain dataset consistency, we linked emotions to the primary dataset. For instance, selecting a "happy" emotion from a male actor involved aligning emotional level sentences with non dangerous ambient nuances, leveraging the similar vocal characteristics in emotional patterns. Ambient noises were randomly chosen while ensuring alignment in events that tend to occur at the same time, or follows to one another, such as angry, sad or scared speakers followed by a glass broken or a gunfire. In the initial tests, 40 audio clips were matched with each corresponding RAVDESS audio danger pair randomly. Importantly, generative models were minimally affected by these variations since they were incorporated during the preprocessing steps.

Both the dangerous environment classifier and the proposed VAE model utilized convolutional layers on pre-processed Mel spectrogram data. The input size for the proposed VAE was 128 by 128 (resized) for both training and testing sets. The proposed VAE's encoder and decoder consisted of two hidden layers, reducing data dimensions from 128 to 64 and then to 32 in the encoder, and restoring it from 32 to 64 and finally to 128 in the decoder. The proposed VAE featured a single output for encoding and reconstructing data. The emotional classifier received 32-sized data from the proposed VAE encoder and included an output layer with a Softmax activation function corresponding to the six mentioned emotion classes reduced to danger and non-danger opposite classes.

B. Experimental Results

In contrast to traditional methods, our model excels by achieving remarkable results with a limited dataset while capturing intricate patterns present in genuine dangerous environments. Unlike conventional speech models that require extensive datasets for comprehensive testing, our model displays flexibility by leveraging robust, limited nuance patterns present while in the presence of danger. This adaptability ensures precise classification without distorting speaker characteristics or imposing specific positional attributes. Generative models, including our proposed model, comprehend data distributions, enabling classification without excessive reliance on additional patterns.

Nevertheless, our model encounters challenges in generalizing learned sentences across diverse data. However, it excels in recognizing similar sentences and/or noises that share common patterns.

The integration of the phonetic-prosodic regularized VAE model with speech, ambient noises, and the dangerous environment classifier results in enhanced classification accuracy compared to the vanilla VAE with our classifier. Notably, the incorporation of well-defined ambient noises such as gunfire and glass broken like, improves the classification with sad and neutral speech by reducing false positives and negatives in the classification. The proposed VAE adeptly reconstructs patterns collaborating with the classifier, automatically regenerating the input data. The classifier benefit from the latent space features of danger, impacting positively in the classification accuracy.

Our model achieves a test accuracy of 0.924 with extended data, surpassing the vanilla VAE accuracy value of 0.909, and outperforming the standard CNN-based model with 0.742. Furthermore, it delivers superior results in fewer epochs, underscoring its efficiency in accurate emotion classification. However, when compared with the RavdessDB dataset, our model with vanilla VAE with the additional classifier achieves a validation accuracy of 0.575, whereas the proposed VAE with additional classifier achieves 0.56. There are some ambient noises such as the open and close of a windows and the metallic stairs steps that could be confused and further misclassified by AI models, due to their similarity content in other danger noises such as glass broken and long distance gunfire. For human earrings could be perfectly estimated, however for AI models, there is much work to perform and great deals to enhance.

In summary, the adaptability, efficient learning, and enhanced accuracy in environmental hazardous classification make our model a promising advancement in this field.

The training loss/accuracy and the validation loss/accuracy of the vanilla VAE model, can be observed in Fig. 2. At the beginning of the accuracy result image we can observe a jumping until getting a good accuracy of 1, which is not observed in our models for training and validation. The

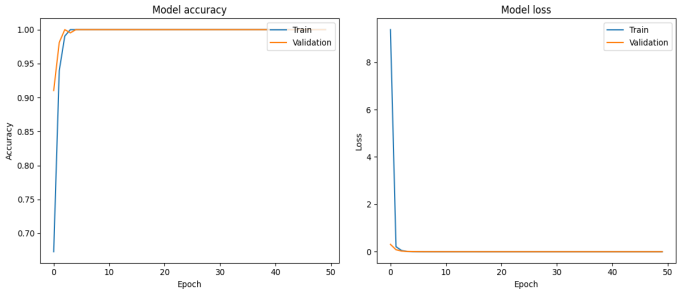


Fig. 2. Training / validation loss and accuracy over epochs of vanilla VAE model.

confusion matrices for the training and validation of the vanilla VAE model and the proposed phonetic-prosodic regularized VAE model with our danger classifier are depicted in Fig. 3 and 4, respectively. The classes "0" and "1" represent "Safe" and "Dangerous", respectively. The training confusion matrix for both models demonstrates accurate predictions for each class. The imbalance in neutral classification is due to the limited number of audios in the neutral class in the Ravdess datasets. During training 287 audios of class "0" (Safe) and 436 audios of class "1" (Danger) are correctly classified, while 97 audios of class "0" and 24 audios of class "1" are misclassifications. For validation, 83 audios of class "0" and 110 audios of class "1" are correctly classified, while 13 audios of class "0" and 6 audios of class "1" are misclassifications. The training loss/accuracy and the validation loss/accuracy of the phonetic-prosodic regularized VAE model with speech and ambient noises, can be observed in Fig. 5.

The confusion matrix for training and validation of the proposed phonetic-prosodic regularized VAE model with speech

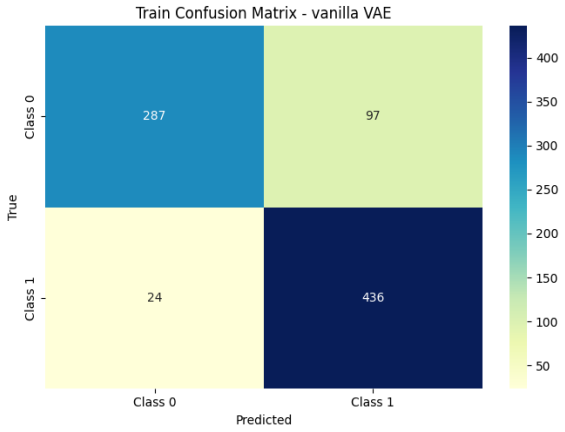


Fig. 3. Confusion matrix for training the vanilla VAE model.

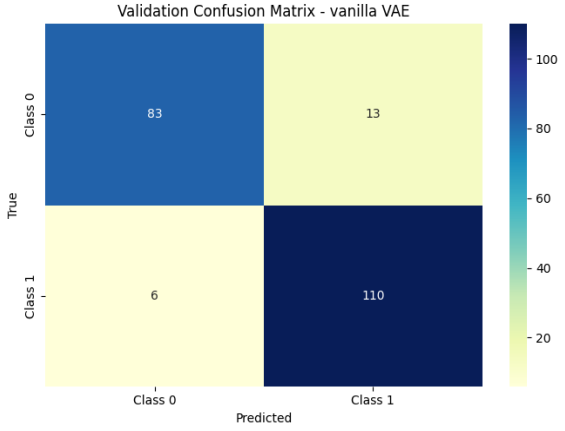


Fig. 4. Confusion matrix for validation of the vanilla VAE model.

and ambient noises, can be observed in Fig. 6 and 7. Phonetic-prosodic regularized VAE model with speech and ambient noises shows better predictions for danger class as "0" and non-danger class as "1". During training 384 audios of class "0" (Safe) and 460 audios of class "1" (Danger) are correctly classified. For validation, 96 audios of class "0" and 116 audios of class "1" are correctly classified, while there are no misclassifications.

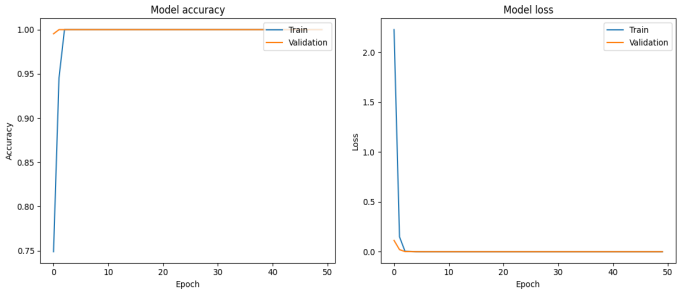


Fig. 5. Training / validation loss and accuracy over epochs of phonetic-prosodic regularized VAE model with speech and ambient noises.

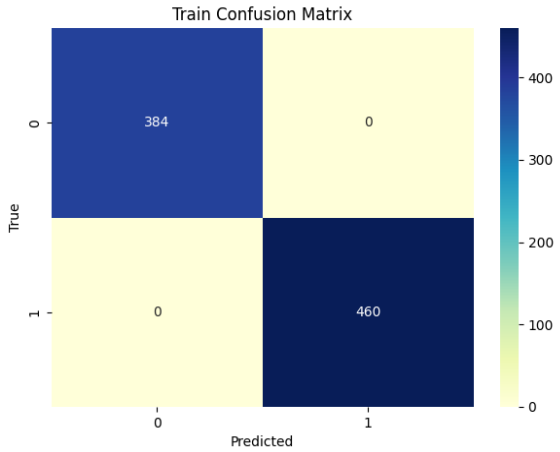


Fig. 6. Confusion matrix for training the phonetic and prosody regularized VAE model with speech and ambient noises.

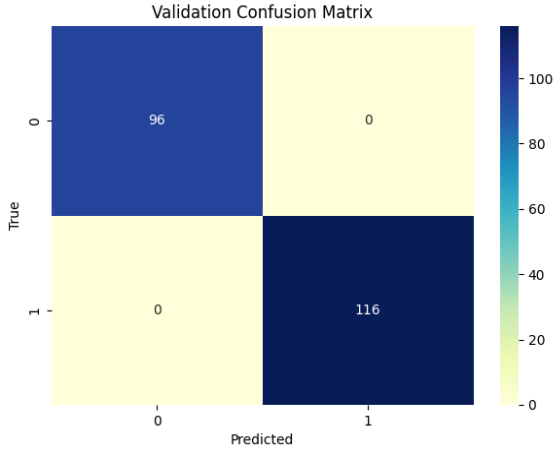


Fig. 7. Confusion matrix for validation of the phonetic and prosody regularized VAE model with speech and ambient noises.

IV. Conclusion

Our proposed model demonstrates notable success in accurate emotion classification, achieving a test accuracy of 0.924 with extended data—an improvement over the vanilla VAE accuracy of 0.909 and the standard CNN-based model's 0.742. The efficiency of our model is further underscored by its ability to deliver superior results in fewer epochs. However, when confronted with the RavdessDB dataset, our model's performance, measured by validation accuracy, shows nuances.

The vanilla VAE with an additional classifier achieves 0.575, while the proposed VAE with an additional classifier achieves 0.56. It is important to note that challenges persist, particularly in discerning ambient noises like the opening and closing of windows and metallic stair steps, which may be prone to confusion and misclassification due to their similarity with other potentially dangerous sounds. The complexity of such audio distinctions poses a challenge for AI models, necessitating ongoing efforts for improvement.

In summary, our study introduces a method leveraging generative models, specifically a variational autoencoder, for identifying hazardous environments through the analysis of emotional speech and ambient noises.

Our model, integrating phonetic and prosody features, addresses disparities between input and expected data. As part of our future research, we aim to explore areas such as background noise analysis, the separation of speech and ambient sounds, and the potential extension of our work to real-time danger processing and analysis.

REFERENCES

[1] Smailnov, Nurzhigit, et al. A Novel Deep CNN-RNN Approach for Real-time Impulsive Sound Detection to Detect Dangerous Events. International Journal of Advanced Computer Science and Applications, vol. 14, no 4, 2023.

[2] Ribino, Patrizia; Lodato, Carmelo. "A distributed fuzzy system for dangerous events real-time alerting". Journal of Ambient Intelligence and Humanized Computing, vol. 10, p. 4263-4282, 2019.

[3] Carbonneau, Marc-André, et al. "Detection of alarms and warning signals on a digital in-ear device". International Journal of Industrial Ergonomics, vol. 43, no 6, p. 503-511, 2013.

[4] Wang, Qing, et al. A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, p. 1251-1264, 2023.

[5] Wang, Qing, et al. The NERC-SLIP system for sound event localization and detection of DCASE2022 challenge. DCASE2022 Challenge, Tech. Rep., 2022.

[6] Sharath Adavanne, Archontis Politis et al. Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks.https://doi.org/10.48550/arXiv.1807.00129, Dec. 17, 2018.

[7] K. Lopatka, J. Kotus et al. "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations". Multimed. Tools Appl.,DOI 10.1007/s11042-015-3105-4, vol. 75, pp.10407-10439, 2016.

[8] Brian Gygi, Valeriy Shafiro; Environmental sound research as it stands today. Proc. Mtgs. Acoust. https://doi.org/10.1121/1.2917563, vol. 1, no. 1, June 2007.

[9] Toni Heittola, Annamaria Mesaros, Tuomas Virtanen, and Moncef Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8677–8681, 2013.

[10] Selina Chu, Shrikanth Narayanan, and CC Jay Kuo, "Environmental sound recognition with time–frequency audio features," IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1142-1158, 2009.

[11] Toni Heittola, Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen, "Context-dependent sound event detection," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2013, no. 1, pp. 1-13, 2013.

[12] Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen, "Acoustic event detection in real life recordings," in 18th European Signal Processing Conference, pp. 1267-1271, 2010.

[13] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Vir- tanen, "Polyphonic sound event detection using multi label deep neural networks," in IEEE International Joint Conference on Neural Networks (IJCNN), 2015.

[14] Elelu, Kehinde; LE, Tuyen; LE, Chau. Collision Hazard Detection for Construction Worker Safety Using Audio Surveillance. Journal of Construction Engineering and Management, vol. 149, no. 1, 2023.

[15] Svatos, Jakub; HOLUB, Jan. Impulse Acoustic Event Detection, Classification, and Localization System. IEEE Transactions on Instrumentation and Measurement, vol. 72, p. 1-15, 2023.

[16] A. Ilic Mezza, G. Zanetti, M. Cobos and F. Antonacci, "Zero-Shot Anomalous Sound Detection in Domestic Environments Using Large-Scale Pretrained Audio Pattern Recognition Models," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, doi: 10.1109/ICASSP49357.2023.10095736., pp. 1-5, 2023.

[17] K. Lopatka, J. Kotus, A. Czyzewski1, "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations", Multimed. Tools Appl., 75, pages 10407-10439, Dec. 2016.

[18] K. Zhou, B. Sisman, et Al., "Vaw-gan for the disentanglement and recomposition of emotional elements in speech", https://doi.org/10.48550/arXiv.2011.02314, Nov., 2020.

[19] A. H. Liu, et Al, "Towards unsupervised speech recognition and synthesis with quantized speech representation learning", in IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), IEEE, pp. 7259-7263, 2020.

[20] Y. Gao, R. Singh, et Al, "Voice impersonation using generative adversarial networks", in IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), IEEE, pp. 2506-2510, 2018.

[21] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization", in Proc. of ICASSP, Brighton, UK, 2019.

[22] X. Li, M. Akagi, "A three-layer perception model for valence and arousal-based detection from multilingual speech", in Proc. Interspeech 2018, pp. 3643-3647, 2018.

[23] U. Tiwari, M. Soni, R. Chakraborty, A. Panda, and S. K. Kop- parapu, "Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7194-7198, 2020.

[24] J. Zhao, S. Chen, "Speech emotion recognition in dyadic dialogues with attentive interaction modeling", in Proc. Interspeech 2019, pp. 1671-1675, 2019.

[25] J. Liu et Al., "Temporal attention convolutional network for speech emotion recognition with latent representation", in Proc. Interspeech 2020, pp. 2337-2341, 2020.

[26] K. Akuzawa, Y. Iwasawa, et Al., "Expressive speech synthesis via modeling expressions with variational autoencoder," in Proc. of Interspeech, Hyderabad, India, 2018.

[27] Xue Feng, Yaodong Zhang, and James Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, pp. 1759–1763, 2014.

[28] M. Blaauw, J. Bonada, "Modeling and transforming speech using variational autoencoders", in Proc. Interspeech 2016, pp. 1770-1774, 2016.

[29] C. H. Wu, et Al, "Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis", in IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 6, pp. 1394-1405, 2009.

[30] Y. Xu, "Speech prosody: A methodological review," Journal of Speech Sciences, vol. 1, no. 1, pp. 85-115, 2011.

[31] G. Zhang, S. Qiu, et Al., "Estimating mutual information in prosody representation for emotional prosody transfer in speech synthesis", in Proc. of ISCSLP, pp. 1-5, 2021.

[32] Latif, Siddique and Rana, Rajib and Qadir, Junaid and Epps, Julien. "Variational autoencoders for Learning latent representations of speech emotion," December 2017.

[33] D. P.. Kingma, M. Welling, "Auto-encoding variational Bayes", in Proc. 2nd International Conference on Learning Representations, 2014.

[34] R. A. Khalil, et Al., "Speech emotion recognition using deep learning techniques: A review", in IEEE Open Access journal, doi 10.1109/ACCESS.2019.2936124, vol. 7, pp. 117327-117345, 2019.

[35] Livingstone SR, Russo FA. The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One, doi: 10.1371/journal.pone.0196391, vol. 13, no. 5, May, 2018.

D2-Net: Dilated Contextual Transformer and Depth-wise Separable Deconvolution for Remote Sensing Imagery Detection

Huaping Zhou[1]
Anhui University
of Science and Technology

Qi Zhao[2*]
Anhui University
of Science and Technology

Kelei Sun[3]
Anhui University
of Science and Technology

***Abstract*—Remote sensing-based object detection faces challenges in arbitrary orientations, complex backgrounds, dense distributions, and large aspect ratios. Considering these issues, this paper introduces a novel method called D2-Net, which incorporates a transformer structure into a convolutional neural network. First, a new feature extraction module called dilated contextual transformer block is designed to minimize the loss of object information due to complex backgrounds and dense targets. In addition, an efficient approach using depth-wise separable deconvolution as an up-sampling method is developed to recover lost feature information effectively. Finally, the circular smooth label is incorporated to compute the angular loss to complete the rotated detection of remote sensing images. Experimental evaluations are conducted on the DOTA and HRSC2016 datasets. On the DOTA dataset, the proposed method achieves 79.2% and 78.00% accuracy in horizontal and rotated object detection, respectively; it achieves 94.00% accuracy in the rotated detection of the HRSC2016 dataset. The proposed model shows a significant performance improvement over other comparative models on the dataset, which verifies the effectiveness of our proposed approach.**

***Keywords*—YOLOv7; dilated contextual transformer; depth-wise separable deconvolution; circular smooth label; remote sensing**

I. Introduction

Due to advances in computer processing power, object detection has developed rapidly over the past decade. This task typically accomplishes by utilizing single-stage detectors, typified by the YOLOs models [1], [2], [3], [4], [5], and dual-stage models exemplified by the RCNN series [6], [7], [8], [9].

Despite significant advances in generic target detection, the mission in remote sensing images (RSIs) faces numerous challenges due to characteristics such as substantial variations in scale, crowded and small targets, arbitrary orientations, and large aspect ratios[10]. Therefore, detection using oriented bounding boxes (OBBs), which can handle object rotation, has become critical in remote sensing applications. Existing rotated object detection models are often constructed with pure convolutional neural networks (CNNs) or CNN-transformer hybrid structures. And the former has a lot of representative work. Pixels-IoU Loss improves performance for complex backgrounds and large aspect ratios[11] but increases training time. Rotational region convolutional neural network (R2CNN) introduces joint prediction of axis-aligned bounding boxes and inclined minimum area boxes to complete text recognition in any direction[12]. A joint image cascade (ICN) and feature pyramid network (FPN) can capture semantic features at multiple scales [13]. Adaptive period embedding (APE) proposed by Zhu et al. represented oriented targets in a novel way and length-independent IoU (LIIoU) suitable for long targets [14]. Kim B et al. developed TricubeNet, which locates oriented targets according to visual cues such as heat maps rather than oriented box offset regression [15].

Although CNNs have achieved impressive performance, they are limited by the difficulty of obtaining long-range dependencies, resulting in deficient performance in remote sensing detection. In contrast, the unique structure of the transformer allows it to compensate well for the shortcomings of CNN. Many hybrid CNN-transformer networks have achieved satisfactory results [16], [17], [18], [19], [20]. The RoI transformer technique utilizes spatial transformations on Regions of Interest (RoIs) and learns the spatial transformation parameters by using OBB annotations as supervision. This approach results in fewer mismatches during detection [16]. To address the boundary loss and spatial receptive field issues in RSIs, Dai et al. developed a rotating object detection transformer-based model (RODFormer) [17]. Another improved detector, CLT-Det, leverages correlation learning and a transformer to tackle the problem of large-scale variation and dense targets [18]. TransConvNet uses a self-attention block and CNN to aggregate broad and specific details, offsetting the CNN's lack of rotational invariance [19]. Li et al. propose an adaptive points learning method that effectively obtains geometric information for instances of arbitrary orientations [20].

The above information suggests that incorporating a transformer module into CNN can help overcome the model's difficulty in global feature modeling. And recent researches show that simple hybrid networks can acquire the same effect as many excellent complex models [21]. Therefore, this paper presents the dilated contextual transformer block (DCoT) combined with efficient layer aggregation networks (ELAN) in YOLOv7[5] to improve the model's feature extraction capability. DCoT extraction provides more feature information with a larger receptive field, allowing shallow location information to combine effectively with deep semantic information, which improves detection ability in complex backgrounds and dense objects. Second, a depth-wise separable deconvolution (DS-DeConv) module is proposed to enable the model to generate more diverse feature information during upsampling,

thereby improving its ability to detect small and dense objects. Finally, the Circular Smooth Label (CSL)[22] is integrated into the baseline YOLOv7[5] to complete the rotation detection process without being affected by boundary discontinuities. Extensive experiments were conducted on DOTA v1.0[10] and HRSC2016[23] datasets to validate the efficacy of the proposed method. The experimental results demonstrate that the proposed model enhances the detection capacity of RSIs. Moreover, it achieves real-time detection with a slight reduction in the number of parameters, striking a balance between accuracy and speed.

The main contributions of this paper can be summarized as follows. First, we use DCoT to improve the model's ability to obtain contextual information, which enhances the model's ability to detect complex backgrounds and dense targets in RSIs. Second, we use DS-DeConv for upsampling, which effectively preserves detailed feature information, enhancing the model's detecting ability of small objects. Finally, CSL is integrated into YOLOv7 to complete the rotated detection of multi-directional objects in RSIs. The proposed model outperforms other comparative models in detection.

The upcoming sections are structured as follows. Section 2 the related work, including pure CNN and CNN-transformer hybrid detection models. Section 3 provides a detailed description of the proposed methods integrated into the D2-Net. Section 4 is the experimental details and analyses of the experimental results. Finally, the conclusion is presented in Section 5.

II. Related Work

A. Pure CNN Detection Models

Depending on whether models generate region proposals, detection models consist of two types: single-stage detection methods and two-stage detection algorithms.

The single-stage detection methods directly predict object class and location without region proposal, resulting in faster inference and lower computational complexity than two-stage models. Redmon J et al. proposed the first generation of YOLO [1], which starts with real-time object detection time. This model views target recognition as a regression task and detects the presence of an object by determining whether the object's center point falls within a particular grid cell, which is obtained by dividing the image into multiple grid cells. Inevitably, it cannot solve problems of dense, small, and large aspect ratio targets and other issues that inspire other researchers to make further progress. To improve the accuracy of small target detection, SSD [24] feeds multiple features extracted from different layers of the feature extraction model to the object prediction module. It also simplifies the training process for targets with different shapes by assigning different scales and aspect ratios to the prior bounding boxes associated with each grid cell. The method used convolutional layers instead of fully connected layers and produced the same results as contemporaneous two-stage detection models. More recently, an enhanced SSD [25] introduces interactive multiscale attention to acquiring more effective feature representation capability. Retinanet [26] incorporates focal loss and effectively addresses the class imbalance problem, resulting in high speed and accuracy performance.

Two-stage detectors also gained significant attention due to their remarkable accuracy and robustness. RCNN[6] treats the detection task as a classification problem. In the first stage, it extracts region proposals from each image, then predicts targets' categories after computing features in CNN. FPN [27] regards layers with consistent feature map sizes as a stage and achieves the top-down integration of multi-scale feature maps through successive stages. It distributes features based on object scale, merging deep-level semantic information with shallow-level fine-grained information to perform more accurately. Mask R-CNN [9] innovates RoI alignment to mitigate date missed owing to feature quantization during the RoI pooling process.

B. Transformer Detection Models

Since transformers were introduced to computer vision, many distinctive models emerged. Vision transformers divide the image into multiple patches, provide them with positional embedding, and then feed the feature information into the head for detection.[28] This allows the model to be independent of image size. DINO improves DETR-like models in terms of performance and efficiency by using a comparative denoising training method, a hybrid query selection method for anchor initialization, and a look-forward double scheme for box prediction.[29] Biformer proposes a novel dynamic sparse attention via bi-level routing for more flexible computational allocation and content awareness, enabling dynamic query-aware sparsity.[30]

C. CNN-Transformer Hybrid Detection Models

The emergence of the Transformer structure compensates for the shortcomings of the pure CNN structure in obtaining long-range dependencies and contextual information, leading to numerous Transformer-related models. However, the pure transformer models have high memory consumption and complexity. So more models fuse the transformer module with CNN by insertion or replacement to achieve a balance. RoI transformer [16] conducts spatial transformations on RoIs, learning transformation parameters supervised by OBB annotations, which solves dense RSI targets and RoI-target mismatches. RODFormer [17] addresses boundary loss and spatial receptive field lack in RSI detection via a structured transformer model. CLT-Det [18] presents a correlation learning detector for solving the problem of large-scale variation and dense targets. TransConvNet [19] merges a self-attention block and CNN, aggregating the detailed and specific information to compensate for the CNN's deficiency in rotational invariance. Li et al. proposed a robust adaptive points learning methodology to extract the geometric information of instances of arbitrary orientations [20].

To summarize, the combination of transformer and CNN can effectively overcome the limitation of CNN structures in capturing features at varying scales and improve the accuracy and robustness of object detection.

III. Methods

In the following parts of this section, we begin with a short introduction to the overall architecture of the proposed D2-Net, taking YOLOv7 [5] as the baseline model. Next, we present

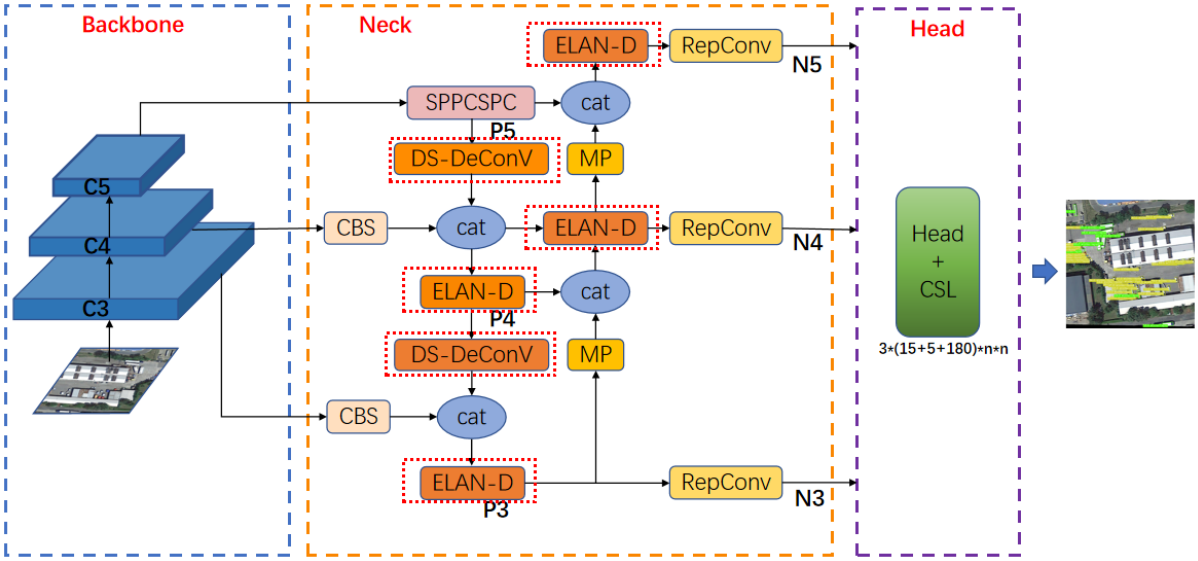


Fig. 1. The overall structure of our network. The SPPCPSC,MP, RepConv are modules of the original YOLOv7. And the detailed composition of each block in Fig. 1 is illustrated in Fig. 2 and Fig. 3.

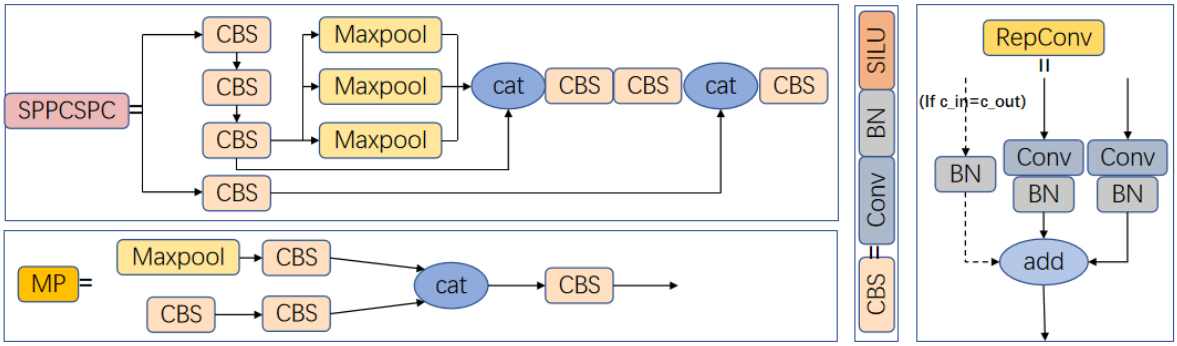


Fig. 2. Detailed block consistency of the neck network. The ELAN-D module is depicted in Fig. 4.

a detailed description of the DCoT block and the depth-wise separable deconvolution. Finally, we briefly discuss the CSL [22], which is integrated into our model to accomplish the task of rotation detection. Fig. 1 and Fig. 2 show the overall structure and detailed block consistency, respectively.

A. The D2-Net Structure

As depicted in Fig. 1, the backbone network extracts feature maps c_i, which are then sent to the neck network, where $i = 3, 4, 5$ represents the level of features, and C_i has a stride of 2^i and is $1/2^i$ pixel density of the input image size $W \times H$. The neck network consists of two modules. The initial component is the FPN [25] architecture, which propagates semantic features from higher to lower resolutions. The second module utilizes the PAFPN [31] module. To compensate for the loss of fine-grained information caused by resolution reduction, an ascending feature merging is employed to transfer location details to feature maps at deeper layers. Furthermore, depth-wise separable deconvolution makes the most suitable up-sample method by itself, and the improved ELAN module is adopted to improve the reception capability of contextual information of the network. Different scales feature maps containing detailed semantic and rich localization information are output to the RepConv block. Finally, the head network with CSL predicts object categories and position information regarding the angular problem as classification.

In our method, after being processed by the improved neck network, the output feature representations with various resolutions achieve a balance between semantic information in deep and shallow spatial details, leading to improve detection performance.

B. Contextual Transformer Block with Dilated Convolution

Drawing inspiration from the self-attention mechanism in Transformer models, numerous scholars have investigated the effectiveness of hybrid networks mixed by CNNs and transformers in computer vision task scenarios [16], [17], [18], [19]. And as existing researches prove, through the simple fusion of CNNs and transformers, object detection models pay more attention to more useful features so that the performances of those models are improved. Therefore, the hybrid network,

including the transformer module, has a good prospect in the RSIs detection task.

The traditional self-attention modules utilize input feature information obtained from various spatial positions to process input data. Nevertheless, these modules acquire knowledge of all possible query-key connections by training on individual query-key pairs. This process occurs independently, without considering the contextual information between their interactions. The CoT [32] architecture can integrate abundant contextual information and Contribute significantly to the visual representation of 2D images. Nevertheless, the standard convolution operation will lose much localization information in feature processing. Therefore, we replace it with dilated convolution to form DCoT, which effectively makes the network increase the receptive field while obtaining more information. Then we displace the last three CBS modules of ELAN with DCoT to form ELAN-D (see Fig. 4), which reduces the calculation amount and FLOPs. By combining the strengths of the Transformer and CNN, the DCoT module can capture both global and detailed local information from input features. This approach improves the network model's ability to represent input information features, leveraging the advantages of each component. It showed the architecture of the DCoT block in Fig. 3.

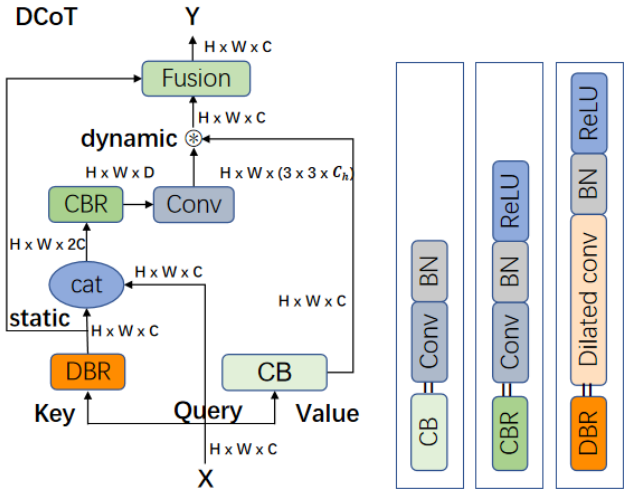


Fig. 3. The detailed structure of the DCoT block and its module. H, W, and C denote the height, width, and number of channels of the input data X, $\circledast$ denotes local matrix multiplication.

For input feature X, it is processed through three pathways, namely Q(queries), K(keys), and V(values), to generate more feature information. The keys undergo dilated convolution to capture local information and increase the receptive field. Then, K is concatenated with X to supplement local information and passed through a CBR module and a standard convolution to generate Q. Finally, Q is multiplied with V and fused with K to obtain the final output Y. The Q, K, and V can be written as:

$$Q = [K, X]W_{CBR}W_C \tag{1}$$

$$K = XW_{DBR} \tag{2}$$

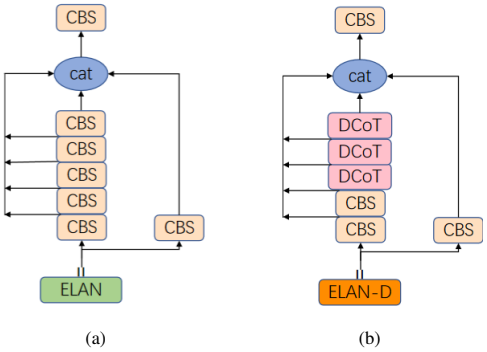


Fig. 4. The architectures of the ELAN block and ELAN-D block. (a) shows the detailed ELAN structure, and (b) shows the detailed ELAN-D block.

$$V = XW_{CB} \tag{3}$$

where X is the input feature, $W_{\square}$ are different convolutional blocks.

C. Depth-wise Separable Deconvolution for Up-sampling

During object detection with deep learning, the resolution of the feature map tends to decrease as the network deepens, leading to a loss of information. Thus, up-sampling is essential for an algorithm. In the YOLO algorithms, nearest neighbor interpolation is employed for up-sampling. However, focusing solely on the nearest pixels has also resulted in image quality and details loss, especially for tiny targets. Deconvolution is also a commonly used up-sampling method. Compared with neighbor interpolation, it performs better than in preserving feature information. However, it produces more parameters as well. Deep separable convolution [33] disassembles traditional convolution into depth convolution and point convolution, which can make the model more efficient and parameter reduction.

In this paper, we propose the DS-DeConv block for up-sampling. With this method, more diverse pixel values can be produced when recovering the feature map's resolution, which makes the Acquired feature map preserve more details and features of the original feature map.

We also introduce group convolution and change the filter size of deconvolution to decrease the parameter quantity caused by deconvolution. Our DS-DeConv method improves network model accuracy in up-sampling with a slight increase in parameters. Fig. 5 illustrates the principal diagram of DS-DeConv, while the number of deconvolution groups is adjusted based on the channel quantities in the network.

D. Rotationally Detection

Currently, bounding boxes in object detection consist of HBBs, rotated bounding boxes, and custom bounding boxes. The characteristics of remote sensing detection include the random and diverse directions of the objects to be detected. And to achieve more accurate detection of these rotating objects, the rotating bounding box is used for it.

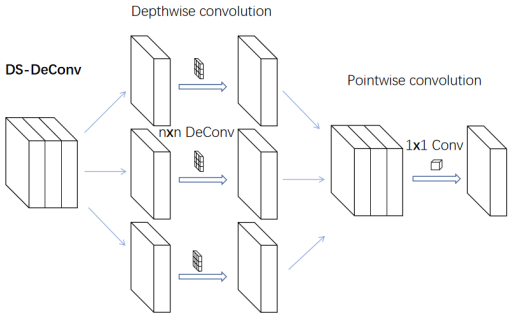


Fig. 5. The structure of depth-wise separable deconvolution.

The rotated detection method based on parametric regression mainly consists of the five parameters and the eight-parameter method. However, in rotation detection, the target parameters for learning are periodic, which causes the learned parameters to be located at the boundary periodicity, resulting in discontinuity issues and an abrupt rise of loss. Therefore, we use CSL [22] to solve the boundary discontinuity problem, as depicted in Fig. 6.

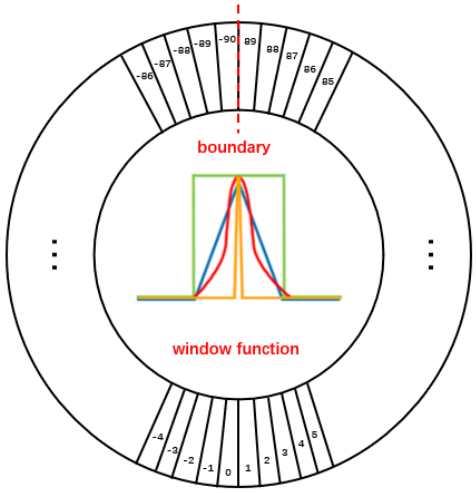


Fig. 6. The schematic diagram of the CSL.

The CSL is expressed as follows:

$$\mathrm{CSL}_{(x)} = \begin{cases} g(x), & \theta - r < x < \theta + r \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $g(x)$, r, and θ represents the window function, radius, and the current bounding box angle, respectively. By converting angle prediction from a regression task to a classification task, the boundary discontinuity issue can be effectively resolved with minimal loss of accuracy.

IV. Experiments and Results Analysis

A. Datasets

1) DOTA Dataset: The DOTA dataset [10] contains 2806 high-resolution aerial images collected from various sensors and platforms and encompasses 15 categories. It is split into three subsets for training, validation, and testing, including 1411 images, 458 images, and 937 images, respectively, containing 188282 instances in total. The image size varies from 800×800 to 4000×4000 pixels.

2) HRSC2016 Dataset: The HRSC2016 dataset[23] includes 1061 remote sensing images from six distinct ports. The dataset is divided into three parts, 436 images for training (a total of 1207 labeled examples), 444 images for testing (a total of 1228 labeled examples), and 181 images for validation (a total of 541 labeled examples). The images have varying resolution, ranging from 300 × 300 to 1500 × 900 pixels.

B. Implementation Details and Evaluation Index

Considering the adverse influence of high and inconsistent resolution images, we reprocess the original data of these two datasets. For the DOTA dataset, we cropped the images to 1024×1024 resolution with 200 pixels overlapping area. Then 15749 images were extracted for training and 5297 images for evaluation, and the final test results are obtained through the official evaluation server. The network is trained with the SGD optimizer in the training process. The lr (learning rate) is 0.001, and momentum and weight decay are 0.937 and 0.0005. We train 300 epochs with batch size 16 on two GeForce RTX 3090 GPUs. For the HRSC2016 dataset, we resized all the images to (768, 768). The network is trained with the SGD optimizer for training. The learning rate is 0.01, and momentum and weight decay are 0.937 and 0.0005. We train 200 epochs with batch size 8 on GeForce RTX 3060 GPU.

We adopt the Average Precision (AP) and the mean AP (mAP @0.5) metric in the comparative experiments to evaluate the multi-class detection accuracy. They can be calculated as follows:

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{5}$$

$$\mathrm{AP} = \int_0^1 Pd_r \tag{6}$$

$$\mathrm{mAP} = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{7}$$

TP is the correctly classified target number, while FP is the background number recognized as target. The accuracy rate P can be defined as the proportion of correctly detected targets among all detection results. The mAP is the average of AP values of all classes. In the ablation experiments, FLOPs and speed are also used to estimate the differences in algorithm capability. Speed is also used to estimate the differences in algorithm capability.

C. Ablation Experiments

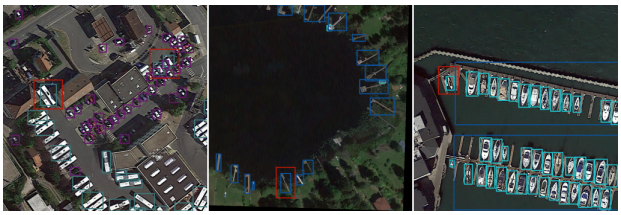
In this section, we choose YOLOv7 as the baseline model to conduct ablation experiments on the DOTA dataset to verify the effectiveness of the introduced DCoT block, DS-DeConv, and CSL. It should be noted that this paper aims to address the problem of rotated RSI detection, so unnecessary ablation experiments on horizontal detection are not shown. The batch

TABLE I. THE RESULT OF THE ABLATION EXPERIMENT

\	YOLOv7	CSL	DS-DeConv	DCoT	FLOPs(G)	Speed(ms)	mAP/HBB(%)	mAP/OBB(%)
①	✓				**103.4**	**90.9**	73.70	\
②	✓	✓			106.5	43.7	75.60(+1.90)	74.71
③	✓	✓	✓		106.5	45.2	77.10(+1.50)	75.12(+0.41)
④	✓	✓		✓	106.4	39.4	76.4(-0.7)	75.76(+0.64)
⑤	✓	✓	✓	✓	106.4	39.4	**79.20(+2.10)**	**77.96(+2.84)**

TABLE II. THE DETAILED RESULT OF THE ABLATION EXPERIMENT. PL: PLANE, BD: BASEBALL DIAMOND, BR: BRIDGE, GFT: GROUND FIELD TRACK, SV: SMALL VEHICLE, LV: LARGE VEHICLE, SH: SHIP, TC: TENNIS COURT, BC: BASKETBALL COURT, ST: STORAGE TANK, SBF: SOCCER-BALL FIELD, RA: ROUNDABOUT, HA: HARBOR, SP: SWIMMING POOL, HC: HELICOPTER.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
①	93.80	73.40	48.00	**72.90**	71.50	88.80	89.50	94.90	72.10	**76.80**	**67.50**	57.60	85.80	62.40	50.50	73.70
②	**98.40**	81.70	**50.90**	59.00	87.50	92.10	97.50	96.80	85.70	70.30	48.10	57.80	**87.60**	64.00	57.10	75.60
③	**98.40**	81.60	48.90	58.60	87.00	91.10	97.20	97.20	82.70	78.70	50.20	**59.10**	86.50	63.80	75.90	77.10
④	97.30	78.40	47.90	65.00	84.60	91.20	97.00	96.60	85.40	75.70	57.80	**59.50**	84.70	58.60	66.90	76.40
⑤	98.30	**84.50**	47.90	61.70	**87.70**	**92.90**	**97.50**	**98.10**	**88.00**	76.40	57.60	58.40	86.30	**66.40**	**86.40**	**79.20**



(a)

(b)

Fig. 7. Some contrastive detection results. (a) is the result of the baseline; (b) is the result of the D2-Net. And the differences are highlighted in red.

size for training was 16, and the performance metrics were evaluated every 10 epochs during the training process. A total of 300 iterations were completed to train both the baseline and improved models. FLOPs, speed, and mAP are used as evaluation indicators in the experiments. Table I shows the results of our improvements and Table II shows detailed AP values of each category conducted on the DOTA dataset. And the bold font is the best result.

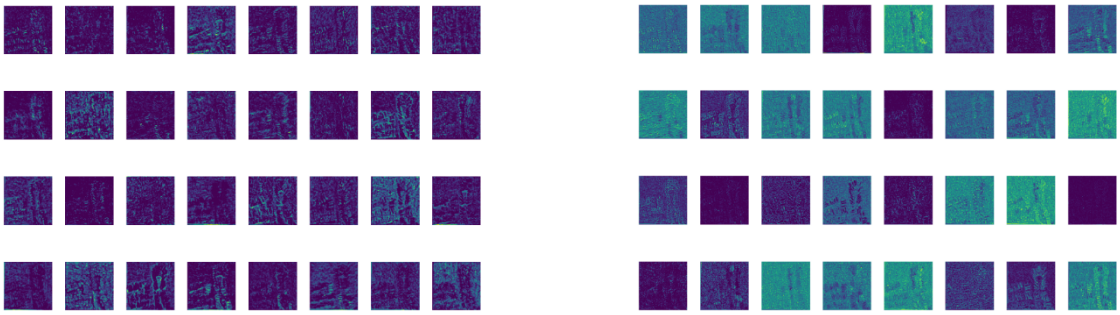
As seen from Table I speed and mAP of the OBB task are commonly lower than those in the HBB task, which is attributed to the angle issue when serving the rotated detection task. Attentively, to ensure the effectiveness of the baseline, its experiments were all performed at $640 * 640$ resolution, while other experiments were conducted at $1024 * 1024$ resolution. And the baseline speed is 40.98 at $1024 * 1024$ resolution. Despite the speed and mAP having decreased, the effect has been improved in the actual detection(see Fig. 7). In the horizontal task, compared with the original YOLOv7, ②③④⑤ showed improvement of 1.9%, 3.4%, 2.7% and 5.5%. Relative to the YOLOv7 with CSL added, ③④⑤ achieved 0.41%, 1.05% and 3.25% improvement. According to Table II, it can be found that the proposed method has greatly improved in the categories of small vehicles, harbors, and ships, obtaining 16.2%, 35.9%, and 8% improvement, respectively, compared with the baseline model.

In Fig. 7, three images are chosen for comparing the detection results from the dataset. The results of the two rows are the baseline model, and the D2-Net model proposed in this paper, respectively. There are plenty of small and dense objects in the leftmost images of Fig. 7(a) and Fig. 7(b). It can be seen from the red highlights that the baseline model loses some targets, while the proposed model detects them very effectively. The background of the middle image is similar to the object, and the baseline's results are affected, while the proposed model works well. The right image contains many targets with large aspect ratios, and the D2-Net is more accurate than the baseline when boxing targets and no targets are lost.

In Fig. 8(a) and Fig. 8(b), to prove the feature extraction capability of the DcoT modules, we made the first 32 feature maps visualization in the same stage of both baseline and the D2-Net. It can be observed that the proposed model can effectively eliminate irrelevant information from the background and has good extraction capability for detecting targets. It is the DCoT modules that enable the network to fully utilize feature information and concentrate on detecting targets with distinguishable features. Fig. 9(a) and Fig. 9(b) show the first upsampling heatmaps of the baseline and D2-Net. The latter preserves more useful feature information around objects and eliminates unnecessary noise. It proves the DS-DeConv's effectiveness when detecting small targets.

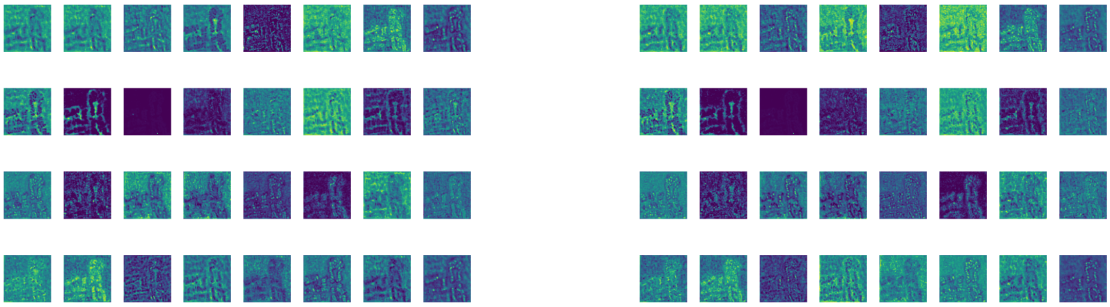
D. Comparison with other OBB Methods

In this section, we choose the YOLOv7 as the baseline. We compare our model performance with other state-of-the-art methods for the DOTA-v1.0 and HRSC2016 datasets. In compared models, RoI Trans [16], RODFormer [17], and CLT-Det [18] adopted a hybrid network using CNNs and transformer blocks, and the others applied pure CNNs structure.



(a) (b)

Fig. 8. The DCoT visualizations of the first 32 features: (a) represents the baseline result, and (b) is the result of the D2-Net.



(a) (b)

Fig. 9. The first upsampling visualizations of the first 32 features: (a) represents the baseline result, and (b) is the result of the D2-Net.

Fig. 10. Visualization of the detection results of our method on the DOTA data set.

TABLE III. OBB TASK PERFORMANCE COMPARISONS ON THE DOTA-V1.0 TEST SET (AP (%) FOR EACH CATEGORY AND OVERALL MAP @0.5 (%). IN THE COLUMN, THE BOLD DENOTES THE BEST DETECTION RESULTS

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
OBB																
R2CNN[12]	80.94	65.67	35.34	67.44	59.52	50.91	55.81	90.67	66.92	72.39	55.06	52.23	55.14	53.35	48.22	60.67
RADet[34]	79.45	76.99	48.05	65.83	65.46	74.40	68.86	89.70	78.14	74.97	49.92	64.63	66.14	71.58	62.16	69.09
Axis Learning[35]	79.53	77.15	38.59	61.15	67.53	70.49	76.30	89.66	79.07	83.53	47.27	61.01	56.28	66.06	36.05	65.98
RoI Trans[16]	88.64	78.52	43.44	75.92	75.92	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
DRN[36]	89.71	82.34	47.22	64.10	76.22	74.43	85.84	90.57	86.18	84.89	57.65	61.93	69.30	69.63	58.48	73.23
RODFormer[17]	89.76	79.64	**56.61**	71.57	78.60	85.29	**89.93**	90.53	87.73	83.05	60.19	60.34	66.03	69.75	64.95	75.60
CLT-Det[18]	89.31	**85.69**	53.97	**77.11**	79.66	79.01	88.55	**90.89**	85.36	86.56	63.92	68.47	75.65	70.65	66.91	77.45
CSL[22]	90.25	85.53	54.64	75.31	70.44	73.51	77.62	90.84	**86.15**	86.69	69.60	68.04	73.83	71.10	68.93	76.17
Ours	**90.54**	82.99	43.20	55.39	**81.71**	**87.40**	89.82	90.80	83.58	**89.71**	**76.62**	**72.27**	75.00	67.57	**82.87**	**77.96**
HBB																
GraphFPN[37]	89.32	68.88	50.41	60.42	70.91	79.45	86.18	90.80	83.11	80.35	53.01	60.98	75.95	64.36	58.71	71.52
SCRDet[38]	90.18	81.88	55.30	**73.29**	72.09	77.65	78.06	90.91	82.44	86.39	64.53	63.45	75.77	78.21	60.11	75.35
Mask OBB[39]	89.69	**87.07**	**58.51**	72.04	78.21	71.47	85.20	89.55	84.71	**86.76**	54.38	**70.21**	78.98	77.46	70.40	76.98
YOLOv7[5]	93.80	73.40	48.00	72.90	71.50	88.80	89.50	94.90	72.10	76.80	**67.50**	57.60	85.80	62.40	50.50	73.70
CGL[40]	89.53	82.85	56.53	**76.52**	79.29	83.39	88.19	90.90	86.67	85.07	63.40	68.23	77.82	**78.77**	50.23	77.16
Ours	**98.30**	84.50	47.90	61.70	**87.70**	**92.90**	**97.50**	**98.10**	**88.00**	76.40	57.60	58.40	**86.30**	66.40	**86.40**	**79.20**

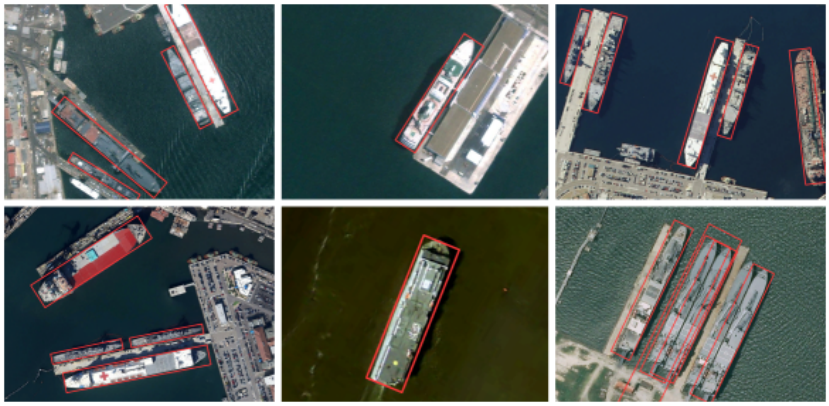


Fig. 11. The visualization of the detection results of our method on the HRSC2016 dataset.

1) Results on DOTA-v1.0: As reported in Table III, The comparative experiments on the DOTA dataset consist of the OBB and HBB tasks. In the OBB task, we achieved the mAP of 77.96%, which gains 1.79% higher than the CSL with CNNs structure, and 0.51% higher than CLT-Det with a hybrid framework. Moreover, the prediction performance on densely distributed small objects, like storage tanks and small vehicles, has improved enormously, reaching 89.71% and 81.71%, which are 3.02% and 2.05% higher than the second best, respectively. Besides, soccer ball fields, large vehicles, and helicopters also perform well, reaching 76.62%, 87.4%, and 82.87%, respectively. In the HBB task, the proposed model is 5.5% (from 73.70 to 79.20%) higher than the baseline. The top-3 mAP is plane, tennis court, and ship, achieving 98.3%, 98.1%, and 97.5%, respectively. In general, the above statement demonstrates the effectiveness of our model, and Fig. 10 visualizes some detection results of our method on the DOTA dataset.

2) Results on HRSC2016: The HRSC2016 dataset consists of plenty of oriented ships. As shown in Table IV, many classical detection algorithms have attained excellent performance in this dataset, such as R2CNN [12], RoI Trans [16], CLT-Det [18], CSL [22], Axis Learning [35], and Oriented R-CNN [41]. Our model uses the A block for enhancing feature extraction and depth-wise separable deconvolution for upsampling. It achieves an mAP value of 94.00% with the 768 × 768 resolution, outdoing several of the mentioned methods. And the visualization of some detection results is depicted in Fig. 11.

TABLE IV. PERFORMANCE COMPARISONS ON THE HRSC2016 OBB TASK. THE BEST RESULT IS HIGHLIGHTED IN BOLD

Methods	mAP	Resolution
R2CNN[12]	73.07	800×800
Axis Learning[35]	78.15	800×800
RoI Trans[16]	86.20	512×800
SLA[42]	89.51	768×768
CLT-Det[18]	89.72	512×800
CSL[35]	89.62	800×800
Oriented R-CNN[41]	90.50	1333×800
Attention-Points[43]	90.59	1333×800
Ours	94.00	768×768

V. CONCLUSIONS

In this paper, we proposed an effective one-stage model called D2-Net for rotated remote sensing image detection based on the YOLOv7 model. we innovate the DCoT block combining dilated convolution with contextual transformer block for feature extraction and enhancing the ability to detect

Objects with tiny sizes and dense distribution of RSIs, which can fully utilize the global and local information of objects and enlarge the receptive field. Then, We designed the DS-DeConv for up-sampling, which mitigates the effects of complex backgrounds and low resolution. It improves the resolution and quality of the up-sampled feature maps, enabling the detector to capture the details and shapes of the targets more effectively. Additionally, the CSL is employed for determining the angle loss and accomplishing the prediction of rotated objects in RSIs. In the end, we conducted experiments on the DOTA and HRSC2016 datasets to prove the effectiveness of D2-Net. Although detection capability surpasses other commonly employed algorithms, the speed and FLOPs has decreased. Thus, we will further enhance the feature representation and improve the model's detection speed with a more lightweight model.

ACKNOWLEDGMENT

The authors thank Anhui University of Science and Technology for its support. Secondly, they thank ITVR-AUST laboratory for its support.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779-788.

[2] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263-7271.

[3] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[4] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[5] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2023, pp. 7464-7475.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580-587.

[7] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440-1448.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, 2015.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961-2969.

[10] G. S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3974-3983.

[11] Z. Chen, K. Chen, W. Lin, J. See, H. Yu, Y. Ke, and C. Yang, "Piou loss: Towards accurate oriented object detection in complex environments," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, 2020, pp. 195-211.

[12] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2CNN: Rotational region CNN for orientation robust scene text detection," arXiv preprint arXiv:1706.09579, 2017.

[13] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III, 2019, pp. 150-165.

[14] Y. Zhu, J. Du, and X. Wu, "Adaptive period embedding for representing oriented objects in aerial images," IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 10, pp. 7247-7257, 2020.

[15] B. Kim, J. Lee, S. Lee, D. Kim, and J. Kim, "TricubeNet: 2D kernel-based object representation for weakly-occluded oriented object detection," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 167-176.

[16] J. Ding, N. Xue, Y. Long, G. S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2849-2858.

[17] Y. Dai, J. Yu, D. Zhang, T. Hu, and X. Zheng, "RODFormer: High-Precision Design for Rotating Object Detection with Transformers," Sensors, vol. 22, no. 7, p. 2633, 2022.

[18] Y. Zhou, S. Chen, J. Zhao, R. Yao, Y. Xue, and A. El Saddik, "CLT-Det: Correlation learning based on transformer for detecting dense objects in remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-15, 2022.

[19] X. Liu, S. Ma, L. He, C. Wang, and Z. Chen, "Hybrid network model: Transconvnet for oriented object detection in remote sensing images," Remote Sensing, vol. 14, no. 9, p. 2090, 2022.

[20] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented reppoints for aerial object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1829-1838.

[21] Y. Zhao, G. Wang, C. Tang, C. Luo, W. Zeng, and Z. Zha, "A battle of network structures: An empirical study of cnn, transformer, and mlp," arXiv preprint arXiv:2108.13002, 2021.

[22] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16, 2020, pp. 677-694.

[23] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," 2017.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, 2016, pp. 21-37.

[25] S. Zhou and J. Qiu, "Enhanced SSD with interactive multi-scale attention features for object detection," Multimedia Tools and Applications, vol. 80, pp. 11539-11556, 2021.

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980-2988.

[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117-2125.

[28] Dosovitskiy, A. et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[29] Zhang, H. et al., "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," arXiv preprint arXiv:2203.03605, 2022.

[30] Zhu, L. et al., "BiFormer: Vision Transformer with Bi-Level Routing Attention," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10323-10333, 2023.

[31] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759-8768.

[32] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[33] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[34] Y. Li, Q. Huang, X. Pei, L. Jiao, and R. Shang, "RADet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images," Remote Sensing, vol. 12, no. 3, p. 389, 2020.

[35] Z. Xiao, L. Qian, W. Shao, X. Tan, and K. Wang, "Axis learning for orientated objects detection in aerial images," Remote Sensing, vol. 12, no. 6, p. 908, 2020.

[36] X. Pan, Y. Ren, K. Sheng, W. Dong, H. Yuan, X. Guo, C. Ma, and C. Xu, "Dynamic refinement network for oriented and densely packed object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11207-11216.

[37] G. Zhao, W. Ge, and Y. Yu, "GraphFPN: Graph feature pyramid network for object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2763-2772.

[38] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, and K. Fu, "Scrdet: Towards more robust detection for small, cluttered and rotated objects," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8232-8241.

[39] J. Wang, J. Ding, H. Guo, W. Cheng, T. Pan, and W. Yang, "Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images," Remote Sensing, vol. 11, no. 24, p. 2930, 2019.

[40] X. Chen, C. Wang, Z. Li, M. Liu, Q. Li, H. Qi, D. Ma, Z. Li, and Y. Wang, "Coupled Global–Local object detection for large VHR aerial images," Knowledge-Based Systems, vol. 260, p. 110097, 2023, Elsevier.

[41] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3520-3529.

[42] Q. Ming, L. Miao, Z. Zhou, J. Song, and X. Yang, "Sparse label assignment for oriented object detection in aerial images," Remote Sensing, vol. 13, no. 14, p. 2664, 2021.

[43] C. T. C. Doloriel and R. D. Cajote, "Improving the Detection of Small Oriented Objects in Aerial Images," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 176-185.

Semantic Embeddings for Arabic Retrieval Augmented Generation (ARAG)

Hazem Abdelazim
School of Computing and Digital Technology
ESLSCA University
Cairo, EGYPT

Mohamed Tharwat
School of Computing and Digital Technology
ESLSCA University
Cairo, EGYPT

Ammar Mohamed
School of Computing and Digital Technology
ESLSCA University
Cairo, EGYPT

***Abstract*—In recent times, Retrieval Augmented Generation (RAG) models have garnered considerable attention, primarily due to the impressive capabilities exhibited by Large Language Models (LLMs). Nevertheless, the Arabic language, despite its significance and widespread use, has received relatively less research emphasis in this field. A critical element within RAG systems is the Information Retrieval component, and at its core lies the vector embedding process commonly referred to as "semantic embedding". This study encompasses an array of multilingual semantic embedding models, intending to enhance the model's ability to comprehend and generate Arabic text effectively. We conducted an extensive evaluation of the performance of ten cutting-edge Multilingual Semantic embedding models, employing a publicly available ARCD dataset as a benchmark and assessing their performance using the average Recall@k metric. The results showed that the Microsoft E5 sentence embedding model outperformed all other models on the ARCD dataset, with Recall@10 exceeding 90%**

***Keywords*—Arabic NLP; large language models; retrieval augmented generation; semantic embedding**

I. Introduction

Retrieval Augmented Generation (RAG), introduced by Facebook Researchers in 2020 [1], is a pivotal AI framework facilitating information retrieval for Generative AI models, thereby enhancing their accuracy and capabilities. RAG empowers Large Language Models (LLMs) by granting them access to external knowledge sources, augmenting the content generation process. This dual functionality entails retrieval, wherein RAG meticulously selects pertinent information from provided sources and generation, whereby LLMs craft contextually relevant responses based on user input.

The advantages of RAG are multi-fold. Firstly, it bolsters the performance by grounding LLMs with factual, up-to-date information from external knowledge repositories. Furthermore, RAG maintains contextual relevance in responses, contributing to a more engaging user experience in conversational AI applications. Its scalability is noteworthy, as RAG models seamlessly handle copious volumes of information, proving invaluable for data-intensive tasks. Additionally, the adaptability of RAG models allows fine-tuning for specific applications [2], rendering them versatile across diverse data and use cases. Customizability is another hallmark, permitting RAG models to specialize in particular domains or subjects through customization and fine-tuning on specific knowledge bases. Due to the importance of such a framework for enterprises, extensive research is currently being pursued to discover new algorithms and techniques to enhance the performance of such models bounded by the context-window limitations of LLMs. Although there is ongoing research to expand the window size for LLM to be able to ingest more data in the prompt, the use of techniques like RAG is still of great practical importance, not only on homogeneous unstructured data but also on heterogeneous data [3].

In principle, at the heart of the information retrieval module is the semantic embedding module which converts a piece of text, whether a query or a context text chunk to a numeric feature vector that embodies all semantic features of the text. The development of word and sentence embeddings is a relatively recent area of research in natural language processing (NLP) and information retrieval.

Most of the semantic models are English language-centred; however, in recent years, Multilingual embedding models were released [4]. There are lots of benchmarks to test the performance of multilingual embeddings [5], which are aggregate but very few focus on language-specific performance, and on the Arabic language in particular. This is the main impetus behind the current research work, which focuses on ten different state-of-the-art embedding models that are capable of embedding Arabic language.All the models are tested using publicly available ARCD (Arabic Reading Comprehension Dataset) [6] and the metric used is average Recall@k for different values of k. A comparative performance is conducted taking into consideration the embedding size for each model.

The rest of the paper is organized as follows: Section II, the Retrieval Augmented generation pipeline is presented, as well as the positioning of the semantic embedding and the information retrieval component within the pipeline. Section III explores related research work in this field, focusing on recent developments. Section IV overviews the 10 semantic embedding models that are used in the experiments will be covered. Section V discusses the 10 embedding models on a standard dataset that are used in for Arabic Reading comprehension (ARCD) and their evaluation using Recall@k performance metric. Also, the impact of the embedding dimension size is analyzed in the comparative results. Section VI concludes the paper.

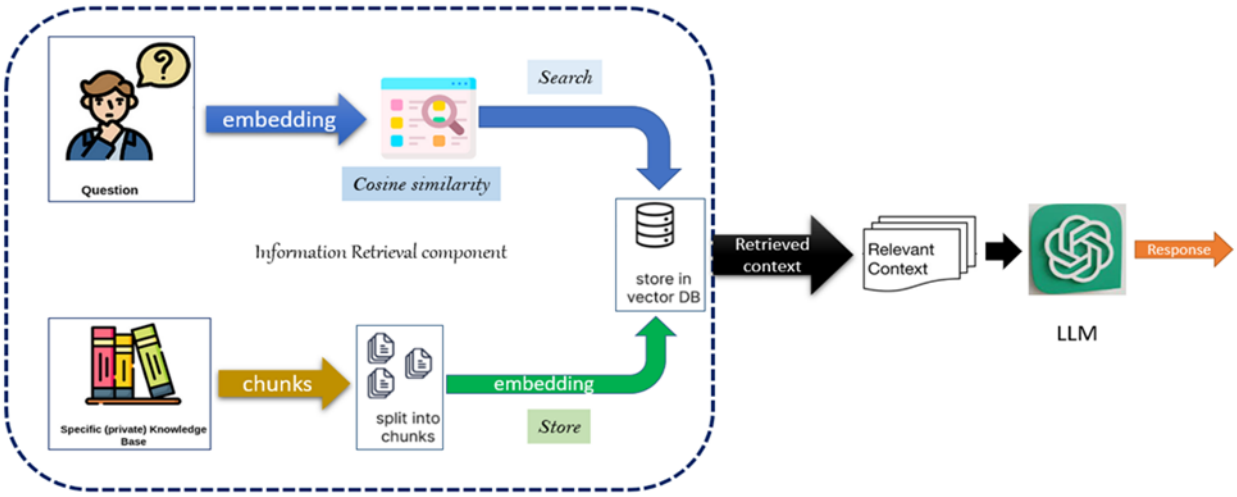


Fig. 1. Retrieval augmented generation.

II. RAG: Retrieval Augmented Generation

The Retrieval Augmented generation pipeline, shown in Fig. 1 is as follows:

A. Phase I: Information retrieval

1) Given a corpus of unstructured text containing documents
2) Given a user text query
3) A semantic embedding model is identified
4) The query is embedded into a feature vector of n dimension (semantic embedding)
5) The corpus is segmented into m text chunks (either disjoint or overlapping)
6) Each text chunk is embedded into a feature vector of n dimension using the same embedding model used in embedding the search query, as shown in Fig. 2.
7) The m- m-vectors are indexed in a Vector DB store
8) Cosine similarity, euclidean or inner product score is computed between the embedded vector of the query
9) 9. Top k relevant chunks are retrieved, which comprise a context for the next phase

B. Phase 2: LLM Comprehension and Response

In this phase, a suitable LLM is identified and selected, whether an open source model (more than 100 LLM models are currently available) , like LLaMA (7b/13b/70b), Falcon, GPT neoX, Bloom, vicuna , FlanT5 , etc.) . However, not all of them support the Arabic Language. The Current Arabic LLM models are:

- OpenAI GPT-turbo-3.5
- Open AI GPT 4.0
- Google Bard
- Microsoft Bing Chat (on top of openAIGPT3.5)
- Google PaLM2 (vertex-ai)
- Jais (UAE Arabic Language Model)

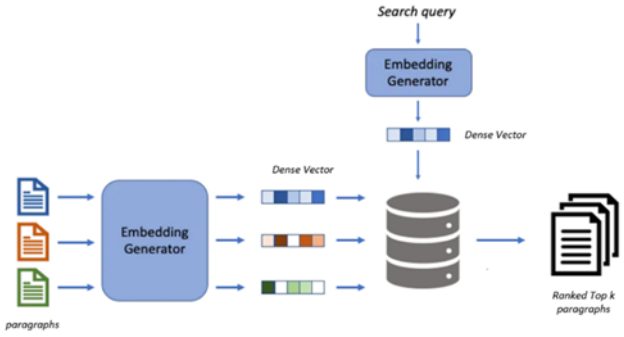


Fig. 2. Semantic embedding.

However, not all of them provide APIs for programmatic tasks, which are provided at cost like (OpenAI GPT3.5-turbo/GPT4.0 Google PalM - vertex-ai). After an Arabic LLM is identified. A prompt is constructed with two variable components: the retrieved Top k text chunks as a context and the input research query. The prompt instructs the Arabic LLM to find an answer to the search query from within the retrieved context. This architecture is widely used in Enterprises for domain-specific deployment of generative AI LLMs. A fundamental component of the overall process is the information retrieval component, as depicted in Fig. 1. The overall efficiency of the system is highly dependent on the performance of the IR component. If the IR component fails to retrieve relevant portions from the corpus, the LLM will not find the proper answer or, even worse, may hallucinate with wrong answers confidently depending on the LLM model settings (like temperature) as well as the prompt engineering. Ensuring an efficient and accurate semantic embedding is of paramount importance for an effective and practical RAG system.

III. Related Work

In the context of Arabic language processing, a lot of research was done in Natural Language Understanding, taking into consideration the different dialectical nature of Arabic

language [7], in addition to research work on Arabic text classification [8] as well as Arabic text similarity using statistical techniques [9]. However, relatively fewer research studies were conducted on semantic embeddings for Arabic text. FastText for Arabic Word Embeddings [11] proved to be an effective method for generating Arabic word embeddings. These embeddings capture subword information, making them valuable for morphologically rich languages like Arabic. Researchers have also explored Word2Vec-based approaches for Arabic word embeddings [12].

Multilingual FastText: Multilingual embeddings have gained attention for their ability to handle multiple languages simultaneously through incorporating language-specific information while sharing a common subword vocabulary across languages [4]. Researchers have also explored cross-lingual embeddings that facilitate knowledge transfer between languages. A new method is proposed that aligns word embeddings across languages, enabling multilingual applications [13].

Sentence-BERT: Sentence embeddings, which capture the semantic meaning of entire sentences, have gained popularity [10] and demonstrated superior performance in various sentence-level tasks in monolingual settings. Multilingual sentence embeddings [14] have been explored for cross-lingual applications based on training sentence embeddings for multiple languages using a shared model. The research work in [9] provided a foundational framework for evaluating critical semantic embedding APIs that play a pivotal role in search and broader information access initiatives. The author in [9] addresses the challenge of the limited accessibility of increasingly large language models by examining the utilization of semantic embedding APIs for information retrieval. Their investigation focused on assessing the capabilities of these APIs in domain generalization and multilingual retrieval using benchmark datasets like BEIR and MIRACL. The study reveals that re-ranking BM25 results using these APIs proves to be cost-effective and most effective in English contexts, offering an alternative to the conventional practice of using them as initial retrievers. For non-English retrieval, the authors suggest a hybrid model with BM25 as the most effective approach, albeit at a higher cost.

For using embeddings in downstream tasks, the authors in [17] focused on Arabic sentiment analysis, particularly on social media platforms like Twitter and Facebook, which have become vital for understanding user opinions and preferences. Sentiment analysis, however, faces challenges in natural language processing (NLP). Recent advancements in deep learning have demonstrated superior performance in NLP-related tasks compared to traditional statistical and lexical-based approaches. A comparative analysis of classic and contextualized word embeddings for sentiment analysis was conducted utilizing both trained and pre-trained versions of the four most commonly used word embedding techniques: GloVe, Word2Vec, FastText, and ARBERT. Deep learning architectures, namely, BiLSTM and CNN, are employed for sentiment classification, and experiments are conducted on benchmark datasets, including HARD, Khooli, AJGT, ArSAS, and ASTD. The results reveal that, in general, embeddings generated by one technique outperform their pre-trained counterparts, with contextualized transformer-based embedding BERT achieving the highest performance, highlighting the significance of word embeddings in Arabic sentiment analysis.

An Arabic reading comprehension dataset (ARCD) [6] addressed the challenge of open-domain Arabic question and answering (QA) with Wikipedia as the knowledge source. Mainly the scarcity of labelled QA datasets and the need for efficient Arabic machine reading comprehension and retrieval. To overcome the lack of Arabic QA datasets, they introduced the Arabic Reading Comprehension Dataset (ARCD), generated by crowd-workers from Wikipedia articles and a machine translation of the Stanford Question Answering Dataset (Arabic-SQuAD). Their open-domain QA system, SOQAL, included two components; the first is a hierarchical TF-IDF component and a neural reading comprehension component based on the pre-trained BERT transformer. Experiments on ARCD demonstrate the effectiveness of their approach, with the BERT-based reader achieving a 61.3 F1 score and SOQAL achieving a 27.6 F1 score in open-domain Arabic question answering.

In the next session, we will dive more into the semantic embedding models used in the current research.

IV. Semantic Embedding Models

Sentence and paragraph embeddings are crucial tools in information retrieval (IR), enabling systems to comprehend and retrieve text based on semantic meaning. These embeddings encode the meaning of sentences and paragraphs into fixed-size vectors [16], [18], allowing for semantic search, document retrieval, question answering, duplicate detection, clustering, summarization, recommender systems, cross-lingual search, and contextual understanding. By representing queries and documents in a continuous vector space, these embeddings enhance the accuracy of IR tasks by retrieving relevant content, even when keyword matching falls short in capturing the nuances of user intent or dealing with extensive and unstructured text collections. In the Question and Answering (QA) setting under study: Given a complete dataset of records (context-paragraphs (cps), question, ground truth answer), the IR problem is to retrieve the most relevant cps to this query. In the current work, since our focus is on the Arabic language, we explored ten embedding models that have multilingual embedding features. The query is embedded, resulting in a fixed-size feature vector, and each of the context paragraphs (cps) is also embedded. The result of embedding is a feature vector that embodies the semantic features of the text (question or context paragraph). A cosine similarity distance metric is calculated between the query, and all the semantic features of all the cps is given by

$$\text{cosine\_similarity}\,(\mathbf{A},\mathbf{B}) = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

The cosine similarity metric is a value between [0,1]; the higher score implies a higher similarity. The essence here is that "most probably the answer of the query lies in the 'context-paragraph' cp with the highest similarity with the input query". This assumption, which is mostly adopted in current QA systems, works well in the majority of situations

with much higher performance than keyword search and retrieval. The following multilingual embedding models were investigated in this research work.

A. Mpnet: Paraphrase-Multilingual-mpnet-base-v2

Mpnet [10] is based on SBERT (Sentence-BERT), which is a modification of the pre-trained BERT network that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings, allowing for efficient comparisons using cosine similarity. The original work BERT [18] and RoBERTa [19] have achieved state-of-the-art performance on sentence-pair regression tasks such as semantic textual similarity (STS). However, they require both sentences to be fed into the network, leading to significant computational overhead. Siamese networks are a type of neural network architecture that can learn to compare two inputs and measure their similarity or dissimilarity. BERT is a pre-trained language model that can encode sentences into fixed-length vectors, but it requires both sentences to be fed into the network simultaneously, which is inefficient for large-scale applications. Sentence-BERT (SBERT) is a modification of BERT that uses Siamese networks to derive sentence embeddings that can be compared using cosine similarity. This way, SBERT can compute the similarity of two sentences without processing them together, reducing the computational cost and enabling semantic similarity search and clustering. SBERT can produce more accurate and consistent embeddings than BERT, as it fine-tunes the model on specific similarity tasks.

B. Google LaBSE

While BERT has proven to be a powerful approach for acquiring monolingual sentence embeddings that excel in tasks related to semantic similarity and embedding-based transfer learning, the realm of BERT-based cross-lingual sentence embeddings was relatively uncharted. A comprehensive Language-agnostic BERT Sentence Embedding (LaBSE) [20], developed by Google researchers with training across 112 languages, including Arabic, using the Tatoeba dataset [13]. This is the multilingual SBERT model used in this research work. The embedding dimension is 768.

C. Openai Ada-embedding

Openai research team [21] delved into the significance of text embeddings, essential for tasks such as semantic search and text similarity assessment, transcending traditional applications. Unlike previous methods that tailored models for specific use cases, they introduced a more unified approach, emphasizing extensive contrastive pre-training on unsupervised data. This strategy produced top-quality vector representations with a wider context window for both text and code, a breakthrough validated across various benchmarks, including MSMARCO [15], Natural Questions, TriviaQA, and code search. Their findings underscored the versatility of these unsupervised text embeddings, demonstrating their potential to excel in state-of-the-art performance across diverse domains, from linear-probe classification to large-scale semantic search.

D. Cohere Multilingual Embedding

Cohere's multilingual text understanding model [16] works by mapping text to a semantic vector space, where texts with similar meanings are positioned close to each other. This allows for a variety of valuable use cases in multilingual settings, such as search, content aggregation and recommendation, and zero-shot cross-lingual text classification. To train the model, Cohere collected a dataset of nearly 1.4 billion question/answer pairs across tens of thousands of websites in hundreds of languages. This dataset is unique because it contains questions actually asked by speakers of said languages, allowing the model to capture language- and country-specific nuances.

E. Meta SONAR:Language-Agnostic Representations

Meta introduced SONAR, a novel fixed-size sentence embedding space with support for multiple languages and modalities [22]. SONAR's single text encoder, spanning 200 languages. Meta stipulated that SONAR outperforms existing sentence embeddings like LASER3 and LabSE in multilingual similarity search tasks. It extends its capabilities to speech segments by employing language-specific speech encoders trained in a teacher-student framework, surpassing existing speech encoders in similarity search tasks. SONAR also provides a text decoder for 200 languages, facilitating text-to-text and speech-to-text machine translation, including zero-shot language and modality combinations [22]. In our findings, we found that this is an overstatement when applied to Arabic Language, as described in section 4.

F. Microsoft E5 - (Small-base-large)

Microsoft researchers [23] presented E5, a family of advanced text embeddings designed for versatile applications across various tasks. E5 stands for EmbEddings from bidirectional Encoder representations. These embeddings are trained using a contrastive approach applied to a large, curated text pair dataset called CCPairs. E5 text embedding models are suitable for tasks like retrieval, clustering, and classification, where a single-vector representation of text is required. It exhibits robust performance in both zero-shot and fine-tuned settings. The authors extensively evaluated 56 datasets using BEIR and MTEB [5] benchmarks. In zero-shot scenarios, E5 surpasses the strong BM25 baseline in the BEIR retrieval benchmark, and when fine-tuned, it achieved the best results in the MTEB benchmark at the time of the publication (Dec. 7th. 2022), outperforming existing embedding models with significantly fewer parameters.

G. HuggingFace DistillBert v1,v2

HuggingFace researchers [24], proposed DistilBERT, a smaller and more efficient language representation model derived from BERT. As transfer learning from large-scale pre-trained models gains prominence in Natural Language Processing (NLP), the challenge lies in deploying these large models on resource-constrained devices or under tight computational budgets. DistilBERT is pre-trained using knowledge distillation techniques, reducing the model size by 40% while retaining 97% of its language understanding capabilities and achieving a 60% increase in speed. To leverage the inductive biases from larger models, they introduced a triple loss mechanism

that combines language modelling, distillation, and cosine-distance losses. DistilBERT proves to be cost-effective for pre-training and demonstrates its suitability for on-device computations through proof-of-concept experiments and comparative on-device studies. In our analysis, we explored distils-base-multilingual-cased-v1 and v2 , denoted in the experiments as hf1 and hf2

V. Experimental Results

A. ARCD Dataset

The dataset used in the benchmark analysis is ARCD (Arabic Reading Comprehension Dataset) [6]. Crowdsourced 1,395 questions with the corresponding context paragraph and ground truth answers. The research involved curation and crowdsourcing, focusing on 155 randomly selected articles from the top 1000 most viewed articles on Arabic Wikipedia in 2018. These articles spanned a wide range of topics, including religious figures, historical figures, sports celebrities, countries, and companies. To ensure the appropriateness of the content, a manual filter was applied to remove any adult material. In total, the project collected 1,395 questions based on 465 paragraphs extracted from the 155 selected articles. Fig. 3 shows a typical record from the ARCD dataset. Following the pipeline in Fig.

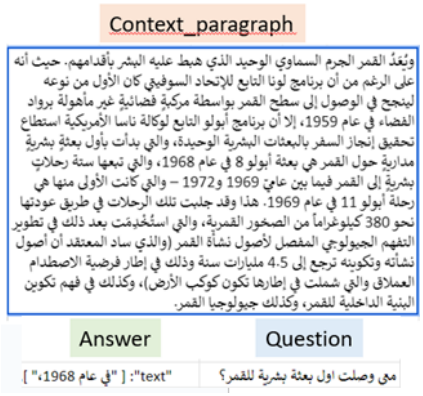


Fig. 3. ARCD record example.

1, the knowledge base (corpus) is constructed based on the concatenation of all Context paragraphs (CPs) of all 1395 questions. Each question and each context paragraph CP is embedded using one of the ten models under study. Faiss (Facebook AI Similarity Search) python library is used for Vector DB indexing and search.

B. Recall@k Performance Metric

Average Recall@k metric is used, where k is the top model retrieval hits with values [1,2,3,4,5,10,15]. The performance results are shown in Fig. 4 and Table I.

The results showed that the Microsoft E5 Family of models had a superior overall performance for all k values, where the top performer was E5 - ML - Large. the second was ada openAI embedding, and worth noting here that E5 is a free, open-source model, while Openai ada is at a cost (0.0001$/1k tokens)

TABLE I. Recall@k for Various Models

	sonar	e5s	e5b	e5l	hf1	hf2	cohere	mpnet	ada	LaBSE
$k = 1$	11%	21%	21%	23%	15%	14%	15%	14%	18%	15%
$k = 2$	23%	43%	42%	46%	27%	26%	28%	27%	36%	29%
$k = 3$	35%	63%	62%	68%	40%	38%	40%	38%	53%	44%
$k = 4$	40%	71%	69%	75%	46%	44%	44%	44%	59%	50%
$k = 5$	44%	76%	74%	79%	52%	49%	48%	49%	64%	55%
$k = 10$	56%	87%	87%	91%	66%	63%	60%	61%	77%	70%
$k = 15$	63%	90%	90%	93%	71%	67%	64%	65%	79%	78%
Embedding	1024	384	768	1024	512	512	768	768	1536	768

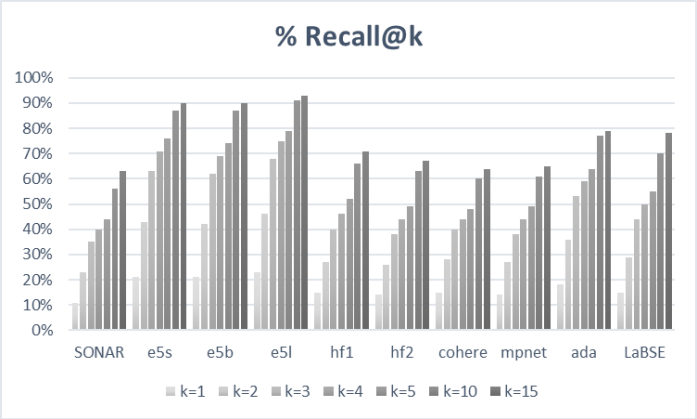


Fig. 4. Average % Recall@k performance.

C. Embedding Dimensions and Model Score

Fig 5 shows the embedding dimension of each model and the top model had 1024, while Openai ada had the maximum embedding dimension of 1536. What's interesting is that the second top performer, E5-small, has an embedding dimension of 384, which is quite impressive. Naturally, the higher the embedding dimension, the higher the capacity to capture better semantic context, which impacts storage and Latency. A simple formula is used to capture the trade-off between model retrieval accuracy and embedding dimension in an overall score:

$$model\_score = \frac{Avg.\ Recall@k\ * 1000}{Embedding\_Dimension} \tag{2}$$

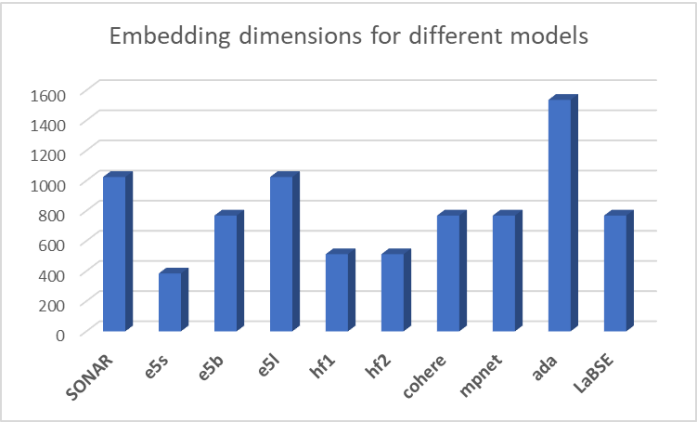


Fig. 5. Embedding dimensions.

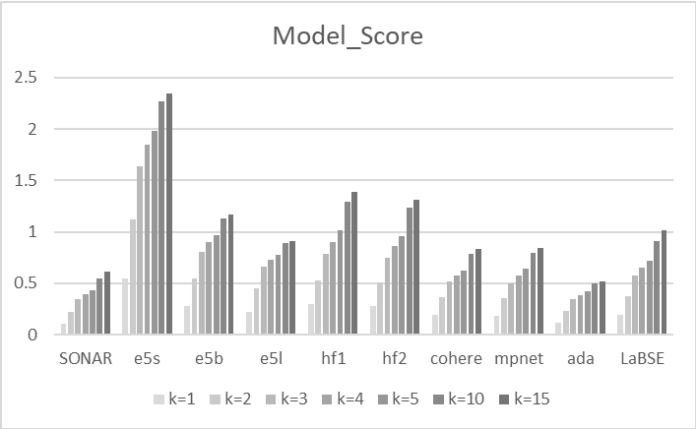


Fig. 6. Overall model_score.

The overall model_score in Fig. 6 shows a clear superior performance for the E5-ml-small model with 384 embedding dimensions, and hence highly recommended for Arabic Language Semantic Information retrieval

VI. Conclusion

In this study, we explored the application of Retrieval Augmented Generation (RAG) models in the realm of the Arabic language, an area that while linguistically rich, often receives less attention in this field. Our focus was particularly on the Information Retrieval component, with a keen eye on the processes of semantic embedding. For our evaluation, we utilized a range of advanced Multilingual Semantic embedding models, employing the ARCD dataset as a benchmark for our assessments. The knowledge corpus is generated from the concatenations of ARCD question contexts. Questions and contexts are embedded using the 10 models understudy, and Recall@k metric is used in the evaluation, where k represents the top retrieval hits based on the cosine similarity distance, and facebook AI similarity search (faiss) library.

The results indicated that the Microsoft E5 Family of models, especially E5-ML-Large (e5l), consistently outperformed the other models across different retrieval hit levels (k values). Notably, the open-source nature of E5 models makes them particularly appealing for a wide range of applications. The second-best performer was the Ada OpenAI embedding, albeit for 0.0001$/1k tokens. Furthermore, we observed that embedding dimensions play a crucial role in model performance. Higher embedding dimensions, such as the 1536 of OpenAI Ada, offer improved semantic context capture but come with storage and latency implications. To account for this tradeoff, we introduced an overall model score that combines model retrieval accuracy and embedding dimension. The E5-ML-Small model (e5s) , with an embedding dimension of 384, emerged as the top performer in this balanced evaluation. In light of these findings, we highly recommend the adoption of the E5-ML-Small model for Arabic Language Semantic Information Retrieval, as it strikes an excellent balance between retrieval accuracy and resource efficiency.

The superior performance of the e5 family of models is attributed to their unique approach to data preparation and training. Unlike conventional methods that rely on small-scale, human-annotated data or large-scale, noisy datasets, the e5 models utilize a specially curated dataset called CCPairs (Colossal Clean text Pairs), which is derived from diverse semi-structured sources.

This research contributes to the broader exploration of RAG models for Arabic language processing and information retrieval, shedding light on valuable avenues for future application of Arabic Language Understanding and Generation.

VII. Discussions and Future Work

A critical aspect warranting further investigation in Retrieval Augmented Generation (RAG) systems and semantic embeddings pertains to the dimensionality of the context window, or embedding size. While reduced embedding dimensions are advantageous for computational efficiency and data storage, they pose challenges in terms of model performance, particularly when processing extensive contexts. Such contexts often exceed the embedding dimension limits, leading to truncation which may adversely impact the model's effectiveness.

Moreover, the choice of tokenizer algorithm, inherently linked to the language being analyzed, presents another variable influencing RAG systems' performance. Tokenizer algorithms vary significantly, and their compatibility and efficiency can differ across languages. This variability underscores the necessity for extensive research into the implications of different tokenizer algorithms, especially in the context of specific languages. Such an investigation could provide valuable insights into optimizing RAG systems for diverse linguistic environments.

Future research should also encompass the exploration of contemporary methodologies in the realm of RAG systems, notably re-ranking strategies and cross-encoder architectures, from a language-specific perspective. This exploration is essential given the evolving nature of large language models and their application across various downstream tasks. In conducting such studies, it will be critical to employ nuanced, language-sensitive metrics, such as Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and normalized Discounted Cumulative Gain at k (ndcg@k). These metrics would offer a more refined evaluation of the models' capabilities in handling language-specific nuances and complexities.

Through these focused areas, we are aim to address the interplay between linguistic characteristics and the technical dimensions of RAG systems, thereby enhancing their applicability and efficiency in diverse linguistic contexts.

References

[1] Lewis, Patrick & Perez, Ethan & Piktus, Aleksandara & Petroni, Fabio & Karpukhin, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". Advances in neural information processing in systems (2020)

[2] Krishna CS. Prompt Generate Train (PGT): A framework for few-shot domain adaptation, alignment, and uncertainty calibration of a retriever augmented generation (RAG) model for domain specific open book question-answering. arXiv preprint arXiv:2307.05915. 2023 Jul 12.

[3] Yu, W.. "Retrieval-augmented Generation across Heterogeneous Knowledge. North American Chapter of the Association for Computational Linguistics (2022).

[4] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

[5] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

[6] Mozannar, H., El Hajal, K., Maamary, E., & Hajj, H. (2019). Neural Arabic Question Answering. Proceedings of the Fourth Arabic Natural Language Processing Workshop, 108–118. Florence, Italy, August 1, 2019. © 2019 Association for Computational Linguistics.

[7] Muhammad Khalifa, Hesham Hassan and Aly Fahmy, "Zero-resource Multi-dialectal Arabic Natural Language Understanding" International Journal of Advanced Computer Science and Applications(IJACSA), 12(3), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120369

[8] Alroobaea R. An Empirical Deep Learning Approach for Arabic News Classification. International Journal of Advanced Computer Science and Applications. 2023;14(6).

[9] Al-Mahmoud RH, Sharieh A. NGram Approach for Semantic Similarity on Arabic Short Text. International Journal of Advanced Computer Science and Applications. 2022;13(11).

[10] Reimers, Nils, and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. November 2019.

[11] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

[12] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing in systems, pp. 3111-3119. 2013.

[13] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

[14] S. Gouws and A. Søgaard. 2015. Simple task-specific bilingual word embeddings. In NAACL-HLT., pages 1386–1390

[15] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L. (2016). "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset." Retrieved November 2016, from https://www.microsoft.com/en-us/research/publication/ms-marco-human-generated-machine-reading-comprehension-dataset/

[16] Ehsan Kamalloo, Xinyu Zhang, Odunayo Ogundepo, Nandan Thakur, David Alfonso-hermelo, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Evaluating Embedding APIs for Information Retrieval. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track), pages 518–526, Toronto, Canada. Association for Computational Linguistics.

[17] Sabbeh, Sahar & Fasihuddin, Heba. (2023). A Comparative Analysis of Word Embedding and Deep Learning for Arabic Sentiment Classification. Electronics. 12. 1425. 10.3390/electronics12061425

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." CoRR, vol. abs/1907.11692.

[20] Mohammed Alsuhaibani. Deep Learning-based Sentence Embeddings using BERT for Textual Entailment. International Journal of Advanced Computer Science and Applications, 14(8), 2023. doi: 10.14569/IJACSA.2023.01408108. URL: http://dx.doi.org/10.14569/IJACSA.2023.01408108

[21] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, Lilian Weng. "Text and Code Embeddings by Contrastive Pre-Training." 2022. [Link](https://arxiv.org/abs/2201.10005)

[22] Duquenne, Paul-Ambroise (MetaAI & Inria), Schwenk, Holger (MetaAI), Sagot, Benoît (Inria). "SONAR: Sentence-Level Multimodal and Language-Agnostic Representations.", August 2023. Meta publisher

[23] Wang, Liang, Nan Yang, Xiaolong Huang,etal , microsoft., "Text Embeddings by Weakly-Supervised Contrastive Pre-training." 2022

[24] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR abs/1910.01108 (2019).

Elevating Android Privacy: A Blockchain-Powered Paradigm for Secure Data Management

Bang Khanh Le*[1], Ngan Thi Kim Nguyen[2], Khiem Gia Huynh[3], Phuc Trong Nguyen[4],
Anh The Nguyen[5], Khoa Dand Tran[6], Trung Hoang Tuan Phan*[7]
FPT University, Can Tho City, Viet Nam[1,3,4,5,6,7]
FPT Polytechnic, Can Tho City, Viet Nam[2]

***Abstract*—The significance of medical test records in diagnosing and treating illnesses cannot be overstated. These records serve as the foundation upon which medical professionals craft precise treatment strategies tailored to a patient's unique health condition and ailment. However, in several developing nations, such as Vietnam, a concerning trend persists: medical test records predominantly exist in vulnerable paper format, entrusted to patients for safekeeping. When patients transition between healthcare facilities, the responsibility of carrying these paper-based medical histories rests with them, introducing a significant risk factor due to the inherent fragility of paper documents, which can be easily damaged by fire or water. The loss of these crucial records can lead to severe disruptions in the diagnostic and therapeutic journey of patients, potentially compromising their well-being. Despite the emergence of various alternatives to address this vulnerability, Vietnam faces multifaceted challenges. These challenges encompass low technological literacy among patients and substantial infrastructural limitations. In response to these pressing issues, this study endeavors to harness the transformative potential of blockchain technology, smart contracts, and Non-Fungible Tokens (NFTs) to effectively mitigate the drawbacks associated with paper-based medical test records. Our comprehensive approach includes meticulous cataloging of current hospital practices, the introduction of a purpose-built blueprint for decentralized record sharing, the proposal of an innovative NFT-backed authentication model, the development of a practical proof-of-concept, and comprehensive platform testing. Through these efforts, we aim to revolutionize the management of medical test records in Vietnam, enhancing accessibility, security, and reliability for both patients and healthcare providers.**

***Keywords*—Medical test result; blockchain; smart contract; NFT; Ethereum; Fantom; Polygon; Binance smart chain**

I. Introduction

The ubiquity of Android applications (apps) in contemporary society is undeniable. These apps have seamlessly woven themselves into the fabric of daily life, offering indispensable utility across various domains. Whether it's navigation through Google Maps, social connectivity via Facebook, or health tracking with Fitbit, Android apps have become indispensable companions to billions of users. As of March 2023, the Google Play store boasts an astounding repository of over 2.6 million apps[1], a testament to the Android ecosystem's enduring popularity and expansion. The significance of Android's prevalence extends beyond mere numbers. According to App Annie's 2021 report, Android users collectively spent an astonishing 3.5 trillion hours on these apps[2]. This level of engagement underscores the integral role Android apps play in modern life. To exemplify, consider Facebook, one of the most pervasive Android apps, with an impressive 2.94 billion monthly active users. Remarkably, 1.96 billion users visit the social networking platform daily[3]. Yet, in tandem with this digital revolution comes a growing concern about privacy. A 2019 Pew Research Center report revealed that approximately 81% of the public feels they have little or no control over the data collected about them, be it by private companies or government entities[4]. This concern extends to mobile apps, as demonstrated by a 2019 survey by NortonLifeLock, which found that 72% of consumers worry about their privacy when using these apps. Concerns range from the specter of identity theft to fears of unauthorized access to personal information[5]. In response to these privacy concerns, Android introduced a permission model for managing access to sensitive data and specific actions within apps[6]. This model empowers users by requiring apps to request permissions, which can be granted or denied during installation or at runtime. App stores like the Google Play Store implement review processes and policies to ensure app compliance with privacy guidelines. This enables users to review app permissions, ratings, and user feedback, equipping them with the tools to make informed decisions regarding privacy risks[7]. Additionally, data safety policies mandate developers to be transparent about their data practices, ensuring data protection[8].

However, traditional methods for safeguarding user privacy within Android apps face significant challenges. Centralized data storage, a prevalent feature in these apps, introduces a "single point of failure" that substantially heightens the risk of data breaches and unauthorized access [1]. Moreover, these traditional methods often lack transparency, rendering it difficult for users to comprehend and control how their data is collected, processed, and shared. This lack of transparency can erode user trust and deter them from using a service due to privacy concerns [2], [3]. Additionally, these methods offer limited user control over personal data, constraining users from managing what data is collected, how it's used, or with whom it's shared [4]. Recognizing the urgency of addressing these challenges, this paper introduces a novel and forward-thinking approach to elevate Android app privacy to new

[1]http://bit.ly/3Nduhcg

[2]https://technologymagazine.com/digital-transformation/app-annie-38-trillion-hours-spent-mobiles-2021

[3]https://bit.ly/3CzGfbo

[4]pewrsr.ch/3PjVOLW

[5]https://bit.ly/447dheI

[6]https://developer.android.com/guide/topics/permissions/overview?hl=en

[7]For instance, TikTok's privacy policy can be found at: https://www.tiktok.com/legal/page/row/privacy-policy/en

[8]https://bit.ly/467Mkct

heights. Our contribution centers on a groundbreaking hybrid architecture that integrates conventional data processing with blockchain technology, tailored specifically to handle sensitive data. The importance of this contribution cannot be overstated. As the Android app ecosystem continues to expand, so do the volumes of data being generated and processed, including sensitive health and medical data. With the growing emphasis on user privacy, our hybrid approach offers a transformative paradigm that ensures advanced security, transparency, and user control. It provides a solution to the persistent privacy concerns pervasive within the Android app ecosystem.

In this paper, we delve into the intricacies of our hybrid architecture, demonstrating how it effectively manages 'dangerous' permissions, particularly those related to motion/health/medical data, which are of paramount importance due to their sensitivity. Through rigorous evaluation on Ethereum Virtual Machine (EVM) compatible platforms such as BNB, Fantom, Celo, and Matic, we identify the optimal platform, with the Fantom platform emerging as a standout choice owing to its low transaction costs and optimal gas limits. Additionally, our paper outlines comprehensive strategies for persuading service providers and Android OS producers to adopt our transformative approach. By doing so, we present a pioneering perspective on the application of blockchain technology to address the persistent privacy concerns that continue to challenge the Android app ecosystem. In the subsequent sections, we provide a detailed account of our approach, the evaluation results, and the strategies for adoption, ultimately contributing to a future where Android users can enjoy the benefits of innovative apps without compromising their privacy.

II. Related Work

In this section, we provide an extensive review of related work in the field of privacy preservation within Android platforms. While our paper introduces a novel approach leveraging blockchain technology for this purpose, it is essential to contextualize our contribution within the broader landscape of existing research.

A. Malware Analysis

Privacy preservation within Android platforms has long been a concern, prompting researchers to explore various methodologies to address these challenges. One prominent avenue of research in this area is malware analysis, which focuses on identifying and mitigating malicious software that may compromise user privacy.

Talha et al. [5] developed APK Auditor, a system designed for detecting malicious apps through permission analysis. This multifaceted system encompassed Android clients, a signature database, and a central server. APK Auditor employed static analysis to capture permission requests and calculate malicious scores, contributing to the early efforts of enhancing Android app security.

In a similar vein, Jianmao et al. [6] introduced MPDroid, an innovative approach that evaluated the risk of target apps based on minimum permissions. The methodology incorporated collaborative filtering and clustering techniques to assess app risk, providing valuable insights into permission-based privacy risk estimation.

Enck et al. [7] proposed TaintDroid, an Android platform extension that significantly advanced privacy preservation. This pioneering system tracked the flow of privacy-sensitive data within third-party apps and promptly labeled apps as privacy violations when personal data was transmitted to third parties. TaintDroid marked a critical milestone in privacy-aware app development and user protection.

Furthermore, Moutaz et al. [8] focused on a set of dangerous permissions identified by Google[9], contributing to the categorization and understanding of potential privacy risks within Android apps.

While these studies made commendable strides in enhancing privacy preservation within Android platforms, they often provided binary detection outcomes (malicious or benign). This binary nature of their analysis might not be sufficient to assist users effectively in their decision-making processes.

To address this limitation, Son et al. [4], [9] proposed an innovative approach centered on risk estimation based on an app's data collection and sharing behavior. Their approach marked a significant shift towards a more nuanced understanding of app privacy risks, enabling users to make informed choices based on the level of risk they were willing to accept. Additionally, their app recommendation system [10] further personalized user app choices based on their privacy preferences, contributing to user-centric privacy preservation.

However, even these advanced risk estimation approaches do not fully address the need for a robust, transparent, and user-controlled platform for managing sensitive data within Android apps. In contrast, our paper introduces a novel paradigm by integrating blockchain technology into Android platforms. This blockchain-based system offers enhanced security, transparency, and user control for handling sensitive data, thereby providing a balanced, secure, and effective method for managing data in Android apps. By retaining traditional processing for 'normal' data, our approach preserves app functionality and data processing effectiveness. Our contribution represents a significant advancement in the ongoing efforts to protect user privacy on Android platforms.

B. Sensitive Data Encryption

Another prevalent approach to privacy preservation within Android platforms is sensitive data encryption. This methodology involves encoding user data in a way that allows only authorized parties to access it, thus ensuring the confidentiality and integrity of sensitive information.

Chen et al. [11] proposed AUSERA, an automated tool designed to detect security vulnerabilities in Android apps. While AUSERA did not specifically focus on encryption, its contribution lay in the identification of security vulnerabilities, which are often associated with privacy breaches.

Zhang et al. [12] examined the vulnerability of Android external storage, an area susceptible to data leakage, and proposed solutions to prevent sensitive information disclosure.

[9]https://developer.android.com/reference/android/Manifest.permission

Their work underscored the importance of protecting sensitive data, especially in scenarios where external storage is involved.

Fan et al. [13] introduced HPDROID, an automated system that identified GDPR compliance violations in mobile health applications. While their primary focus was on compliance, it indirectly contributed to privacy preservation by highlighting the critical need for robust data protection mechanisms within health-related apps.

Mia et al. [14] conducted a comprehensive comparative study on HIPAA technical safeguards assessment of Android mHealth applications. This study provided valuable insights into healthcare data privacy, emphasizing the need for stringent security measures within healthcare apps.

Haupert et al. [15] evaluated the state of Android app hardening and identified vulnerabilities in a leading Runtime Application Self-Protection (RASP) product. Their work demonstrated the need for continuous improvement in security mechanisms to protect sensitive data from emerging threats.

Chen et al. [16] and Sengupta et al. [3] proposed blockchain-based systems specifically designed for preserving medical data privacy within Android platforms. These pioneering contributions recognized the unique challenges associated with healthcare data and demonstrated the potential of blockchain technology in safeguarding sensitive medical information.

Balasubramanium et al. [17] conducted a comprehensive survey on data privacy and preservation using blockchain in healthcare organizations. This survey illuminated emerging trends and challenges in healthcare data privacy, underscoring the importance of innovative solutions like blockchain.

While these studies have made valuable contributions to privacy preservation within Android platforms, they often rely on developers' security knowledge and diligence. Moreover, some studies focus primarily on healthcare data and do not address the broader range of sensitive data handled by Android apps. Others lack a specific implementation tailored for Android platforms.

In contrast, our proposed blockchain-based system offers a robust, transparent, and user-controlled platform for managing sensitive data in Android apps, applicable to a wide array of data types beyond healthcare. By integrating blockchain technology, we address the limitations of existing approaches and contribute significantly to the ongoing endeavor to preserve user privacy on Android platforms.

III. Background

A. Android OS and Permission System

Android, an open-source operating system developed by Google, is ubiquitous in modern society, powering a wide array of mobile devices such as smartphones, tablets, and wearables. Built upon the Linux kernel, Android offers a robust application framework, enabling developers to create innovative apps and games within a Java-based environment. This extensive ecosystem boasts over 2.6 million apps available on the Google Play Store as of March 2023, highlighting its profound impact on our digital lives. Central to Android's design is its intricate permission system, meticulously crafted to safeguard user privacy. This system plays a pivotal role in dictating how apps interact with sensitive user data and device resources. When developers create an Android app, they must explicitly define in the app's manifest file what types of permissions the app requires to function effectively. These permissions govern a diverse range of access, from location data to contacts, camera, microphone, and beyond.

Permissions in the Android ecosystem are categorized into different levels, with particular emphasis on two primary types: Normal and Dangerous permissions. The distinction between these permission categories is crucial and forms the cornerstone of user privacy and security:

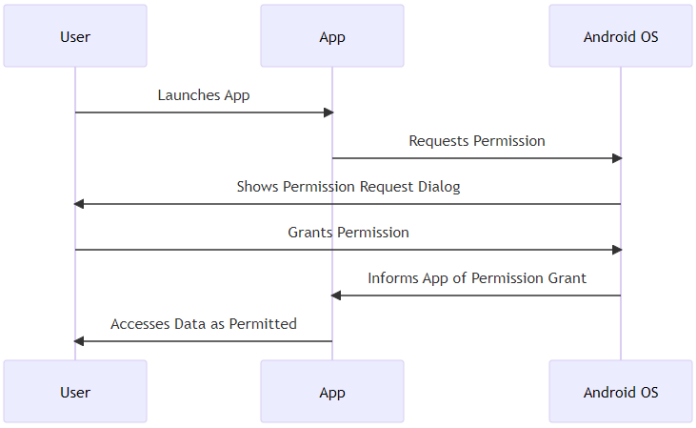


Fig. 1. The user data collection via the permission system in android OS.

- Normal Permissions: These permissions encompass scenarios where apps need to access data or resources outside their own sandboxed environment, with minimal risk to user privacy or the functioning of other apps. An example of a normal permission is the ability to set the device's time zone. Such permissions are typically granted without explicit user consent at installation.
- Dangerous Permissions: In contrast, dangerous permissions involve access to highly sensitive user data or device resources, potentially affecting user privacy and the integrity of other apps. Permissions like accessing the user's contacts, camera, or location fall into this category. Android mandates explicit user consent for granting dangerous permissions, often soliciting consent dynamically as the app requires access during runtime, rather than at installation.

The Android permission system operates as a crucial pillar in maintaining a balance between app functionality and user privacy. It empowers users to make informed decisions regarding which permissions to grant to individual apps, ensuring their data remains secure and their privacy respected.

Fig. 1 illustrates the intricate workflow involved in an app's data collection via the Android permission system. This process embodies the essence of user-centric privacy preservation within the Android ecosystem.

In summary, Android's widespread adoption and sophisticated permission system have fostered an environment where

user data privacy is paramount. However, challenges remain in ensuring transparent, secure, and user-controlled data management, especially concerning sensitive information.

B. Blockchain Technology

Blockchain technology has emerged as a disruptive force, initially popularized through its association with digital currencies like Bitcoin. Its inherent properties have extended far beyond the realm of cryptocurrencies, finding applications across diverse industries. At its core, a blockchain consists of a chain of blocks, each containing a list of transactions. These blocks are cryptographically linked, with each new block referencing the previous one, creating an immutable ledger. Immutability ensures that once data is recorded on the blockchain, it cannot be retroactively altered, providing unparalleled data integrity. One of the foundational characteristics of blockchain technology is its decentralization. Unlike traditional centralized databases, where a single entity controls access and data, a blockchain operates on a network of distributed nodes. Each node holds a complete copy of the blockchain and participates in validating and recording new transactions. This decentralized architecture eliminates single points of failure and dramatically enhances security. Transparency and auditability are fundamental attributes of blockchain technology. All transactions on the blockchain are visible to every participant in the network. Moreover, every transaction is permanently recorded, creating an unchangeable audit trail. This transparency builds trust and accountability, assuring participants that no transaction can be tampered with or erased.

Smart contracts, another hallmark feature of blockchain, are self-executing agreements with contract terms directly encoded into code. These contracts automatically execute transactions when predefined conditions are met, removing the need for intermediaries. In the context of Android privacy preservation, blockchain technology offers a range of benefits:

- Enhanced Data Security: Blockchain replaces centralized data storage with a decentralized, tamper-resistant mechanism. Advanced cryptographic algorithms and distributed consensus protocols secure user data against unauthorized access, manipulation, or breaches. This significantly mitigates the risk associated with centralized data storage and bolsters privacy preservation in Android systems.
- Improved Transparency and Auditing: Traditional methods for managing user data in Android apps often suffer from a lack of transparency, causing trust issues among users. Blockchain's transparency and immutability address this problem. Every transaction is transparent and can be audited, ensuring app developers and service providers adhere to declared privacy policies and data handling practices. Smart contracts provide an auditable trail of all data interactions, increasing transparency and fostering user trust.
- User Control and Consent: One significant drawback of conventional methods is the limited control users have over their personal data. Blockchain technology significantly enhances user control over personal data. With self-sovereign identity solutions built on blockchain, users can manage their own identity data, decide what information to share, and with whom, and for what purpose. These features offer users the ability to selectively consent to data collection and sharing, substantially preserving their privacy.
- Data Integrity and Provenance: The immutable nature of blockchain ensures the integrity and provenance of data. Each data transaction is recorded permanently and cannot be altered, providing a verifiable and auditable history of data transactions. This feature allows users to verify the authenticity and accuracy of their data, enhancing the trustworthiness of Android apps, particularly in scenarios where data traceability and accountability are paramount.
- Secure Data Sharing: Blockchain technology can also facilitate secure and controlled data sharing within the Android ecosystem. Users can grant or revoke access to their data at any time with blockchain-based consent management systems. This flexibility ensures that data is only shared with authorized parties and only for approved purposes, further enhancing privacy preservation in Android applications.
- Trust and Collaboration: The distributed and transparent nature of blockchain technology inherently fosters trust and collaboration among participants. By providing a shared and immutable record of data transactions, blockchain can help establish trust between users, app developers, and service providers in the Android ecosystem. This enhanced trust can encourage responsible data handling practices, promote fair data exchanges, and motivate collaborative efforts in privacy preservation, leading to a more secure and user-centric Android platform.

In summary, the intersection of Android OS and blockchain technology holds great promise for addressing the challenges associated with privacy preservation in the Android app ecosystem. The combination of Android's extensive user base and the robust security and transparency offered by blockchain creates a compelling framework for reimagining how sensitive data is managed and protected in mobile applications.

IV. Approach

Our proposed approach, depicted in Fig. 2, represents a comprehensive integration of blockchain technology with the permission management process in Android systems. This integration is meticulously designed to bolster privacy preservation and provide users with heightened control over their sensitive data.

To elucidate the intricacies of our approach, we break it down into a series of steps, each serving a crucial role in ensuring robust privacy protection and user empowerment:

Step 1: User Application Installation The journey begins when a user decides to install an application, with a particular focus on medical applications. Medical apps are chosen for their inherent sensitivity, as even small fragments of medical data hold the potential for privacy breaches, potentially compromising individual identities.

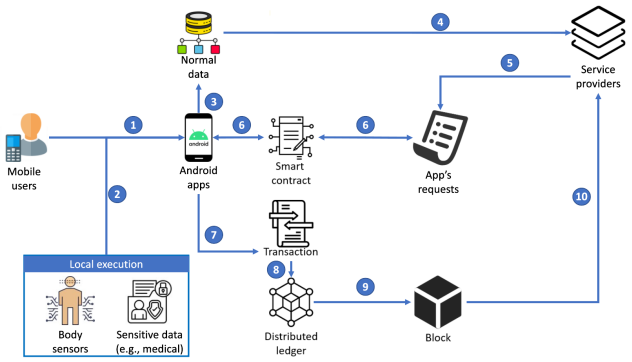


Fig. 2. Architecture of our proposed privacy-preserving model.

Step 2: Dangerous Permission Request When a medical application requests access to sensitive medical data through Android's dangerous permissions system, it potentially gains access to the user's comprehensive or aggregated health information. This marks the second step in our process.

Step 3: Normal Data Request In cases where the application requests access to non-sensitive, normal data, it follows the conventional permissions procedure outlined in Section III. Such data can be collected without the need for additional security measures.

Step 4: Secure Data Transmission Should the service provider necessitate the transmission of the collected normal data to their servers (referred to as "global execution"), the fourth step of our approach comes into play. Our system ensures the secure and encrypted transmission of this data to the designated service provider's servers.

Step 5: Sensitive Data Request Alternatively, if an application seeks sensitive data, such as those accessed via dangerous permissions like `BODY_SENSORS` and `BODY_SENSORS_BACKGROUND` for medical data collection, it is mandated to provide detailed information regarding the purpose of usage, the specific data types required, and the possibility of third-party involvement. This meticulous process is justified by (i) the availability of most of this information in the application's privacy policy and (ii) compliance with data privacy regulations, including the General Data Protection Regulation (GDPR)[10], which mandates secure data handling processes by applications.

Step 6: User Validation via Smart Contract The user, on their mobile device, actively participates in the decision-making process concerning their data in the sixth step. This active involvement is facilitated through the use of smart contracts, which enhance transparency and decentralize control, placing it squarely in the hands of the user.

Step 7: Transaction Logging Upon user validation, the decision is meticulously logged as a transaction, creating a formal agreement between the user and the service providers. This transactional logging process harkens back to the conventional permission system. However, it takes a significant leap forward by incorporating machine learning techniques, which enable future decisions to reference these initial agreements. This reduces the need for recurrent user interventions [18].

[10]https://gdpr-info.eu/

Step 8: Secure Storage in Distributed Ledger All these transactions, replete with user decisions and service provider agreements, are securely stored within a distributed ledger (DLT) in the eighth step. This distributed ledger serves as an immutable repository of records, rendering every transaction traceable and tamper-resistant.

Step 9: Encryption and Non-Fungible Tokens (NFTs) The ninth step involves the encryption of user responses using the service provider's public key, employing elliptic curve cryptography [19]. This ensures that only the intended service provider can decrypt and access the user's responses. To further bolster data security and uniqueness, we utilize Non-Fungible Token (NFT) technology to encapsulate these responses. Each NFT is designed to be distinct and non-interchangeable, safeguarding the integrity and confidentiality of user data.

Step 10: Data Transfer via NFTs The final step in our approach is the seamless transfer of these NFT-encapsulated user responses to the service provider. This culminating interaction furnishes the service provider with the requisite and authorized data, while concurrently creating a traceable, indelible record of the transaction on the blockchain.

Our architectural approach presents numerous advantages over the traditional Android permission system:

- Enhanced Transparency and Auditability: The integration of blockchain technology ensures that every transaction is meticulously recorded and verifiable, offering a heightened level of transparency and auditability.
- User Empowerment and Control: The use of smart contracts places the reins of control firmly in the hands of the user. Users can distinctly specify which data can be accessed, fostering greater empowerment and autonomy.
- Data Integrity and Secure Sharing: By leveraging NFTs and elliptic curve cryptography, our approach guarantees data integrity and secure data sharing. This ensures that sensitive data remains confidential and unaltered.
- Trust Building and Regulatory Compliance: Our innovative approach facilitates trust-building between users and service providers. Furthermore, it promotes collaborative data sharing while adhering to stringent data privacy regulations.
- Scalability and Efficiency: The inherent scalability and efficiency of blockchain technology position our approach as a sustainable choice for future privacy-preserving systems, capable of meeting evolving demands and challenges.

In conclusion, our comprehensive integration of blockchain technology with the Android permission process represents a significant leap forward in privacy preservation and user-centric control over sensitive data. By embedding transparency, security, and efficiency into the core of our approach, we aim to redefine data privacy in the digital age.

V. Experiments

Our research endeavors to introduce a blockchain-based platform that operates as a distributed ledger, with a primary

focus on the intricacies of steps 8 and 9, as illustrated in Fig. 2. This phase marks a pivotal step toward our overarching research objective.

A. Methodology

In the development of our platform, we meticulously considered various blockchain platforms that offer support for the Ethereum Virtual Machine (EVM). The selected platforms, including Binance Smart Chain (BNB Smart Chain)[11]; Polygon[12]; Fantom[13]; and Celo[14], were chosen over open-source alternatives such as those within the Hyperledger ecosystem (e.g., Hyperledger Fabric) due to their compatibility with EVM and their widespread adoption within the decentralized application (DApp) development community. As a critical facet of our research, we implemented our approach across four distinct blockchain platforms that support the Ethereum Virtual Machine (EVM): BNB Smart Chain (BNB), Polygon (MATIC), Fantom (FTM), and Celo (CELO). An essential contribution of this study is the comprehensive collection and analysis of transaction fees associated with these platforms, utilizing their respective testnet coins. All implementations have been made publicly accessible, underscoring our commitment to contributing to the broader blockchain community. Within our evaluation, we paid meticulous attention to transaction fees and gas limits, both of which have profound implications for the operational costs and efficiency of deploying DApps on blockchain networks. Transaction fees represent the costs associated with processing a transaction and fluctuate depending on factors such as transaction complexity and network congestion. Gas limits, on the other hand, dictate the maximum amount of computational resources (gas) a user is willing to expend on a transaction, safeguarding against inadvertent overspending.

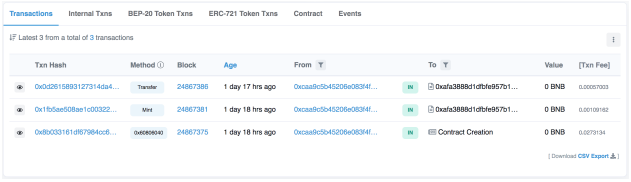


Fig. 3. Transaction information (e.g., BNB Smart Chain).

Our evaluation involved the assessment of three fundamental functions on each of these platforms:

1) **Creating User Responses:** This pivotal function embodies the user's consent level, signifying their stance on data collection, whether it involves denying data access, permitting partial access, or granting full access to medical data, as expounded in Section IV.
2) **Creating Non-Fungible Tokens (NFTs):** The user's response is encrypted using the service provider's public key and subsequently encapsulated into an NFT. This process ensures both the integrity and confidentiality of user responses.
3) **Transferring NFTs:** Subsequently, the NFT is transferred to the service provider, solidifying an immutable record of the user's response.

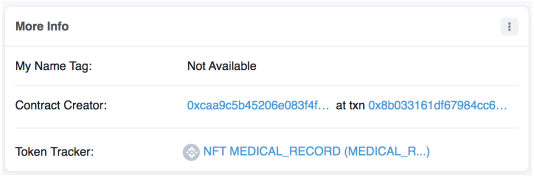


Fig. 4. NFT creation

While encryption plays a pivotal role in our approach, we have intentionally refrained from delving into the intricacies of encryption methodologies within this paper due to the substantial scope involved. A more exhaustive exploration of encryption methodologies and their analysis will be presented in forthcoming iterations of this research. Fig. 3 offers a glimpse into one of our evaluations, specifically detailing a successful deployment on the BNB Smart Chain. Analogous procedures and assessments were meticulously carried out across the remaining three platforms. Smart contracts, developed using the Solidity programming language, constituted the core of our evaluations. We scrutinized the execution costs of these contracts within the testnet environments of the respective platforms, with a primary objective of ascertaining the most cost-efficient platform for deploying our system. We delved into the intricacies of contract creation, NFT generation (as depicted in Fig. 4), and NFT ownership address updates, a process fundamentally revolving around retrieving and transferring NFTs (refer to Fig. 5).

B. Transaction Fee Analysis

In our investigation, we conducted an in-depth analysis of transaction fees incurred during the operations of Contract Creation, NFT Creation, and NFT Transfer across four prominent blockchain platforms: BNB Smart Chain, Fantom, Polygon (MATIC), and Celo. The results of this comparative analysis are summarized in Table I. On the BNB Smart Chain, the transaction fees for Contract Creation, NFT Creation, and NFT Transfer amounted to 0.0273134 BNB ($8.43), 0.00109162 BNB ($0.34), and 0.00057003 BNB ($0.18), respectively. The Fantom platform exhibited lower transaction costs, with Contract Creation priced at 0.00957754 FTM ($0.001849), NFT Creation at 0.000405167 FTM ($0.000078), and NFT Transfer at 0.0002380105 FTM ($0.000046). The Polygon (MATIC) platform showcased even more economical costs, with Contract Creation incurring 0.006840710032835408 MATIC ($0.01), NFT Creation incurring 0.000289405001852192 MATIC (almost negligible in USD terms), and NFT Transfer incurring 0.000170007501088048 MATIC (also nearly negligible in USD terms). Finally, the Celo platform reported fees of 0.007097844 CELO ($0.004) for Contract Creation, 0.0002840812 CELO (almost negligible in USD terms) for NFT Creation, and 0.0001554878 CELO (also nearly negligible in USD terms) for NFT Transfer.

This comprehensive analysis highlights the substantial variability in transaction fees across the four platforms, underscor-

[11]https://github.com/bnb-chain/whitepaper/blob/master/WHITEPAPER.md
[12]https://polygon.technology/lightpaper-polygon.pdf
[13]https://whitepaper.io/document/438/fantom-whitepaper
[14]https://celo.org/papers/whitepaper

TABLE I. TRANSACTION FEES (IN NATIVE TOKENS AND USD EQUIVALENTS)

	Contract Creation	Create NFT	Transfer NFT
BNB Smart Chain	0.0273134 BNB ($8.43)	0.00109162 BNB ($0.34)	0.00057003 BNB ($0.18)
Fantom	0.00957754 FTM ($0.001849)	0.000405167 FTM ($0.000078)	0.0002380105 FTM ($0.000046)
Polygon	0.006840710032835408 MATIC ($0.01)	0.0002894050018521 92 MATIC (almost negligible)	0.000170007501088048 MATIC (almost negligible)
Celo	0.007097844 CELO ($0.004)	0.0002840812 CELO (almost negligible)	0.0001554878 CELO (almost negligible)

ing the critical importance of platform selection in optimizing the cost-effectiveness of deploying our approach.

C. Gas Limit Assessment

TABLE II. GAS LIMITS FOR OPERATIONS

	Contract Creation	Create NFT	Transfer NFT
BNB Smart Chain	2,731,340	109,162	72,003
Fantom	2,736,440	115,762	72,803
Polygon	2,736,284	115,762	72,803
Celo	3,548,922	142,040	85,673

In addition to transaction fees, gas limits represent a vital facet when deploying smart contracts on Ethereum-based platforms. Gas limits delineate the upper bound of computational resources (gas) allocated to a transaction, serving as a safeguard against inadvertent overspending. Table II summarizes the gas limits associated with various operations, including Contract Creation, NFT Creation, and NFT Transfer, across the four blockchain platforms under consideration: BNB Smart Chain, Fantom, Polygon (MATIC), and Celo.

On the BNB Smart Chain, the gas limits for Contract Creation, NFT Creation, and NFT Transfer stand at 2,731,340, 109,162, and 72,003, respectively.

In the Fantom platform, the corresponding gas limits are marginally higher, with Contract Creation at 2,736,440, NFT Creation at 115,762, and NFT Transfer at 72,803.

Polygon (MATIC) records similar gas limits, with Contract Creation at 2,736,284, NFT Creation at 115,762, and NFT Transfer at 72,803.

Lastly, the Celo platform exhibits the highest gas limits among the four platforms. Contract Creation incurs a gas limit of 3,548,922, NFT Creation requires 142,040 gas, and NFT Transfer necessitates 85,673 gas.

This detailed analysis underscores the substantial variations in gas limits across the four platforms, further emphasizing the need for meticulous consideration in selecting the most suitable platform for deploying our innovative approach.

VI. DISCUSSION

The paradigm of secure data management and privacy preservation on Android devices has been a persistent challenge in the digital age. Our proposed approach, which integrates blockchain technology into the Android permission process, represents a transformative shift in how we address this challenge. In this discussion, we delve deep into the implications and significance of our approach, considering its potential to elevate Android privacy to new heights.

A. Blockchain as the Guardian of Privacy

The integration of blockchain technology into Android's data permission management brings about a fundamental transformation. One of the core advantages is the transparency and immutability inherent to blockchain. Every transaction, every access request, and every decision made by users are recorded in an unalterable ledger. This ledger serves as a comprehensive audit trail, granting users unprecedented visibility into how their data is accessed and used. The user-centric transparency aligns with the overarching theme of elevating Android privacy, allowing users to exercise greater control over their sensitive data. Moreover, blockchain introduces the concept of decentralized control. Through smart contracts, users actively participate in granting or denying access to their data. This not only empowers users but also shifts the locus of control from centralized authorities to individual users. The notion of users as stewards of their data is a critical step toward achieving robust data privacy.

B. User-Centric Data Control

Central to our approach is the pivotal role of smart contracts in the decision-making process. Users validate data access requests through these contracts, granting consent or withholding it. This mechanism puts the power back in the hands of users, enabling them to specify precisely which data can be accessed and under what circumstances. This level of fine-grained control is a substantial departure from the traditional Android permission system, which tends to be binary and all-encompassing. Our approach aligns perfectly with the ethos of user-centric data control. Users are no longer passive participants; they are active decision-makers in the data access process. This newfound agency fosters trust between users and service providers, as users have confidence that their data is used according to their wishes. It also complies with emerging data privacy regulations, such as the General Data Protection Regulation (GDPR), which emphasize user consent and control over personal data.

C. Security and Integrity via NFTs and Cryptography

Data security is a paramount concern in Android privacy, especially when dealing with sensitive information like medical data. Our approach addresses this concern through the use of Non-Fungible Tokens (NFTs) and elliptic curve cryptography. NFTs encapsulate user responses in a unique and non-interchangeable manner. This uniqueness guarantees that user responses remain distinct and unforgeable. It bolsters data integrity and confidentiality, assuring users that their data is protected from tampering or unauthorized access. Elliptic curve cryptography further fortifies data security. By encrypting user responses with the service provider's public key, we ensure that only the designated service provider can decrypt and access the data. This cryptographic layer adds an additional barrier to unauthorized data access.

Fig. 5. NFT transfer.

D. Trust Building and Regulatory Compliance

The elevation of Android privacy through blockchain-powered data management also contributes to trust-building between users and service providers. Users are more likely to engage with applications and services when they are confident that their data is handled with care and transparency. This trust-building aspect is essential for the sustained growth of the Android ecosystem. Additionally, our approach aligns seamlessly with data privacy regulations like GDPR. Compliance with these regulations is not just a legal requirement but also an ethical obligation. By facilitating user consent, control, and transparency, our approach inherently complies with these stringent privacy regulations. This alignment ensures that service providers can operate with confidence in a regulatory landscape that increasingly prioritizes data privacy.

E. Scalability and Efficiency

The scalability and efficiency of blockchain technology make our approach a viable and sustainable choice for the future. As Android ecosystems evolve and expand, the need for scalable privacy solutions becomes increasingly pronounced. Blockchain's inherent capacity to handle a growing volume of transactions positions our approach as a robust solution for the long term. Moreover, the efficiency of our approach is underscored by the use of various blockchain platforms, such as Binance Smart Chain, Fantom, Polygon (MATIC), and Celo. The comparative analysis of transaction fees and gas limits across these platforms offers insights into cost-effective deployment options. The adaptability of our approach to multiple blockchain environments ensures that it can seamlessly integrate into a diverse range of Android applications and services.

In conclusion, our blockchain-powered paradigm for secure data management on Android devices represents a groundbreaking approach to elevate Android privacy. By introducing transparency, user-centric control, data security, trust-building, and scalability, we lay the foundation for a new era of Android privacy that aligns with evolving user expectations and regulatory requirements.

VII. Conclusion

In this paper, we have introduced a groundbreaking approach aimed at enhancing privacy and data security for Android applications, with a specific focus on the sensitive domain of medical apps. Leveraging the capabilities of blockchain technology, our novel approach has been designed to fundamentally transform the way privacy preservation is addressed within Android platforms. Our comprehensive architecture ensures robust controls on permissions, enforces transparency in data transactions, and places users at the forefront of data management, thereby significantly elevating Android privacy standards. The central premise of our approach is the integration of blockchain technology into the Android ecosystem, revolutionizing how user data is accessed, utilized, and protected. We have presented a detailed architectural framework that harnesses the power of blockchain to regulate access to sensitive user data. This framework extends beyond the conventional Android permission system, offering a multi-step process that heightens the security and transparency of data interactions.

Through our rigorous evaluation on four Ethereum Virtual Machine (EVM)-supported platforms, namely Binance Smart Chain (BNB), Polygon (MATIC), Fantom (FTM), and Celo (CELO), we have demonstrated the feasibility and effectiveness of our approach. Notably, our evaluation revealed that the Fantom platform emerged as the most suitable option for our work. The low transaction costs and optimal gas limit settings make it an attractive choice for implementing our privacy-preserving framework. However, we acknowledge that the cryptocurrency market is subject to fluctuations, and these results may evolve over time. In our discussion, we have recognized that the successful deployment of our approach hinges on the acceptance and adoption by key stakeholders, including service providers and Android OS producers. We have outlined strategies to garner support from these critical actors, underlining the value of our approach in achieving enhanced privacy and data security.

Looking ahead, our future work will delve into the measurement and enhancement of encryption methodologies to further fortify data privacy and mitigate potential risks. As technology continues to advance, our research makes a substantial contribution to the ongoing discourse surrounding user privacy and data security within the Android ecosystem. We firmly believe that our work lays the essential foundation for the development of more transparent, secure, and user-friendly Android applications. By aligning with the evolving expectations of users and complying with ever-stricter data privacy regulations, our approach is poised to lead the way towards a future where Android privacy reaches new heights.

Acknowledgment

Our sincere appreciation is extended to Engineer Le Thanh Tuan and Dr. Ha Xuan Son, whose expertise and guidance have been indispensable in the brainstorming, execution, and assessment phases of this project. Additionally, the continuous support from FPT University Cantho Campus, Vietnam, has been instrumental in the fruition of this work.

References

[1] E. Bandara *et al.*, "A blockchain and self-sovereign identity empowered digital identity platform," in *2021 International Conference on*

Computer Communications and Networks (ICCCN). IEEE, 2021, pp. 1–7.

[2] S. Al-Natour *et al.*, "An empirical investigation of the antecedents and consequences of privacy uncertainty in the context of mobile apps," *Information Systems Research*, vol. 31, no. 4, pp. 1037–1063, 2020.

[3] A. Sengupta *et al.*, "User control of personal mhealth data using a mobile blockchain app: design science perspective," *JMIR mHealth and uHealth*, vol. 10, no. 1, 2022.

[4] H. X. Son, B. Carminati, and E. Ferrari, "A risk assessment mechanism for android apps," in *2021 IEEE International Conference on Smart Internet of Things (SmartIoT)*. IEEE, 2021, pp. 237–244.

[5] K. A. Talha *et al.*, "Apk auditor: Permission-based android malware detection system," *Digital Investigation*, vol. 13, pp. 1–14, 2015.

[6] J. Xiao *et al.*, "An android application risk evaluation framework based on minimum permission set identification," *Journal of Systems and Software*, vol. 163, p. 110533, 2020.

[7] W. Enck *et al.*, "Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones," *ACM Transactions on Computer Systems (TOCS)*, vol. 32, no. 2, pp. 1–29, 2014.

[8] M. Alazab *et al.*, "Intelligent mobile malware detection using permission requests and api calls," *Future Generation Computer Systems*, vol. 107, pp. 509–521, 2020.

[9] H. X. Son, B. Carminati, and E. Ferrari, "A risk estimation mechanism for android apps based on hybrid analysis," *Data Science and Engineering*, vol. 7, no. 3, pp. 242–252, 2022.

[10] ——, "Priapp-install: Learning user privacy preferences on mobile apps' installation," in *Information Security Practice and Experience: 17th International Conference*. Springer, 2022, pp. 306–323.

[11] S. Chen *et al.*, "Ausera: Automated security vulnerability detection for android apps," in *37th IEEE/ACM International Conference on Automated Software Engineering*, 2022, pp. 1–5.

[12] H. Zhang *et al.*, "Protecting data in android external data storage," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 1. IEEE, 2019, pp. 924–925.

[13] M. Fan *et al.*, "An empirical evaluation of gdpr compliance violations in android mhealth apps," in *2020 IEEE 31st international symposium on software reliability engineering (ISSRE)*. IEEE, 2020, pp. 253–264.

[14] M. R. Mia *et al.*, "A comparative study on hipaa technical safeguards assessment of android mhealth applications," *Smart Health*, vol. 26, p. 100349, 2022.

[15] V. Haupert *et al.*, "Honey, i shrunk your app security: The state of android app hardening," in *Detection of Intrusions and Malware, and Vulnerability Assessment: 15th International Conference*. Springer, 2018, pp. 69–91.

[16] Z. Chen *et al.*, "A blockchain-based preserving and sharing system for medical data privacy," *Future Generation Computer Systems*, vol. 124, pp. 338–350, 2021.

[17] S. Balasubramanium *et al.*, "A survey on data privacy and preservation using blockchain in healthcare organization," in *International Conference on Advance Computing and Innovative Technologies in Engineering*. IEEE, 2021, pp. 956–962.

[18] H. X. Son *et al.*, "In2p-med: Toward the individual privacy preferences identity in the medical web apps," in *International Conference on Web Engineering*. Springer, 2023, pp. 126–140.

[19] D. Hankerson *et al.*, *Guide to elliptic curve cryptography*. Springer Science & Business Media, 2006.

A Deep Transfer Learning Approach for Accurate Dragon Fruit Ripeness Classification and Visual Explanation using Grad-CAM

Hoang-Tu Vo, Nhon Nguyen Thien, Kheo Chau Mui
Software Engineering Department
FPT University, Cantho city, Vietnam

***Abstract*—Dragon fruit, known for its rich antioxidant content and low-calorie attributes, has garnered significant attention as a health-promoting fruit. Its economic value has also surged due to increasing consumer demand and its potential as an export commodity in various regions. The classification of dragon fruit ripeness is a pivotal task in ensuring product quality and minimizing post-harvest losses. This research article presents a comprehensive study on the classification of ripe and unripe dragon fruits (Hylocereus spp) using the Densenet201 model through three distinct approaches: as a classifier, feature extractor, and fine-tuner. To explain the outcomes of the image classification model and thereby enhance its performance, optimization, and reliability, this study employs advanced visualization techniques. Specifically, it utilizes Grad-CAM (Gradient-weighted Class Activation Mapping) and Guided Grad-CAM techniques. These techniques offer insights into the model's decision-making process and pinpoint regions of interest within the images. This approach empowers researchers to iteratively validate the model's accuracy and enhance its performance. The utilization of Densenet201 as a classifier, feature extractor, and fine-tuner, coupled with the insights from Grad-Cam and Guided Grad-Cam, presents a holistic approach to enhancing dragon fruit ripeness classification. The findings contribute to the broader discourse on agricultural technology, image analysis, and the optimization of classification models.**

***Keywords*—Dragon fruit classification; ripeness classification; densenet201 model; Grad-CAM visualization; guided grad-CAM; visual interpretation; Explainable AI; XAI; deep learning; pre-trained models; model fine-tuning; transfer learning**

I. Introduction

Dragon fruit is a tropical fruit that is widely grown in many countries around the world, including Vietnam [1]. It contains numerous nutrients beneficial to human health, such as vitamin A, vitamin C, and protein [2], [3]. While the ripeness of dragon fruit can be observed visually and harvested manually by humans, the necessity arises for the integration of equipment and robots in the field due to the vast planting area and technological advancements. The automated system is capable of efficiently harvesting significant quantities of dragon fruit within a brief timeframe, leading to time savings in the harvesting process. Therefore, an automatic dragon fruit ripeness grading system is really necessary.

Explaining deep learning models is of paramount importance in today's rapidly evolving technological landscape. Deep learning, with its complex architectures and black-box nature, has demonstrated remarkable capabilities across various domains. However, this complexity often comes at the cost of interpretability, creating challenges in understanding how and why these models arrive at their decisions. Explaining these models is a critical step to building trust, ensuring fairness, and enabling effective adoption. Interpretability empowers stakeholders to comprehend the factors influencing predictions. By shedding light on the inner workings of deep learning models, explanations help bridge the gap between advanced machine learning techniques and human comprehension, fostering collaboration between data scientists, domain experts, and end-users.

This research paper provides a comprehensive examination of how to classify ripe and unripe dragon fruits (Hylocereus spp) using the Densenet201 model. The study explores three distinct approaches to employing the model: as a classifier, a feature extractor, and a fine-tuner. To expound upon the outcomes of the image classification model and, in turn, enhance its performance and reliability, advanced visualization techniques are applied. Specifically, the study makes use of Grad-CAM (Gradient-weighted Class Activation Mapping) and Guided Grad-CAM techniques.

II. Related Works

Lately, the domains of image recognition have witnessed extensive utilization of Deep Learning (DL) and Machine Learning (ML) techniques. Including medicine [4], [5], [6], [7]. In self-driving cars [8], [9], [10], [11]. In agriculture [12], [13], [14], [15]. Especially when it comes to automating fruit sorting [16], [17], [18], etc.

Currently, there is a lot of applied research on machine learning, and deep learning on dragon fruit for many different purposes such as: In this study [19], Minh Trieu, N., & Thinh, N. T. present an automated system for classifying dragon fruit, which relies on a convolutional neural network (CNN). This classification system integrates machine learning and image processing through a convolutional neural network model to discern the external characteristics of dragon fruits. The paper [20] presents an automated dragon fruit classification system through a combination of KNN, CNN, and ANN models for identification, feature extraction, and classification. This study [21] devises dragon fruit grading and sorting techniques via machine learning algorithms (CNN, ANN, and SVM). Zhou et al. in this paper [22] presents a novel dragon fruit detection method, utilizing YOLOv7 to locate and classify the dragon fruit and further detect the endpoints of the dragon fruit. Vijayakumar, D. T., & Vinothkanna, M. R. in the paper

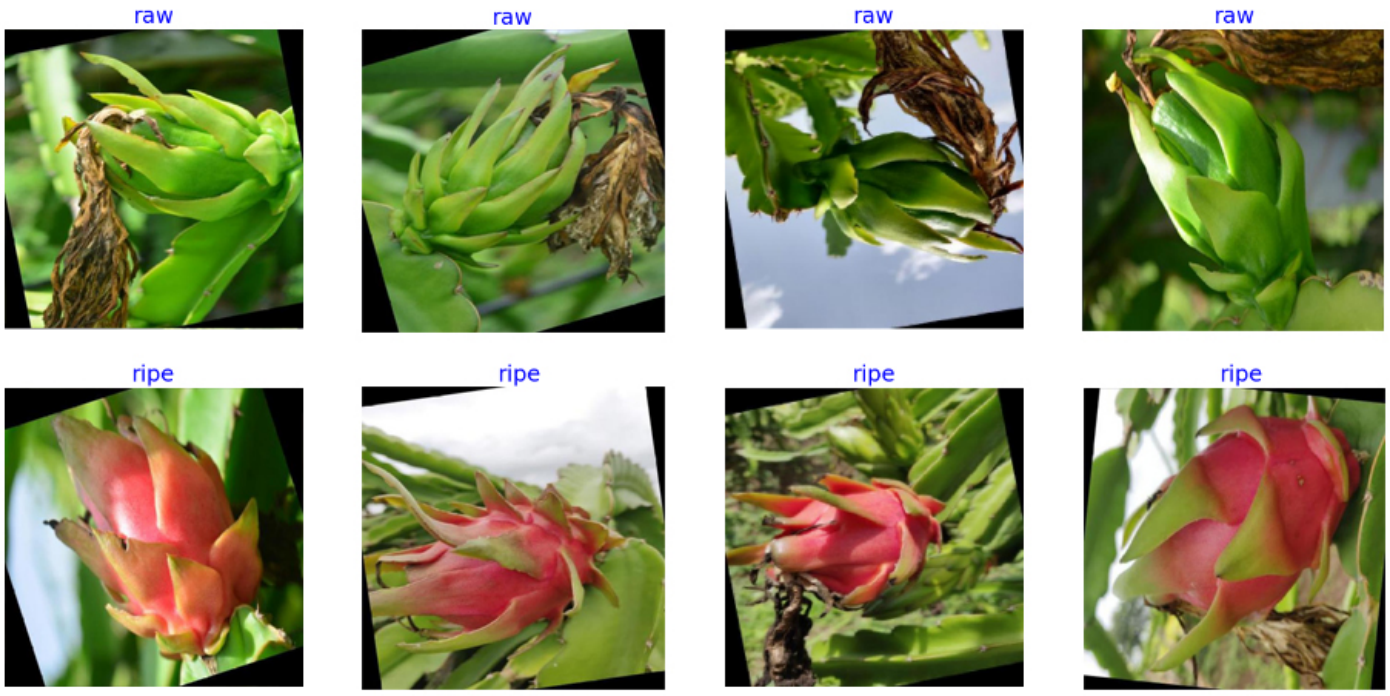


Fig. 1. Some images of dragon fruit dataset.

[23] introduces the utilization of the RESNET 152 deep learning convolutional neural network to identify dragon fruit mellowness, signifying the optimal time for harvest. This study [24] applies color and texture feature extraction techniques, utilizing color moments and gray level co-occurrence matrices (GLCM), to develop a system for recognizing three types of dragon fruit stems through digital image processing, employing Support Vector Machine and k-Nearest Neighbors methods for comparison.

In today's context, with the increasing complexity of deep learning models, there is an even greater need to understand the decisions made by these models. The role of Explainable Artificial Intelligence (XAI) [25], [26] is to provide transparency, clarity, and understanding to the decision-making process of deep learning models and bring the gap between the "black-box" nature of AI and human understanding. XAI has achieved numerous successes across various domains, including: In medicine [27], [28], [29]; In agriculture: [30]; In traffic classification: [31].

The purpose of this research article is to conduct a thorough investigation into the classification of ripe and unripe dragon fruits utilizing the Densenet201 model. The study explores three distinct methodologies: employing the model as a classifier, feature extractor, and fine-tuner. In addition, using Grab-CAM [32], [33], [34] and Guided Grad-Cam (this involves performing an element-wise product between Grad-CAM and Guided Backpropagation [35]) to interpret models for the purpose of evaluating the decision and effectiveness of deep learning models detecting features in a dragon fruit image.

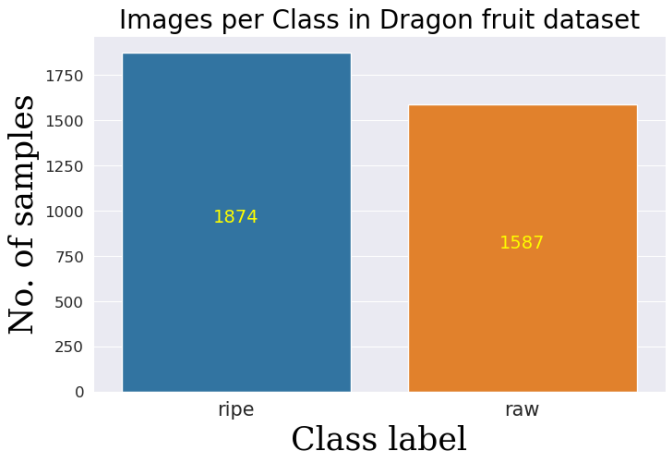


Fig. 2. A dataset distribution.

III. Methodology

A. Data Collection and Preparation

In this investigation, we utilized a dataset containing 3,461 images, sourced from [36] with two classes: Ripe and Raw. Fig. 1 illustrates sample images of dragon fruit from this dataset, and the dataset distribution is visually depicted in Fig. 2. Prior to model training and evaluation, a preprocessing step is performed on the images, resizing them to 224x224 and applying an image preprocessing function. The data set is partitioned into three distinct sets, comprising a training set, a validation set, and a testing set, with a distribution ratio of 6:2:2.

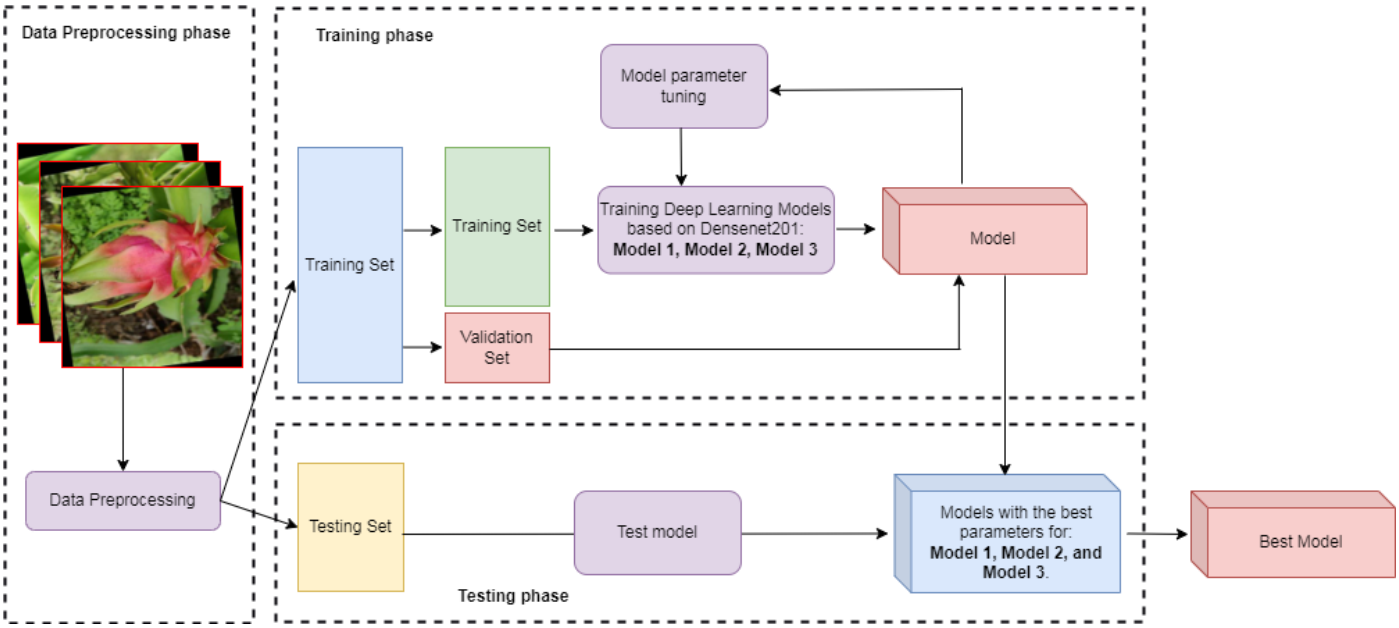


Fig. 3. The training and testing process involves three dragon fruit ripeness models based on densenet201.

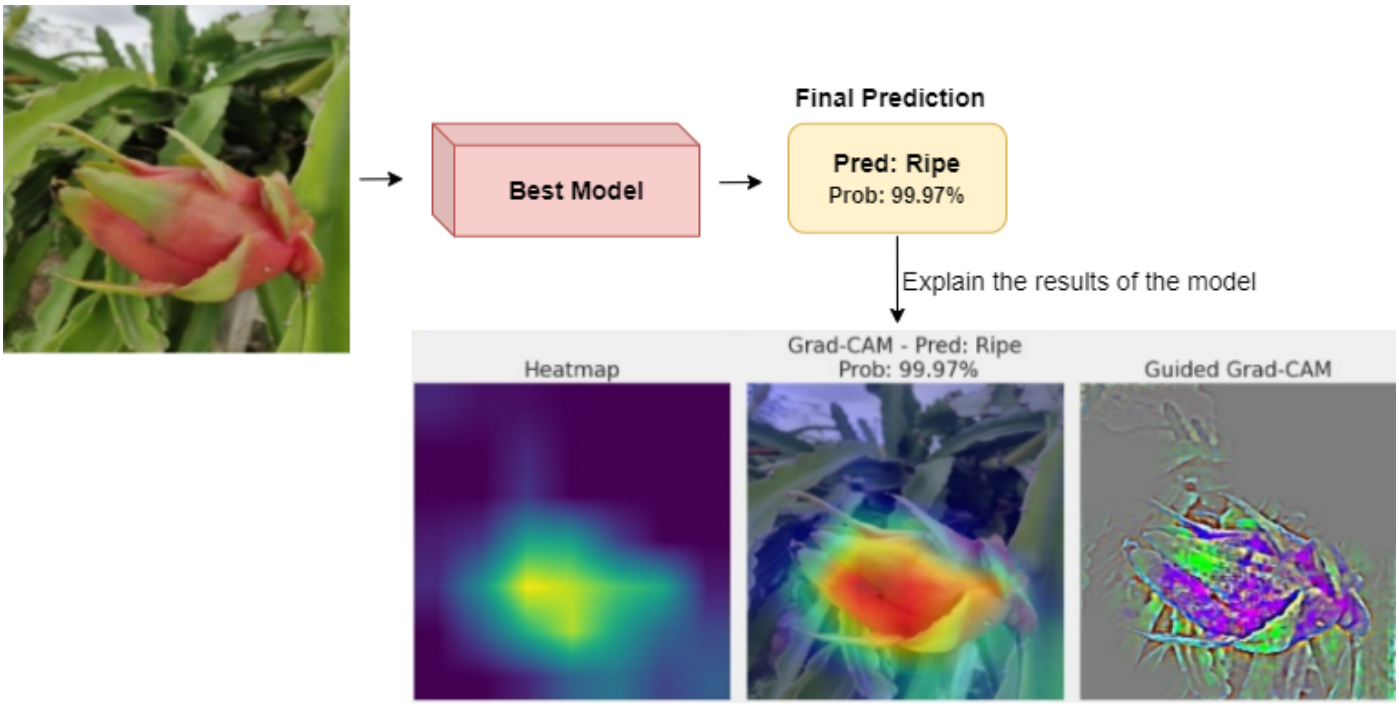


Fig. 4. Classify the dragon fruit ripeness and explain the results of the model.

B. Transfer Learning for Dragon Fruit Classification

Transfer learning [37], [38], a pivotal concept in the realm of machine learning, is a technique that leverages knowledge gained from one task to improve performance on a different but related task. Rather than starting from scratch, transfer learning enables models to take advantage of patterns and representations learned from a source domain and apply them to a target domain with limited labeled data. This approach has revolutionized the field by dramatically reducing the need for vast datasets and extensive computational resources, making it feasible to tackle new problems even when data is scarce. Transfer learning's ability to extract and transfer valuable insights across tasks has been instrumental in advancing the efficiency, accuracy, and generalization of machine learning models, thereby accelerating progress across various domains and enabling AI systems to learn more like humans – by

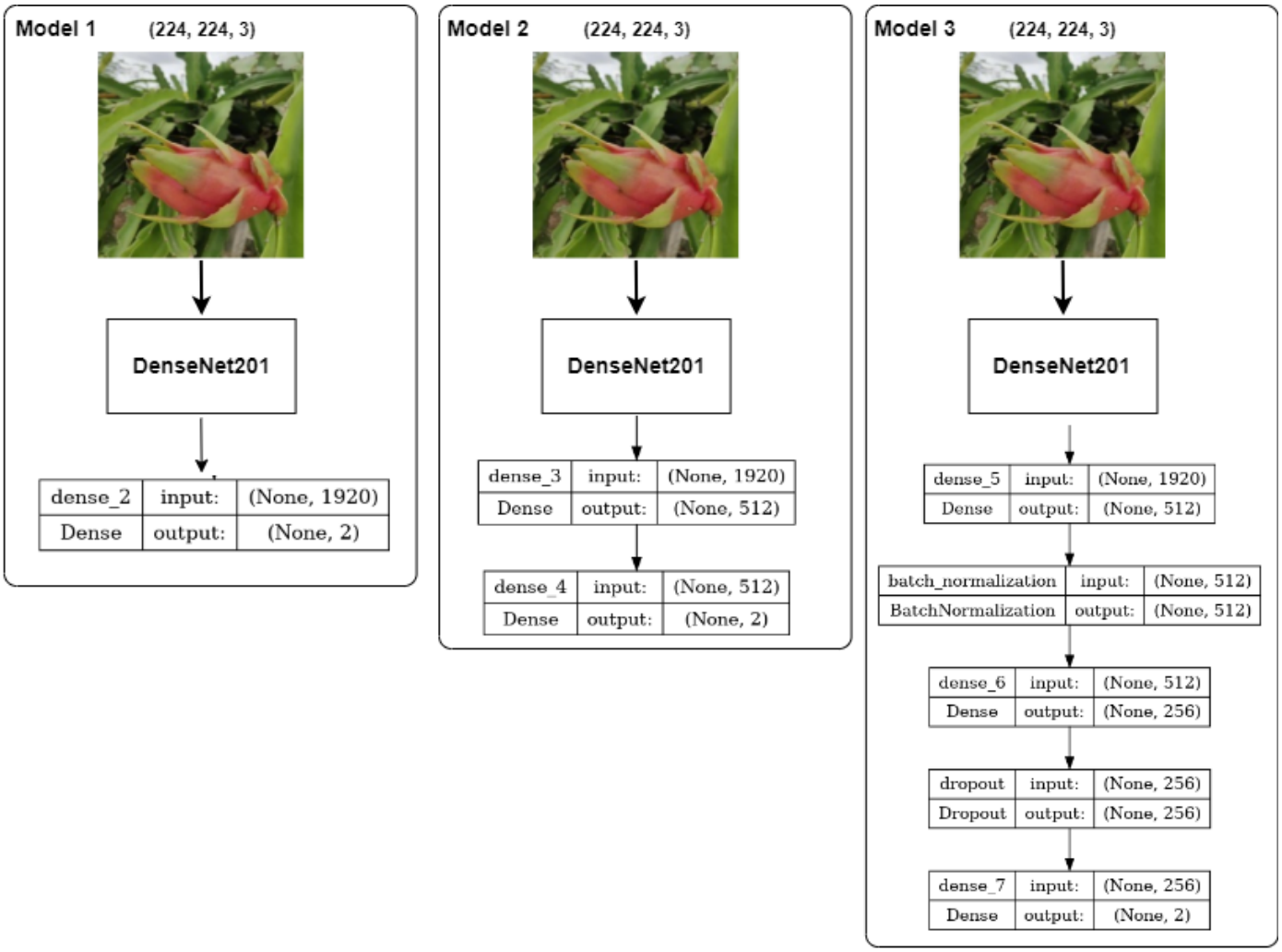


Fig. 5. The architecture of the dragon fruit ripeness classification is constructed using denseNet201.

building upon prior knowledge to solve new challenges.

C. Proposed Model

This research paper presents a comprehensive investigation into the categorization of ripe and unripe dragon fruits utilizing the Densenet201 [39] model in three separate ways: as a classifier, feature extractor, and fine-tuner. Model 1 (as a classifier) is built upon the DenseNet201 architecture. The pre-trained layers are frozen to preserve their acquired features. The final softmax layer in the DenseNet201 model has been modified to classify two classes, corresponding to ripe and raw dragon fruit. Model 2 (as a feature extractor) is a modified DenseNet201-based model for the binary classification of ripe and raw dragon fruit. It has two new entirely connected layers, with 512 neurons in the first layer and two neurons in the final. Both the base pre-trained layers and the fully connected layers have been completely retrained. Model 3 (as a feature extractor with fine-tuning) is a base model initialized using the DenseNet201 architecture. A fully connected layer with 512 neurons is added on top of the base model's output. Batch normalization is applied to the output of the previous layer, helping to stabilize and accelerate training. Another fully connected layer with 256 neurons is added, along with various regularization techniques. A dropout layer is introduced, which helps in preventing overfitting. A final fully connected layer with 2 neurons and softmax activation is added to produce class probabilities. The pre-trained layers are frozen to preserve their acquired feature. Fig. 3 shows The training and testing process involves three Dragon Fruit Ripeness models based on Densenet201. Fig. 4 displays Classify the Dragon Fruit Ripeness and explain the results of the model. The architecture of the Dragon Fruit Ripeness Classification is constructed using DenseNet201 shown in Fig. 5.

D. Performance Evaluation Measures

In this study, various evaluation metrics, including Accuracy, F1-score, Precision, and Recall, were utilized to assess the effectiveness of the deep learning (DL) models. Accuracy served as a measure of overall performance, while Precision and Recall evaluated the model's ability to correctly predict positive instances. The F1-score provided a balanced perspective by considering both Precision and Recall, enabling informed judgments on the model's effectiveness. Through the

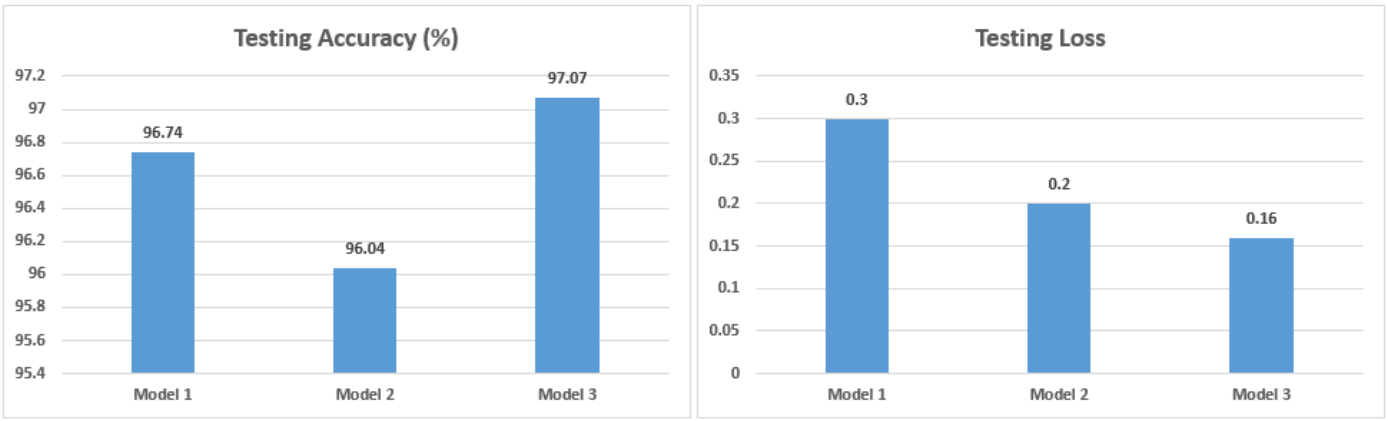


Fig. 6. Confusion matrix of the recommended models. (left) The number of predictions, (right) The percentage.

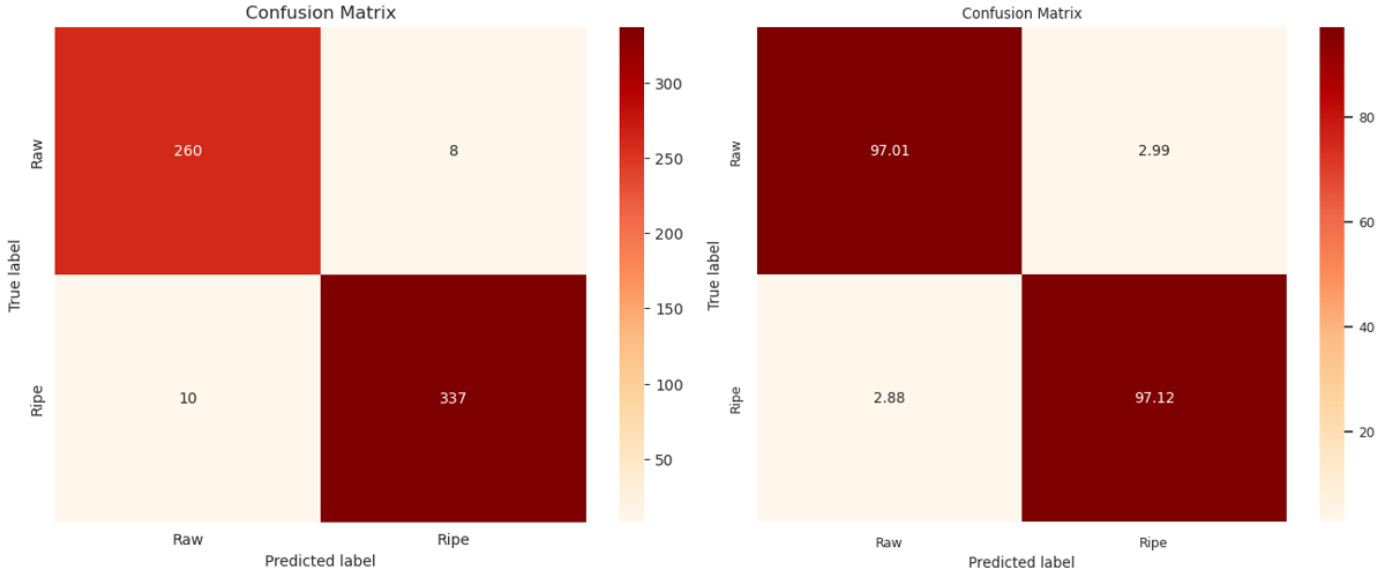


Fig. 7. Confusion matrix of Model 3. (left) The number of predictions, (right) The percentage.

utilization of various evaluation metrics, a thorough comprehension of the model's performance was attained.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F_1 - Score = \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

In which, TP represents True Positive, TN signifies True Negative, FP represents False Positive, and FN stands for False Negative.

IV. Results

A. Environmental Settings

The experimental results were obtained by conducting the experiments on the Kaggle platform. The system used for the experiments had 13GB of RAM and a GPU P100 with 16GB of memory. The models were trained for a total of 35 epochs, and a batch size of 32 was used during the training process.

B. Evaluation Overall

The results presented in Fig. 6 reveal that the majority of the models exhibit remarkably high accuracy and low loss. Notably, Model 3 stands out with the lowest test loss of 0.1606 and the highest accuracy of 97.07% among the three models. It demonstrates excellence on the test dataset. Additionally, both Model 1 and Model 2 perform exceptionally well, achieving test accuracies of 96.75% and 96.04%, respectively. However, they do have significantly higher test losses compared to Model 3.

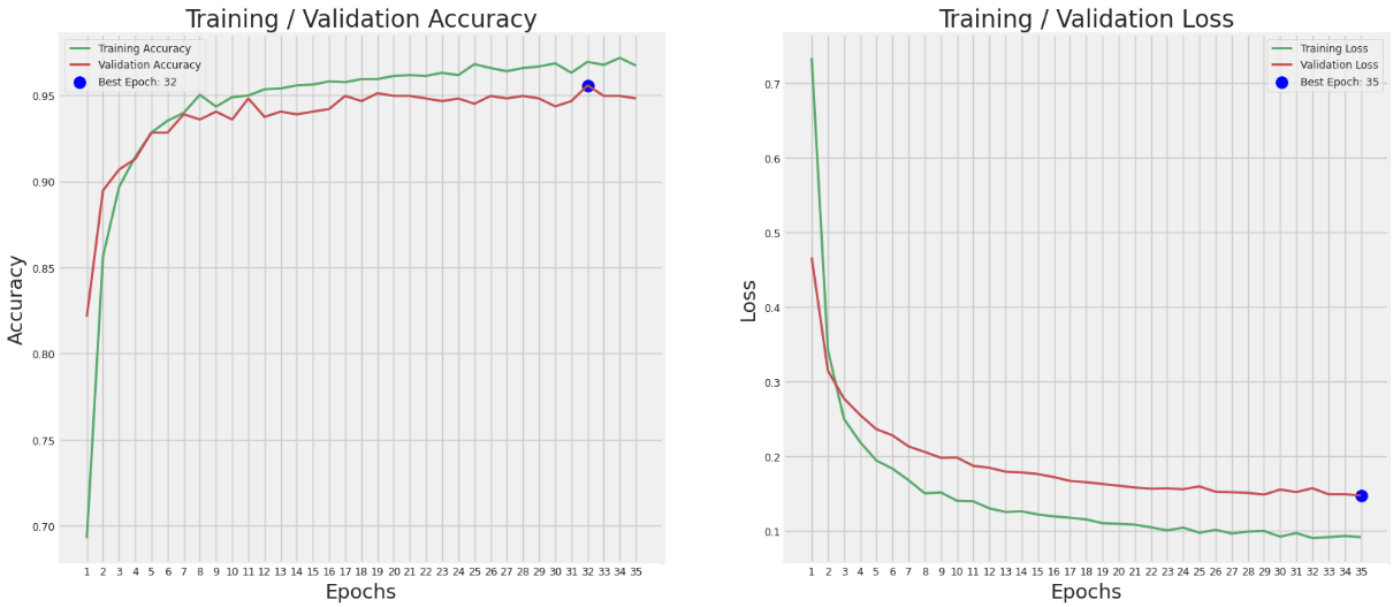


Fig. 8. Loss and accuracy plots of the model 3.

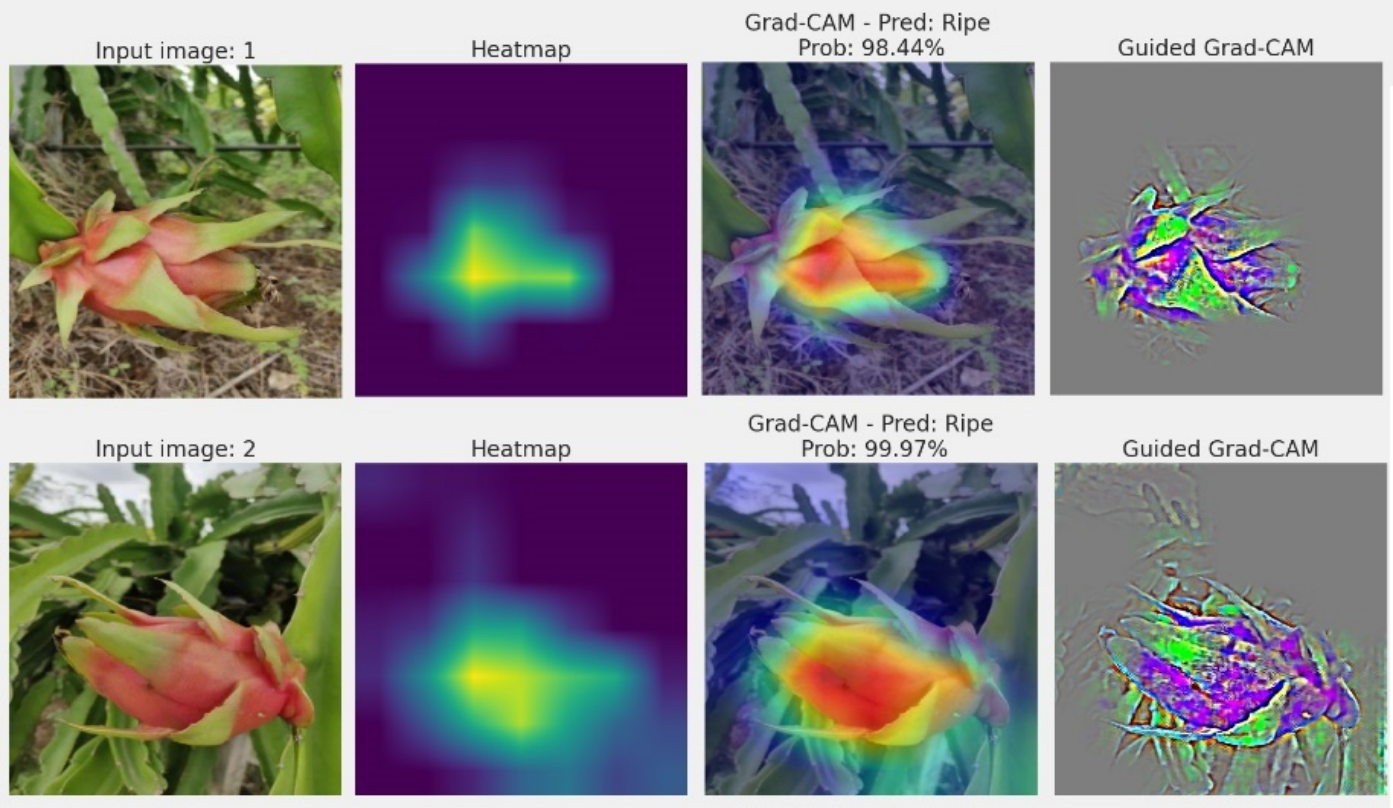


Fig. 9. Examples of the ripe feature explained using Grad-CAM from model 3.

The accuracy of Model 3 reaching 97.07% on the test set is outstanding and represents an exceptional level of performance. It means that the model can correctly classify almost all instances in the test data. Such a high level of accuracy is a strong indicator of the model's ability to generalize well and make precise predictions on previously unseen data. Fig. 7 illustrates the confusion matrix for Model 3.

Furthermore, Fig. 8 presents an overview of the performance metrics, including loss and accuracy, which were assessed throughout both the training and validation stages of Model 3.

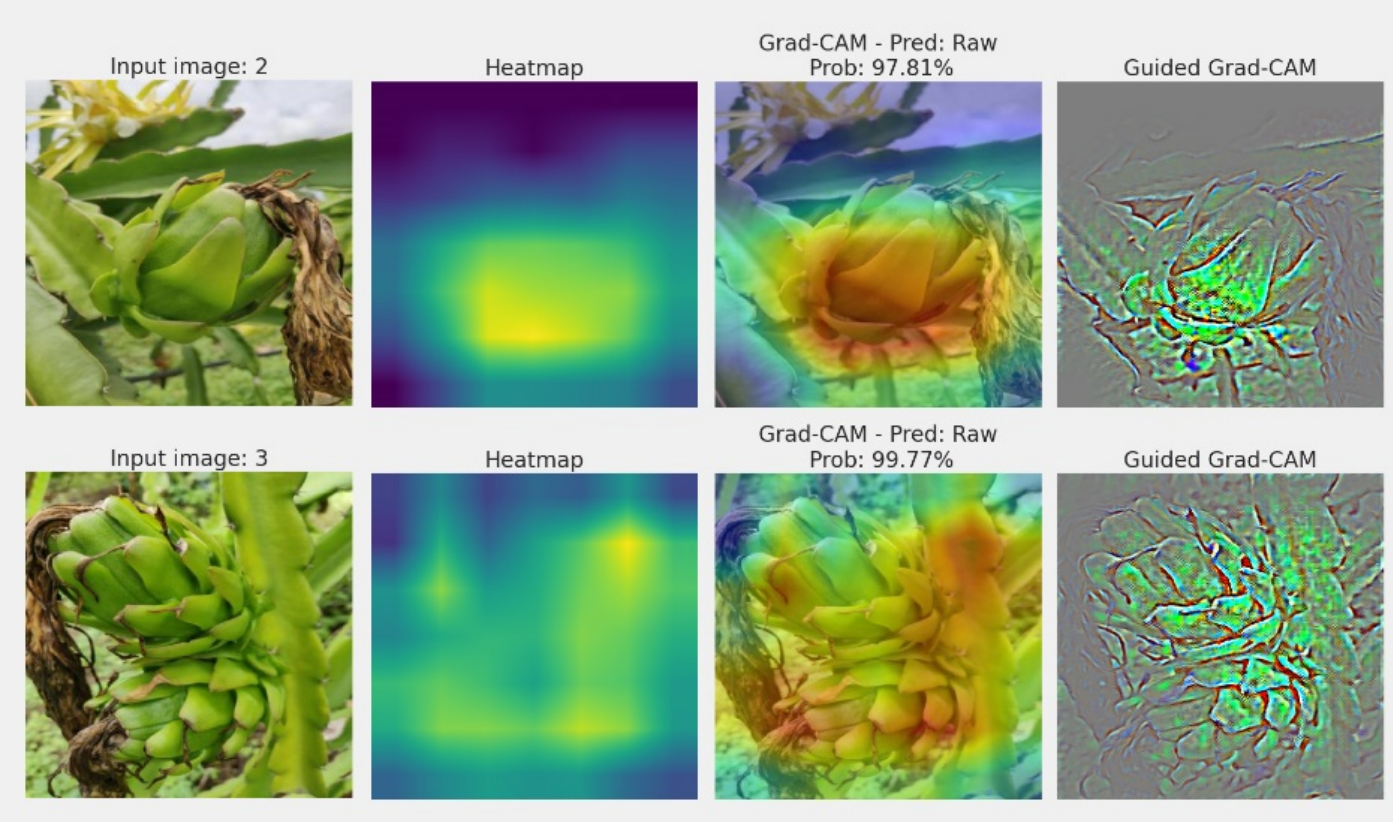


Fig. 10. Examples of the raw feature explained using Grad-CAM from model 3.

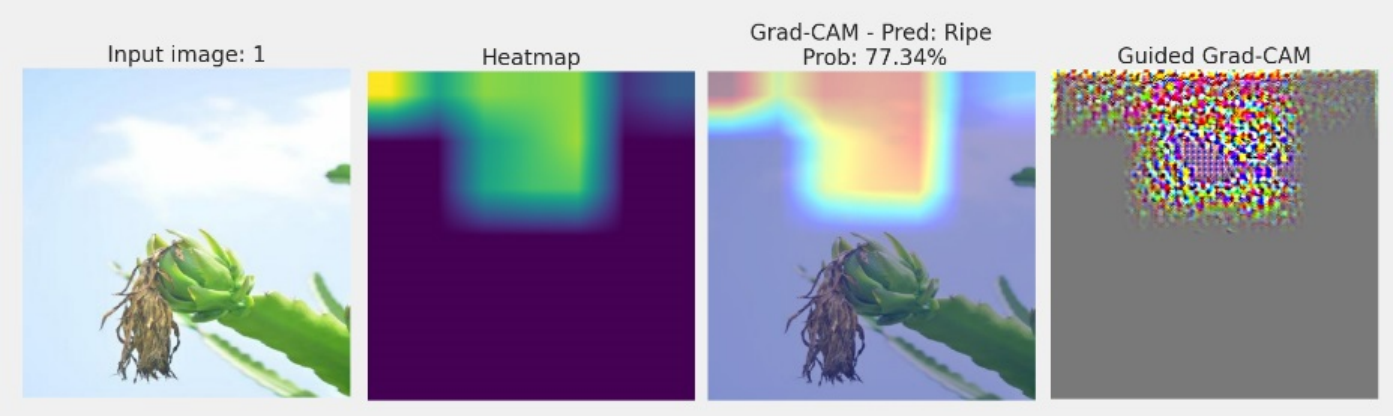


Fig. 11. Examples of ripe feature misclassification are explained using Grad-CAM from model 3.

The classification report Table I provides a detailed analysis of the evaluation metrics for each grape disease class. The classification report is a summary that assesses a model's classification performance. It includes precision, recall, and F1-score metrics for each class label. The report helps to evaluate the model's accuracy in correctly classifying instances for each class, with higher scores indicating better performance. In this case, the model achieved perfect scores for all classes, demonstrating excellent accuracy in its classification task.

TABLE I. THE MODEL RESULTS OF THE CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
Raw	0.96	0.97	0.97	268
Ripe	0.98	0.97	0.97	347

C. Visualizing the Interpretation of Model Predictions Using Grad-CAM

To better understand the significant regions in the images that the model focuses on for making predictions, the research team employed Grad-CAM on corner images. Specif-

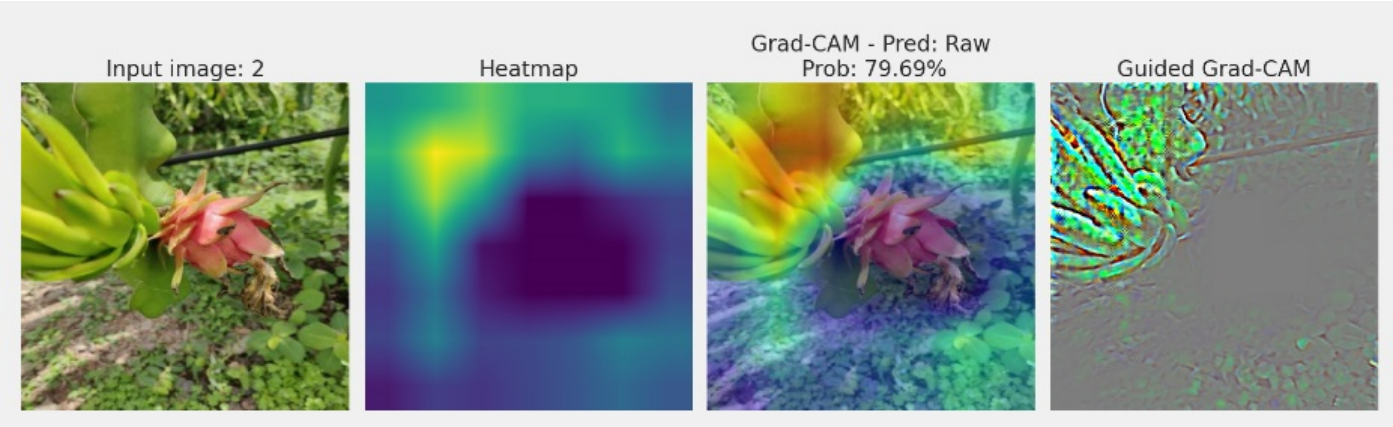


Fig. 12. Examples of raw feature misclassification are explained using Grad-CAM from model 3.

ically, when applying Grad-CAM to an image, it generates a "heatmap" that highlights important positions in the image. This heatmap indicates the areas of the image that the model is paying attention to while making predictions. High values on the heatmap usually correspond to important regions relevant to the classification decision. In addition, the research team also used Guided Grad-CAM to better understand the parts of the image that the model is interested in to make predictions.

Based on the results from Fig. 9 and Fig. 10, it is evident that the accurate classification by the model relies on specific regions unique to each type. The heatmap also highlights key areas in the image that the model focuses on while making predictions that align with its own predictions.

Grad-CAM and Guided Grad-CAM are also useful in comparing misclassified results to understand why the model may have "misinterpreted" certain images. For example, Fig. 11 and 12 explain why the model is misclassified.

V. Conclusion

This study conducts an extensive investigation into the classification of ripe and unripe dragon fruit, employing the Densenet201 model across three distinct approaches: as a classifier, feature extractor, and fine-tuner. All three proposed models yield exceptionally impressive outcomes. Particularly noteworthy, Model 3 (functioning as a feature extractor with fine-tuning) stands out with the highest accuracy, achieving 97.07% among the three models. Furthermore, both Model 1 and Model 2 showcase exceptional performance. To delve deeper into understanding the significant areas within images that the model emphasizes for prediction, the research team applied Grad-CAM to corner images. Additionally, the team employed Guided Grad-CAM to enhance comprehension of image regions that capture the model's attention for prediction purposes. Both Grad-CAM and Guided Grad-CAM prove to be invaluable tools in the comparative analysis of misclassified results, providing insights into the potential reasons for the model's "misinterpretation" of specific images. The study team will use HiResCAM [40] in the future for model explanation. HiResCAM serves the same functions as Grad-CAM but with the added benefit of highlighting only the regions used by the model.

References

[1] Edmundo M Mercado-Silva. Pitaya—hylocereus undatus (haw). *Exotic fruits*, pages 339–349, 2018.

[2] Md Farid Hossain, Sharker Md Numan, and Shaheen Akhtar. Cultivation, nutritional value and health benefits of dragon fruit (hylocereus spp.): A review. *International Journal of Horticultural Science & Technology*, 8(3), 2021.

[3] Truc-Linh Le, Nga Huynh, Pablo Quintela-Alonso, et al. Dragon fruit: A review of health benefits and nutrients and its sustainable development under climate changes in vietnam. *Czech Journal of Food Sciences*, 39(2):71–94, 2021.

[4] Fei Wang, Lawrence Peter Casalino, and Dhruv Khullar. Deep learning in medicine—promise, progress, and challenges. *JAMA internal medicine*, 179(3):293–294, 2019.

[5] Francesco Piccialli, Vittorio Di Somma, Fabio Giampaolo, Salvatore Cuomo, and Giancarlo Fortino. A survey on deep learning in medicine: Why, how and when? *Information Fusion*, 66:111–137, 2021.

[6] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.

[7] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps: Automation of Decision Making*, pages 323–350, 2018.

[8] Qing Rao and Jelena Frtunikj. Deep learning for self-driving cars: Chances and challenges. In *Proceedings of the 1st international workshop on software engineering for AI in autonomous systems*, pages 35–38, 2018.

[9] Jianjun Ni, Yinan Chen, Yan Chen, Jinxiu Zhu, Deena Ali, and Weidong Cao. A survey on theories and applications for self-driving cars based on deep learning methods. *Applied Sciences*, 10(8):2749, 2020.

[10] Truong-Dong Do, Minh-Thien Duong, Quoc-Vu Dang, and My-Ha Le. Real-time self-driving car navigation using deep neural network. In *2018 4th International Conference on Green Technology and Sustainable Development (GTSD)*, pages 7–12. IEEE, 2018.

[11] Zhenchao Ouyang, Jianwei Niu, Yu Liu, and Mohsen Guizani. Deep cnn-based real-time traffic light detector for self-driving vehicles. *IEEE transactions on Mobile Computing*, 19(2):300–313, 2019.

[12] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90, 2018.

[13] Nanyang Zhu, Xu Liu, Ziqian Liu, Kai Hu, Yingkuan Wang, Jinglu Tan, Min Huang, Qibing Zhu, Xunsheng Ji, Yongnian Jiang, et al. Deep learning for smart agriculture: Concepts, tools, applications, and opportunities. *International Journal of Agricultural and Biological Engineering*, 11(4):32–44, 2018.

[14] Luís Santos, Filipe N Santos, Paulo Moura Oliveira, and Pranjali Shinde. Deep learning applications in agriculture: A short review. In *Robot 2019: Fourth Iberian Robotics Conference: Advances in Robotics, Volume 1*, pages 139–151. Springer, 2020.

[15] Muhammad Hammad Saleem, Johan Potgieter, and Khalid Mahmood Arif. Automation in agriculture by machine and deep learning techniques: A review of recent developments. *Precision Agriculture*, 22:2053–2091, 2021.

[16] Chandra Sekhar Nandi, Bipan Tudu, and Chiranjib Koley. An automated machine vision based system for fruit sorting and grading. In *2012 Sixth International Conference on Sensing Technology (ICST)*, pages 195–200. IEEE, 2012.

[17] Rashmi Pandey, Sapan Naik, and Roma Marfatia. Image processing and machine learning for automated fruit grading system: A technical review. *International Journal of Computer Applications*, 81(16):29–39, 2013.

[18] CS Nandi, B Tudu, and C Koley. Machine vision based techniques for automatic mango fruit sorting and grading based on maturity level and size. *Sensing technology: current status and future trends II*, pages 27–46, 2014.

[19] Nguyen Minh Trieu and Nguyen Truong Thinh. Quality classification of dragon fruits based on external performance using a convolutional neural network. *Applied Sciences*, 11(22):10558, 2021.

[20] Nguyen Minh Trieu and Nguyen Truong Thinh. A study of combining knn and ann for classifying dragon fruits automatically. *Journal of Image and Graphics*, 10(1):28–35, 2022.

[21] Pallavi U Patil, Sudhir B Lande, Vinay J Nagalkar, Sonal B Nikam, and GC Wakchaure. Grading and sorting technique of dragon fruits using machine learning algorithms. *Journal of Agriculture and Food Research*, 4:100118, 2021.

[22] Jialiang Zhou, Yueyue Zhang, and Jinpeng Wang. A dragon fruit picking detection method based on yolov7 and psp-ellipse. *Sensors*, 23(8):3803, 2023.

[23] Dr T Vijayakumar and Mr R Vinothkanna. Mellowness detection of dragon fruit using deep learning strategy. *Journal of Innovative Image Processing*, 2(1):35–43, 2020.

[24] Lutfi Hakim, Sepyan Purnama Kristanto, Dianni Yusuf, Mohammad Nur Shodiq, and Wahyu Ade Setiawan. Disease detection of dragon fruit stem based on the combined features of color and texture. *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, 5(2):161–175, 2021.

[25] David Gunning. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web*, 2(2):1, 2017.

[26] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

[27] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of imaging*, 6(6):52, 2020.

[28] Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. Using stylegan for visual interpretability of deep learning models on medical images. *arXiv preprint arXiv:2101.07563*, 2021.

[29] Loveleen Gaur, Mohan Bhandari, Tanvi Razdan, Saurav Mallik, and Zhongming Zhao. Explanation-driven deep learning model for prediction of brain tumour status using mri image data. *Frontiers in genetics*, 13:448, 2022.

[30] Luyl-Da Quach, Khang Nguyen Quoc, Anh Nguyen Quynh, Nghe Nguyen Thai, and Tri Gia Nguyen. Using gradient-weighted class activation mapping to explain deep learning models on agricultural dataset. *IEEE Access*, 2023.

[31] Seyoung Ahn, Jeehyeong Kim, Soo Young Park, and Sunghyun Cho. Explaining deep learning-based traffic classification using a genetic algorithm. *IEEE Access*, 9:4738–4751, 2020.

[32] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[33] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

[34] Harsh Panwar, PK Gupta, Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Prakhar Bhardwaj, and Vaishnavi Singh. A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images. *Chaos, Solitons & Fractals*, 140:110190, 2020.

[35] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[36] first. rp dataset. https://universe.roboflow.com/first-zr3cq/rp-rombw, apr 2023. visited on 2023-07-31.

[37] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[38] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[39] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[40] Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020.

Identification of Air-Writing Tamil Alphabetical Vowel Characters

Rukshani Puvanendran
Dept. of ICT
University of Vavuniya
Vavuniya, Sri Lanka

Vijayanathan Senthooran
Dept. of ICT
University of Vavuniya
Vavuniya, Sri Lanka

***Abstract*—In recent years, there has been a lot of focus on gesture recognition because of its potential as a means of communication for cutting-edge gadgets. As a special category of gesture recognition, air-writing is the practice of forming letters or words in the air using one's fingers or the movements of one's hands. The primary objective of this study is to propose a classification framework with feature extraction techniques to enhance the recognition of vowel characters in the Tamil language. The data collection and classification procedure involved a set of 12 distinct letters. A methodology has been developed to facilitate the analysis of various configurations for the purpose of evaluation. To get useful features from the 2-second time window data segments, this study uses a one-dimensional convolutional neural network (1D CNN). In our approach, we employ five machine-learning methods to conduct our evaluation. These methods include Naive Bayes, Random Forest, K-Nearest Neighbor, Support Vector Machine, and Decision Tree. The classification algorithms are considered to be superior based on the results obtained from our dataset in this experiment. The results of the tests show that the suggested K-nearest neighbors (KNN) algorithm works very well when used with a k-1 and 0.6:0.4 split ratio for training and testing. Specifically, the KNN model achieved an accuracy rate of 91.67%. The present study builds upon previous research by utilizing applications that have been employed in prior studies. However, a unique aspect of our system is the integration of cutting-edge technology, which utilizes collected sensor data to classify the characters. The examination of the window size has the potential to enhance accuracy and performance.**

***Keywords*—Air-writing; Tamil alphabetical vowel; convolutional neural network; feature extraction; machine learning**

I. Introduction

The recent advancement in technology breaks the barriers to communication between users and computers. The communication between users and computers includes emotion and gesture recognition. Gestures offer a complementary mode of interaction for standard human-computer communication. New types of user interfaces made possible by gestures are especially well suited for portable and wearable computer systems. As a subset of gesture recognition, air-writing recognizes characters and numbers written in the air using the hand. Touchscreens and other technological gadgets have become fashionable in today's information-driven society, and a majority of individuals are comfortable using them [1]. As a result, air-writing recognition systems are gradually becoming more adaptable using wearable sensors and state-of-the-art techniques [2], [3], [4], [5].

Languages consist of alphabets that are combined to make meaningful words and are utilized for communication [6]. Therefore, these alphabets possess both symbolic and phonetic characteristics [6], [7]. The recognition of the air-writing characters of a particular language can be distinguished based on their symbolic structure [7], [8]. Most alphabets in any language consist of several complex motions, which means that sometimes a single sensor may not be enough to recognize these alphabets.

Usually, traditional handwriting styles of languages encompass two primary forms: cursive and print letters. Air-writing character recognition is done using information gathered from six-degree free-range hand motions. In air-writing each isolated letter is written in a virtual space. For user interfaces that do not allow users to type on a keyboard or write on a trackpad or touch screen, as well as for text input for controlling smart systems and many other applications, air-writing is extremely helpful [9], [10], [11]. Air-writing has been found to be a highly effective method for various applications, including but not limited to device control, entertainment, health care, and education.

The categorization of air-writing recognition approaches can be delineated into two distinct classes: Vision-Based Recognition (VBR) algorithms and Sensor-Based Recognition (SBR) algorithms [3], [12], [13]. The VBR algorithms are designed to execute gesture recognition tasks using image data acquired through a camera device. According to Amma's research, it has been observed that achieving accurate classification in image processing tasks can be feasible [4]. However, it is important to note that this process often demands substantial computational resources. Specifically, significant computational efforts are necessary for extracting relevant information from images during both the training and inference stages. This finding underscores the importance of considering the computational aspect when designing and implementing image processing algorithms [12]. The study conducted by Amma et al. [4], sheds light on the potential challenges associated with image-based classification tasks and highlights the need for efficient computational strategies in order to achieve optimal results. The techniques employed in SBR (Sensor-Based Recognition) are fundamentally rooted in the utilization of sensors. Accelerometers, gyroscopes, flex sensors, electromyography (EMG), Radio Frequency Identification (RFID), and the integration of these sensors have been widely employed in various applications [5], [14].

In the context, of hand gesture identification using SBR,

accelerometers are widely used sensors in various applications because of their practicality, affordability, and durability [13], [15]. They have been extensively utilized in numerous prediction systems [16]. Wearable technology, such as smart phones, provides an optimal platform for the collection and monitoring of data [10], [17]. Furthermore, it has been reported that placing the sensor component on the wrist is the optimal placement position for a high degree of accuracy [5]. In some SBR-based studies, high classification accuracy for gesture recognition has been observed [2].

However, many of these techniques do not support the recognition of a sequence of gestures. Only isolated gestures can be recognized [19]. Furthermore, some SBR techniques are used for gesture recognition, which may incur high computation complexities [18]. The Recurrent Neural Networks (RNNs) and their variants, such as the Long Short-Term Memory (LSTM) algorithm, are effective for the recognition of gesture sequences with low computation costs [12], [19], [20], [21].

Many emerging algorithms have been proposed as the availability of temporal data has grown substantially in recent years [22]. To show time series as feature vectors, people often use dynamic time warping (DTW), simple statistics, complex mathematical methods, and other similar methods. For classification, they use algorithms that range from shallow learning to deep neural network models [23]. In addition, models like recurrent and convolutional neural networks (CNNs) incorporate feature engineering internally and automatically [24]. The One-Dimensional (1D) CNNs are effective for several applications, such as the classification of ECG signals [19], human activities, and internet traffic [16], [18], [25], [26], [27]. Researchers have found that using CNNs to classify time series is better than other methods in a number of important ways. This is because CNNs are very good at ignoring noise and can pull out very useful, deep features that are not affected by time [28]. In the context of CNN, one-dimensional convolutional neural networks (1D CNNs) excel at extracting useful features from smaller (fixed-length) subsets of a larger data set [29] and at analyzing audio signals, cyber security, NLP, and time series data from sensors [18], [30].

The majority of previous studies focused on the recognition of the English alphabet, where most of the alphabets are mostly drawn with lines that could subsequently be identified with the changes of the sensors [31], [32]. Some of the previous studies dealt with the designated devices for data collection and identification of the characters in second-language learning [33], [34]. Moreover, no prominent dataset is shared in publicly available sources.

Sensor-based air-writing has significant importance for the Tamil language alphabets. The recognition of handwritten words from a digital writing pad using sensor-based techniques can greatly benefit Tamil language learners and users. Jayanthi & Thenmalar (2023) proposed a method for recognizing handwritten words from a digital writing pad using the MMU-SNet algorithm [8]. The Tamil script is conventionally written in a horizontal manner, progressing from left to right. Its fundamental repertoire of characters encompasses 247 letters, including 12 vowels, 18 consonants, 12 vowels by 18 consonants (216), and 1 unique character [7]. The Tamil script is commonly recognized for its distinctive rounded shapes, leading to its colloquial designation as the "round alphabet" [6]. This approach can accurately capture the unique characteristics of Tamil alphabets, which have a smaller number of lines, angles, curves, and bends.

The study mainly improves the motion gesture recognition approach for six degrees of freedom (DOF) for air-written Tamil language characters. In the initial phase of the study, in spite of the motion movements, isolated air-writing characters are analyzed and identified. Motion sensing is done with sensors that are attached to the back of the palm. This led to a great deal of interest in the potential of wearable technologies for air-writing. A comprehensive list of features is extracted from the dataset of 3-axis signals from five different sensors with 1D CNN.

The present study focuses on the investigation of the recognition process pertaining to isolated characters that are written in a continuous single-stroke manner. This study employed one specific dataset, which was collected as part of the study. To sum up, this study presents the following contributions to the field of air-writing:

The authors have

- Collected the sensor data for 12 vowels of the Tamil alphabet.
- Developed a light feature extraction approach using 1D CNN from the collected signal data and populated the dataset for the classification.
- Evaluated the performance of the classification models with few configurations.

By leveraging this study, this method enables the recognition of Tamil alphabets written in the air, providing a valuable tool for Tamil language learners and users.

The structure of this article is as follows: We go over the relevant earlier work in the part after that. The data collection techniques for air-writing, the feature extraction process, and the methods for modeling are described in Section III. Section IV presents the experimental design results and discussions. Section V contains the conclusion of the study.

II. Related Works

Air-writing refers to the process of writing characters or words in free space using finger or hand movements [31]. It is a form of gesture recognition where gestures correspond to characters and digits written in the air [32]. Air-writing has been studied in various contexts, including its effectiveness in language learning and its potential applications in handwriting analysis and recognition.

In previous studies, certain research endeavors have employed a specifically engineered apparatus to carry out the process of data acquisition. In contemporary times, a plethora of wearable devices, including but not limited to smartphones and smartwatches, have been integrated with various sensors such as accelerometers, gyroscopes, magnetometers, and other similar components. Various methodologies have been put forth in scholarly works to address the task of air-writing recognition, wherein it is regarded as a spatial-temporal signal [14].

The wearable input system for handwriting recognition proposed by Amma et al. in [4] utilizes an accelerometer and a gyroscope to effectively capture and analyze each of the complex gestures involved in air-writing. Another study conducted by [5], a range of alterations to the Hidden Markov Models (HMMs) were implemented in order to facilitate the identification and interpretation of air-writing patterns that were generated through the utilization of the Leap Motion Controller.

Agrawal et al. [2] introduced the PhonePoint Pen system on the Nokia N95, which leverages the inherent accelerometers present in mobile phones for the purpose of recognizing handwritten English characters. The findings of this study indicate that the identification of English characters can be achieved with a mean accuracy rate of 91.9%.

Furthermore, Xia et al. [10] presented a MotionHacker system that utilizes a smartwatch application to capture and analyze the dynamics of hand movement, thereby demonstrating the potential for motion sensor-based handwriting.

The proposed system in [15] introduces a novel approach to handwritten character recognition. It leverages the power of 3D accelerometer signal processing in conjunction with Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models. In the experimental setup, a participant carrying a MYO armband. The data in the dataset corresponds to handwritten English lowercase letters (a–z) and digits (0–9) that were written in a freestyle manner. The findings indicate that the proposed system exhibits superior performance compared to other existing systems, surpassing them by a margin of 0.53%.

Wang and Chuang in [13] proposed a novel digital pen that utilizes an accelerometer for the purpose of recognizing handwritten digits and gesture trajectories. Subsequently, the collected acceleration signals are subjected to a process of feature extraction, wherein a hybrid approach was employed to identify the most discriminant features. This involved the utilization of kernel-based class separability to determine the salient features, followed by the application of Linear Discriminant Analysis (LDA) to reduce the dimensionality of the dataset.

Two air-writing recognition methods were discussed in this study [1] dynamic time-warping and a Convolutional Neural Network. To assess the efficacy of the methodologies, a total of 15 instances of the English alphabet letters were obtained through airborne inscriptions and subsequently captured utilizing a smart-band device. These data sets were procured from a diverse pool of 55 individuals. The findings from the comprehensive evaluation yield an exceptional accuracy rate of 89.2%.

The study [35], introduces a novel framework for air-writing using a convolutional neural network (CNN) that is dependent on a generic video camera. The utilization of transfer learning with the recently obtained data leads to a notable enhancement in the recognition accuracy. The framework that was proposed demonstrated recognition rates of 97.7%, 95.4%, and 93.7% in person independent evaluations conducted on English, Bengali, and Devanagari numerals, respectively.

Additionally, a novel approach is proposed for character recognition of air-written text through the utilization of a 2D-CNN model for seven datasets [27]. In recent years, there has been significant progress in the field of handwritten character recognition, largely attributed to the utilization of deep convolutional neural networks (CNNs). These advanced neural network architectures have demonstrated remarkable accomplishments, surpassing established benchmarks by substantial margins. Currently, there is a lack of a standardized dataset in the field of air-writing research, which hinders the ability to assess the effectiveness of various methodologies. In recent times, a number of datasets have been made available for the purpose of recognizing air-writing characters.

The study [25], was performed using four different datasets (WISDM, HARDS, SBRHA, PAMAP2) with various characteristics, as previously used by similar recognition research. However, this includes the investigation to test the selected features with a variety of sensor placement locations, such as waist, wrist, chest, and ankle, sampling frequency, and the performance of the dataset. The purpose of [20] is to investigate the impact of window lengths with orientation invariant heuristic features on the performance of 1D-CNN-LSTM using data from 42 participants to recognize six human activities: sitting, lying down, walking, and running at three different speeds using information gathered from the Samsung Galaxy s7 smartphone's accelerometer sensor with the pre-installed Ethica application.

There are multiple benefits associated with the utilization of one-dimensional convolutional neural networks (1D CNNs) [28]. These advantages include notable performance on datasets with limited amounts of data, reduced computational complexity in comparison to two-dimensional CNNs and other deep learning architectures, expedited training processes, and a strong capability to extract pertinent features from sequential data and time series, such as signal data [26], [36].

Attributes capable of capturing patterns over a sliding window are formulated in the feature extraction phase. Time-domain features, frequency-domain features, and time-frequency-domain features are the three broad categories established by applying elementary statistical techniques to the features.

The impact of the one-dimensional convolutional neural network (1D CNN) technique on the field of air-writing recognition has been substantial. Numerous studies have been conducted to investigate the application of deep convolutional neural networks (CNNs) in the field of air-writing recognition, and these investigations have consistently proved the efficacy of such networks [20], [29].

Additionally, CNNs can be employed to analyze 1D signals, including electrocardiograms (ECG), electroencephalograms (EEG), and electromyograms (EMG), in healthcare applications. Lastly, CNNs can be utilized for machine fault detection purposes, among other applications [23]. The use of 1D CNN algorithms in sensor-based air-writing has made significant contributions to the field. These algorithms have been employed to recognize and interpret air-written characters and gestures accurately. The researchers investigated air-writing recognition using a 2D CNN model [32]. They extensively studied different interpolation techniques on publicly available air-writing datasets and developed a method to recognize air-

written characters. Their findings highlighted the importance of choosing the proper interpolation technique for accurate recognition.

Few researchers have developed a simple yet effective air-writing recognition approach based on deep CNNs [9]. Their method utilized a 1D CNN architecture to recognize air-written characters. The results showed that the suggested method was very good at reading handwritten characters, showing that 1D CNNs could be useful in this area. The authors proposed an approach for air-writing recognition based on deep convolutional neural networks (CNNs).

Furthermore, [21] developed a wearable IMU-based human activity recognition algorithm for clinical balance assessment using a 1D-CNN and GRU ensemble model. Although this study focused on human activity recognition, the use of a 1D-CNN model demonstrates the potential of this algorithm in capturing and analyzing sensor data for various applications, including air-writing.

Furthermore, the use of 1D CNNs has been explored in other domains as well. applied a 1D CNN algorithm for the detection of water pH using visible near-infrared spectroscopy [28]. They interpreted the learning mechanism of the 1D CNN through visual feature maps generated by the convolutional layers. This demonstrates the versatility of 1D CNNs in various applications, including air-writing recognition.

In the field of handwriting analysis and recognition, air-writing has also been explored. Researchers have highlighted the importance of analyzing handwriting not only on paper but also in the air before the pen touches the paper. Significant differences have been observed between these two writing conditions [37]. Additionally, studies have focused on developing recognition systems for air-writing using deep learning and trajectory-based approaches. These systems utilize techniques such as deep neural networks and depth sensors to accurately recognize and track air-written characters [32], [11].

In a nutshell, the use of 1D CNN algorithms has made significant contributions to air-writing recognition. The studies mentioned above have demonstrated the effectiveness of deep CNNs in accurately recognizing air-written characters. The utilization of one-dimensional convolutional neural networks (1D CNNs) in the analysis and interpretation of intricate patterns in air-writing movements has facilitated progress and innovation in this particular domain.

Furthermore, the detection and tracking of fingertip movements in air-writing are crucial for accurate recognition. In [31], developed a fingertip detection and tracking algorithm specifically for air-writing recognition. This algorithm outperformed state-of-the-art approaches, achieving a mean precision of 73.1

Sensor-based air-writing offers several advantages for Tamil language learners and users. It provides a more interactive and immersive learning experience, allowing learners to practice writing Tamil alphabets without the need for physical writing materials. This can be particularly beneficial for learners who may have limited access to writing resources or prefer a more hands-on approach to learning. Additionally, sensor-based air-writing systems can provide real-time feedback and evaluation, helping learners improve their writing skills and accuracy.

In conclusion, sensor-based air-writing methods like reading handwritten words from a digital writing pad and finding and following fingertip movements are very important for the Tamil language alphabets. These techniques enable accurate recognition and analysis of Tamil alphabets written in the air, offering valuable tools for language learning, practice, and evaluation.

Our study differs from the existing studies in multiple aspects. Most of the prior research only used statistical equations to extract small subsets of features from the time domain, frequency domain, and time-frequency domain. Some of these studies also relied on publicly available datasets that were used for the purpose of the research. In contrast, this paper utilizes 1D-CNN to extract a complete set of features from five sensors' raw data. This investigation follows a multi-stage process and employs state-of-the-art techniques. In addition, unlike previous work, which relied on publicly available datasets, this study used data on Tamil alphabet recognition that was collected independently.

III. Methodology

A flowchart for the proposed air-writing recognition system is shown in Fig. 1. An approach is employed in our research for the purpose of 6-DOF character recognition.

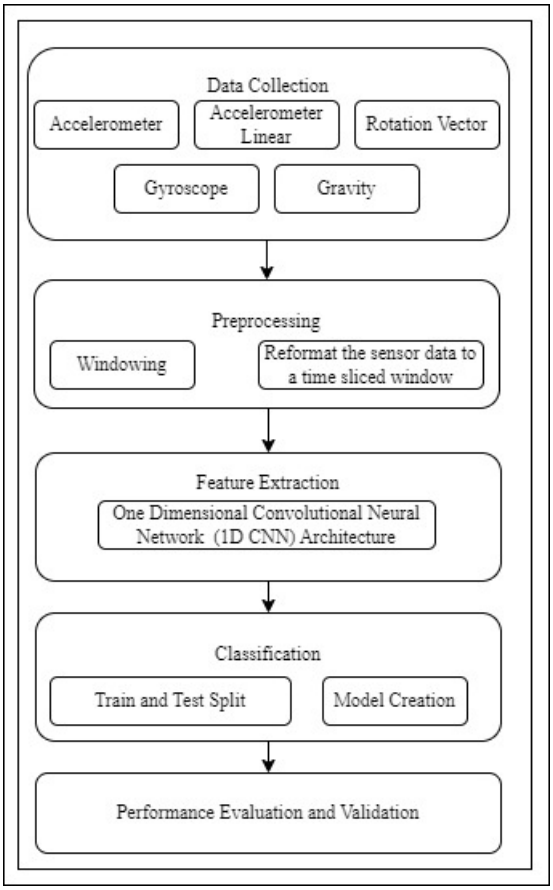


Fig. 1. The research flow.

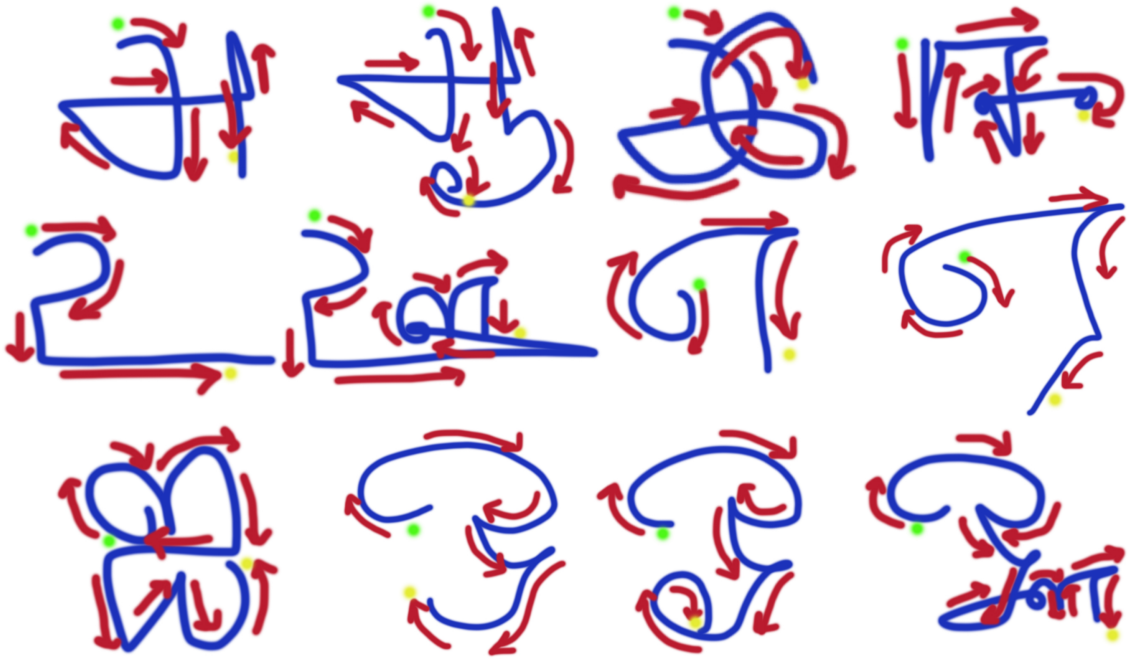


Fig. 2. The isolated characters of tamil vowel alphabets are written using a single continuous stroke, where the paths of the stroke are shown with arrows. The starting and ending points are denoted with green and yellow color dots, respectively.

A. Writing with Six Degree of Freedom (DOF)

Writing in the air differs significantly from writing by hand on paper or another surface. The writing is naturally reproduced in the air in uni-stroke without any pen-up or pen-down information, instead writing in a box in the imagined space without tactile input. Air-writing is tracked with a continuous stream of sensor data.To our best knowledge, we are the first to evaluate the rounded alphabet, i.e., Tamil language vowel recognition with axis-independent motions.

A total of five participants, all of whom were right-handed, were selected for the purpose of collecting air-writing data. Participants were instructed to maintain consistency in their writing style, including the use of a standardized writing scale, stroke order for each letter, and speed of writing. Each participant was given the opportunity to engage in a practice session before data collection commenced. Each instance of isolated motion characters was meticulously documented on different occasions by each participant.

The tracking system utilizes a smartphone device, which is monitored through an independent application. The application utilizes the input features to effectively acquire and record the sensor data associated with the task of writing. Subsequently, participants were provided with detailed instructions regarding the functionality and operation of the mobile application, as well as a comprehensive overview of the data collection procedure. Each participant carried an OPPO A12 smartphone attached to their right hand with the pre-installed application that can be recorded with five sensor data values with an interval of 10 ms. In order to mitigate the inclusion of extraneous movements, participants were instructed to execute the software with their non-dominant hand. Each letter was recorded separately. This was done in order to give the subjects the option to rest between recordings. and even change positions. The dataset pertaining to air-writing encompasses a comprehensive collection of 600 distinct samples for 12 vowels.

In order to enhance the machine's ability to recognize and expedite the user's writing process, the letters are subjected to a simplification process. Isolated characters are written using a single continuous stroke. The path of uni-stroke writing for different letters of Tamil vowel letters is illustrated in Fig. 2.

B. Preprocessing

Each individual sample, which corresponds to a single air-written letter, is regarded as a collection of distinct signals (time series). Based on the sensor data, it is possible to represent a writing motion as a spatio-temporal pattern. An analysis of the sensor values of the aforementioned sensors was performed in the spatio-temporal domain, wherein the relationship between sensor data is illustrated. The visual representation of the sensor data for the 12th vowel ('oow') is illustrated in Fig. 3. It serves as a tool for researchers to analyze and interpret patterns in the data collected.

C. Reformat Sensor Data into a Time-sliced Representation and Data Sampling

The variability in the duration of each signal's recording arises from the inherent characteristics of the writing and recording procedures. In accordance with the experimental design, a predetermined length, denoted as "l," has been established to ensure uniformity in the length of all signals

across all samples. Achieving a suitable time window is a challenging task. Each window is considered non-overlapping as the characters are isolated. The raw data from each of the windows is considered an occurrence.

The sensor data from five sensors is accumulated into a single CSV format. We have tried truncating to make the signals consistent in length. This study selected an arbitrary window length as the intermediate time taken to write each character.

We have used the programming language Python to execute the required preprocessing of the data and extract the features, respectively. We have used packages named Pandas and Numpy for performing feature extraction.

D. The Motivation for using CNN

Significant advancements have been made in recent years in the domains of sensors, smart sensors, artificial intelligence, and their integration. In the field of machine learning, there is a current trend towards the utilization of Artificial Neural Networks (ANN), Deep Neural Networks (DNN) which are the prevailing techniques in the deep learning domain, and Convolutional Neural Networks (CNN). These concepts have gained popularity, indicating an increasing scientific interest and subsequent contributions from the scientific community. [19]

CNNs, also known as convolutional neural networks, are a prominent class of deep learning models that are widely used in various domains. These models are characterized by their architecture, which typically includes multiple layers of convolutions. These layers are designed to leverage a collection of adaptable multi-dimensional filters. These filters are systematically moved across the various axes of the input sample [35].

In various domains, including industrial, clinical, and environmental sectors, it is frequently observed that one-dimensional (1D) data is prevalent. Examples of such sensor data include electrocardiogram readings, temperature measurements, environment-control variables, motion data, and power consumption data. One potential approach for modeling this information is the utilization of one-dimensional convolutional neural networks (1D CNN) [28], [29], [36].

In the context of a one-dimensional convolutional neural network (1D CNN), the filters are applied by sliding them along a single axis of the input data. In the present scenario, it is observed that all dimensions of the filter's size, with the exception of one, are predetermined to align with the sizes of the fixed axes. In recent times, their utilization has extended to the realm of Temporal Sequence Recognition and Time Series Classification as well, as evidenced by the growing interest in the area. Temporal Sequence Classification refers to the task of accurately categorizing sequences of data that are captured by inertial measurement sensors. The objective is to assign these sequences to specific, pre-defined activities that occur over extended periods of time [26], [36]. This classification process takes place within a continuous stream of data, requiring robust and reliable algorithms to achieve accurate results.

In the present investigation, the decision was made to employ Convolutional Neural Networks (CNNs) due to two primary justifications. Traditional machine learning methods typically depend on the extraction of explicit features. In contrast, it is worth nothing that deep neural networks, particularly convolutional neural networks (CNNs), possess the ability to directly consume the raw input data without the need for explicit feature extraction procedures. Hence, it is plausible to hypothesize that a methodology that has demonstrated efficacy in the domain of temporal data analysis and computer vision holds promise for achieving favorable outcomes in our specific context.

Next, the set of preprocessed samples is introduced into the architecture, and a mini-batch learning process is implemented. The architectural design of our one-dimensional convolutional neural network (1D CNN) classifier draws inspiration from the model. The architectural design consists of convolutional layers that are strategically interspersed with pooling layers. In this study, it is observed that all of the convolutional layers in the experimental setup employ filters with a length of 3200 and 1600. Additionally, the Rectified Linear Units (ReLU) activation function is utilized across these layers.

Given the nature of our task, which involves classifying data into distinct vowels, we have opted to utilize categorical cross entropy as our chosen loss function. This particular loss function is well-suited for categorical classification tasks.

In the context of sensor technology, irrespective of the specific domain of application, it is frequently observed that data pertaining to the one-dimensional (1D) shape is prevalent [20], [29].

E. 1D CNN Architecture for Creating Feature Vector

The conventional architecture of a Convolutional Neural Network (CNN) consists of three primary layers, namely the convolutional layer, pooling layer, and fully connected layer. The layers in question employ various components and techniques, including convolutions, activation functions, pooling, dropout, batch normalization, and fully connected blocks, among others. These elements can be flexibly combined in numerous configurations.

A 1D CNN (Convolutional Neural Network) can be constructed based on the following considerations:

- The input data is formatted in one dimension. The data for this study was obtained from five sensors installed on a sliced window.
- The convolutional layers are tasked with performing feature extraction operations. The extraction process involves the application of convolution operations to the input data, with the resulting convolutions being passed as input to the subsequent layer. The convolutional operations are determined by a set of filters, a specified kernel size, padding, and stride. These operations result in the creation of a feature map, which is obtained by applying a ReLU activation function.
- Pooling layers are typically applied following a convolutional layer, serving the purpose of retaining the information produced by the feature maps. The techniques on these layers include max pooling.

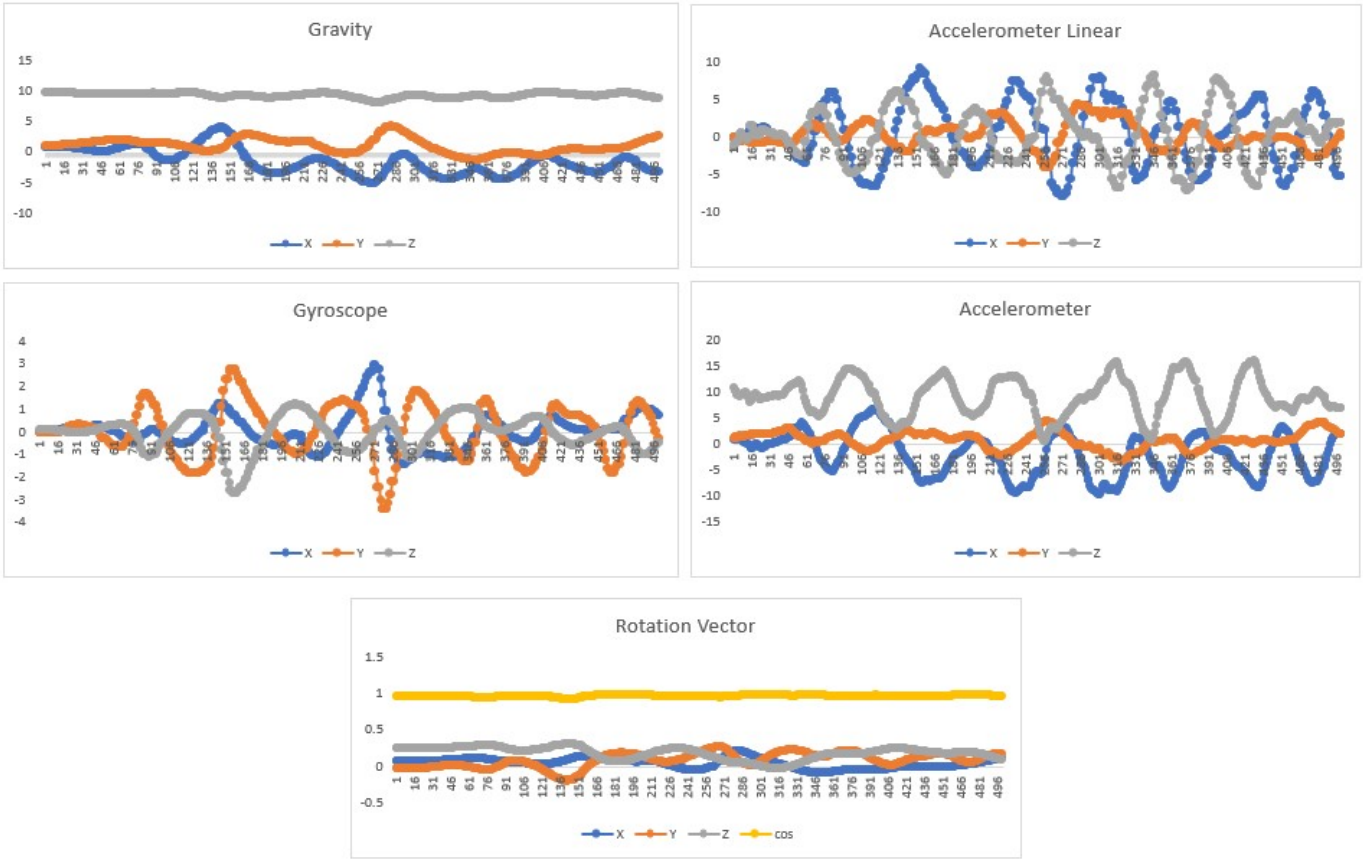


Fig. 3. The variation of the sensor data from five different sensors for the last letter of Tamil language vowel alphabet.

Model: "sequential_5"

Layer (type)	Output Shape	Param #
conv1d_15 (Conv1D)	(None, 3200, 32)	128
conv1d_16 (Conv1D)	(None, 3200, 64)	6208
conv1d_17 (Conv1D)	(None, 3200, 128)	24704
max_pooling1d_5 (MaxPooling1D)	(None, 1600, 128)	0
dropout_5 (Dropout)	(None, 1600, 128)	0
flatten_5 (Flatten)	(None, 204800)	0
dense_15 (Dense)	(None, 256)	52429056
dense_16 (Dense)	(None, 512)	131584
dense_17 (Dense)	(None, 128)	65664

Total params: 52657344 (200.87 MB)
Trainable params: 52657344 (200.87 MB)
Non-trainable params: 0 (0.00 Byte)

Fig. 4. The 1D CNN architecture.

The term "dropout" refers to an individual who leaves an educational institution before completing their program of study. The dropout technique is employed as a means of mitigating overfitting, and it can be implemented in both fully connected and convolutional layers. The technique involves randomly disabling certain connections within the neural network. This process ensures that individual neurons are able to be removed, thereby preventing them from exerting excessive influence on the model's output. Alternatively, the technique enhances the generalizability of the model.

The application of the corresponding experiments, which have resulted in the 1D CNN architecture for the problem, is presented in Fig. 4. The input layer will consist of 300 inputs, which correspond to the sensor data for Tamil vowel characters.

F. Population of the Dataset

In order to facilitate analysis, it is necessary to extract features from the signal windows, as the raw data itself is not suitable for direct analysis. This process enables the identification and capture of patterns within the data. Previous studies in the field of recognition have utilized both time domain and feature domain features in their analysis. 1D CNN uses filters on each window to extract features automatically. 1D CNN maps these internally extracted features to different Tamil vowel alphabets. The extracted features are tabulated with their vowel labels using the label encoding technique.

However, when the 1D CNN feature vector is combined with the machine learning algorithms, the extracted features from the 1D CNN are used as inputs for the selected machine learning algorithms.

G. Split up the Data Set into Train and Test Sets

To facilitate the training process of our network, we initially partition the dataset into two distinct subsets: a training set and a testing set. The division is performed in such a way that the training set comprises 80% of the data, while the remaining 20% is allocated to the testing set.

H. Train the Dataset with Machine Learning Algorithms for Tamil Air-writing Alphabet Recognition

In this modeling and dataset training stage, we develop a model for classifying the features into the Tamil language vowel alphabet, we utilize various classifiers of machine learning. The use of Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT) and k-Nearest Neighbors (KNN) model for classification. These classification-based algorithms are based on performance metrics that include accuracy, precision, recall, and f1 score.

It is worth noting that our dataset is balanced, where each vowel is represented equally, ensuring a fair and unbiased evaluation of our model's performance. Therefore, we have selected accuracy as our performance measure, which will allow us to assess the effectiveness of our model in correctly classifying the data.

I. Validate the Performance of the Trained against the Test Data using Cross-validation and Confusion Matrix

In the conducted experiments, the relevant algorithm metrics that were evaluated included the following measures: The accuracy of the model is being evaluated with 80:20 split dataset. In the present scenario, the research pair has designated classification accuracy as the appropriate performance metric. The accuracy was assessed by measuring the performance in four different splits that were done on the original dataset. The datasets were designated as training, and testing sets, with a split of 80%, 20%, 70%, 30%, 60%, 40% and 50%, 50% from the original dataset, respectively. The objective was to develop a model that effectively generalizes learning from data. Consequently, a solution that exhibited consistent and homogeneous accuracy values was deemed favorable, rather than solely prioritizing the highest accuracy values within individual sub-datasets.

In summary, the experiments assessed the impact of the variant on the accuracy of the data sets with the parameter tuning of the selected model.

IV. Results and Discussions

A. Preprocessing of Sensor Data

The sensor data from five sensors is analyzed with regard to the variation in the data values. However, each of the letters lapses at different time intervals. We have chosen 2s as the preliminary window for the time-sliced segmentation, as shown in Fig. 5. As many of the vowel letters end up in the long lines or curve lines as shown in Fig. 2, the additional time can be ignored.

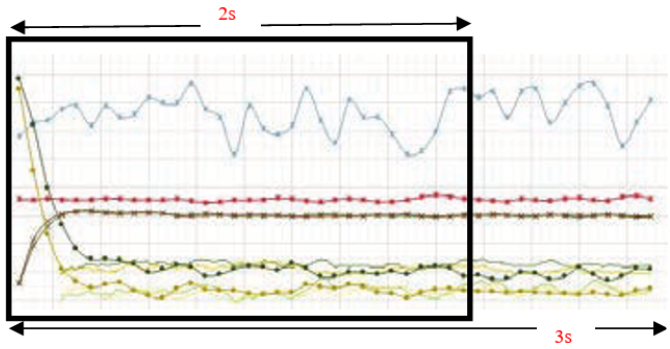


Fig. 5. The windowing process: the sensor data is sliced into the time window of 2s.

B. Feature Engineering Through 1D CNN

Each of the sliced windows of the sensor data was passed into the above-mentioned 1D CNN to create feature vectors. A total of 128 features were manipulated in the feature extraction stage, as illustrated in Fig. 4. Each of the feature rows in the feature matrix is labeled using a label encoder. The full feature set for the classification was populated in this manner.

C. Training and Testing of Dataset using Machine Learning Algorithms

As the initial step in this classification stage, the split training dataset is used for training, while the testing data is used for testing. The classification model was fitted with the training dataset. We have used five classification algorithms, such as Random Forest, Support Vector Machine, Decision Tree, Naive Bayes, and K-Nearest Neighbour. The analysis was first initiated with the 80:20 dataset split. It was found that we achieved the highest accuracy of 90.83% from the KNN algorithm. Fig. 6 shows the accuracy of the model in both testing and training. The accuracy for the training seemed to be 100% for RF, KNN and DT. However, in terms of testing accuracy, only KNN outperforms the others. We performed tests with numerous configurations. The first

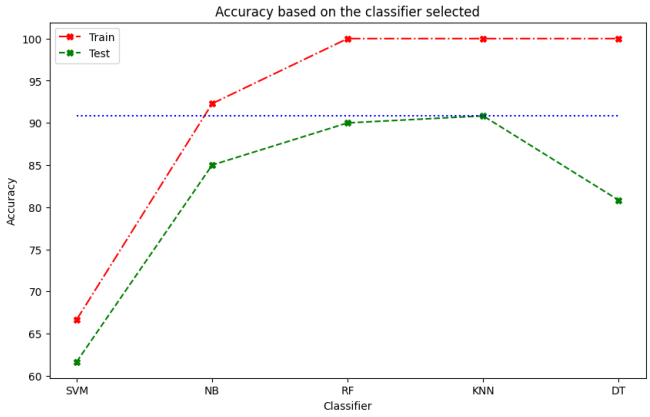


Fig. 6. The accuracy of the models (SVM,RF, NB, DT and KNN) during training and testing at 80:20 dataset split.

performance evaluation of the model is further extended using the accuracy for the different values of k. It was observed with

the variation of the k value, ranging from 1 to 30. The accuracy of the model reduces with an increase in the k value. However, there are some fluctuations at certain places where a small rise prevails in the fall. Fig. 7 clearly illustrates the variation in accuracy with the k value. The accuracy for classifying the air-written characters with k = 1 and k = 2 depicts the maximum accuracy of 90.83%. This means that with k = 1 or k = 2, the classification is more accurate compared with the other k values.

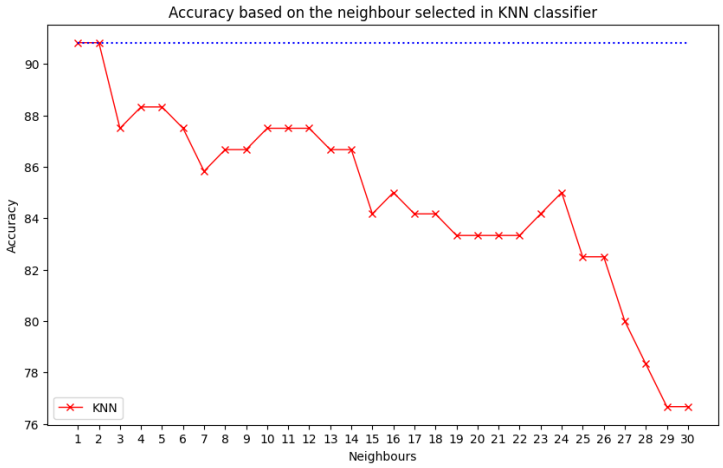


Fig. 7. The variation in accuracy with the change in k values.

Next configurations presented with the split percentage of 0.8, 0.7, 0.6 and 0.5 as training data and 0.2, 0.3, 0.4 and 0.5 as testing data, which had been taken into account for the evaluation of metrics. In this evaluation the k value is taken as 1, as it was obtained to be the highest accuracy. In Fig. 8, to classify the alphabets as per the samples used for training and testing, it has been observed that 0.6:0.4 achieves the maximum accuracy of 91.67% with the value k = 1, which shows the frequency of samples that are correctly classified to a specific air-written character within all the samples that are to be tested. Though the accuracy with the percentage of samples for 0.8, 0.7, and 0.6 increases, there is a sudden fall at the percentage of 0.5 of training data.

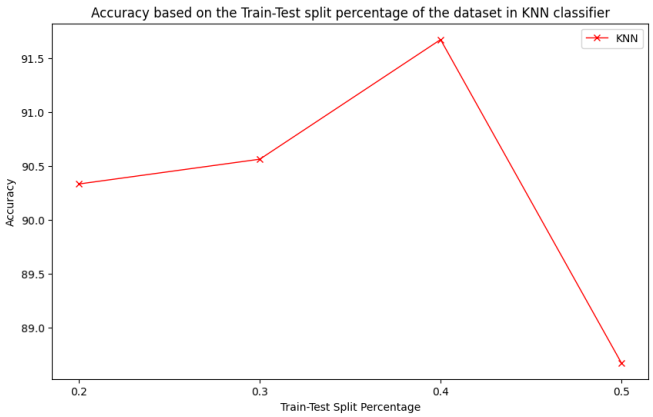


Fig. 8. The variation in accuracy with the change in the split percentage for KNN model, when k=1.

To provide the significance of the proposed model, its performance is being compared with multiple configurations. In this regard, the approaches are investigated that have been discussed previously.The confusion matrix of the model for the percentage samples of 0.6:0.4 and k=1 of KNN algorithm is shown in Fig. 9.

The accuracy, precision, recall, and F1-score of the classes of air-written characters are shown in Fig. 10. It was observed that the 3rd, 4th, 5th, 7th, and 10th characters can be easily distinguished from the others. The 6th, 11th, and 12th letters resembling the other letters show a lower accuracy rate. The evaluation has been performed with respect to the metrics used for the comparison.

The following are the findings of the evaluation:

- KNN outperforms the other classification algorithm.
- The value for k is chosen as 1 while varying from 1 to 30.
- The best test and train split is 0.6:0.4 for maximum accuracy.

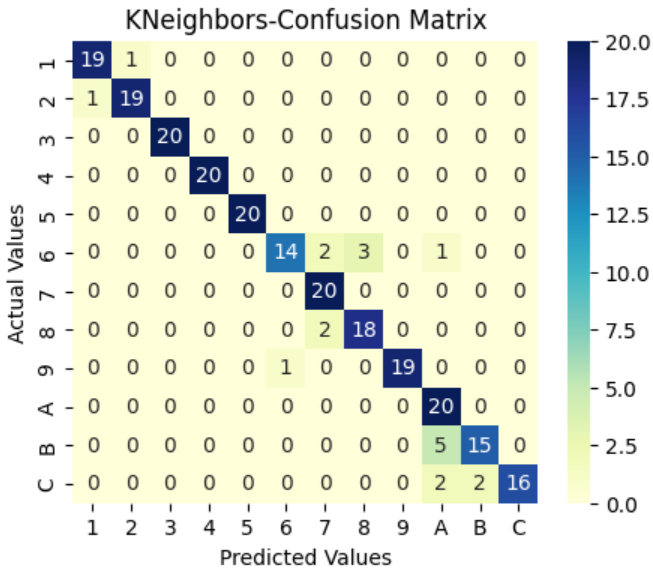


Fig. 9. The confusion matrix for KNN model when k = 1, with the 60:40 split percentage.

As this study has fully focused on the classification of the Tamil air-writing characters as of now, this is the first paper dealing with this objective.

D. Validation of the Performance

The validation of the model was tested with GridSearchCV with 10 folds for the above-identified configuration, and it was achieved with 93.33%. This study is summarized with the achievement of this accuracy in the initial phase of identifying the Tamil language air-written characters of vowel letters. We have achieved a considerable amount of achievement in the context of air-writing.

The process of identifying air-written Tamil characters poses numerous obstacles as a result of the distinct attributes

	precision	recall	f1-score	support
0	0.950000	0.950000	0.950000	20.000000
1	0.950000	0.950000	0.950000	20.000000
2	1.000000	1.000000	1.000000	20.000000
3	1.000000	1.000000	1.000000	20.000000
4	1.000000	1.000000	1.000000	20.000000
5	0.933333	0.700000	0.800000	20.000000
6	0.833333	1.000000	0.909091	20.000000
7	0.857143	0.900000	0.878049	20.000000
8	1.000000	0.950000	0.974359	20.000000
9	0.714286	1.000000	0.833333	20.000000
10	0.882353	0.750000	0.810811	20.000000
11	1.000000	0.800000	0.888889	20.000000
accuracy	0.916667	0.916667	0.916667	0.916667

Fig. 10. The performance evaluation report for KNN model when k = 1, with the 60:40 split percentage.

exhibited in handwritten Tamil, including variances in dimensions, styles, and angles of orientation. Moreover, the task of recognizing air-written characters presents intrinsic challenges due to the need to capture three-dimensional trajectories and employ suitable approaches for precise detection. This difficulty arises from the fact that many letters exhibit various features such as holes, loops, and curves. The study was conducted using a restricted dataset due to the unavailability of publicly accessible datasets. This work has the potential for further expansion through the comprehensive identification of all Tamil alphabetical characters. Additionally, enhancing the accuracy can be achieved by increasing the dataset and modifying the window size.

V. Conclusions

This paper focuses on developing a classification framework that employs a selection of features to assist us in recognizing Tamil vowel characters. A total of 12 different letters were used with the data collection and classification procedures. We have used a methodology for analyzing multiple configurations for evaluation. Initially, 1D CNN is used for feature extraction from the 2s of time window data segments. An evaluation is based on the results of five machine-learning methods: Naive Bayes, Random Forest, K-Nearest Neighbor, Support Vector Machine, and Decision Tree. The experimental results show that the proposed KNN achieves high accuracy with k = 1 and 0.6:0.4 train test split percentage with 91.67% accuracy. However, in validation using GridSearchCV this model has achieved an accuracy of 93.33% with 10 folds. Our findings were based on applications utilized in prior studies; however, as a novelty, the system provided employs collected sensor data to classify the characters by integrating the framework with cutting-edge technology. It is possible to increase the accuracy performance by examining the window size.

References

[1] Yanay, Tomer, and Erez Shmueli. "Air-writing recognition using smartbands." Pervasive and Mobile Computing 66, 2020

[2] Agrawal, S., Constandache, I., Gaonkar, S., Roy Choudhury, R., Caves, K., & DeRuyter, F. Using mobile phones to write in air. Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services - MobiSys '11. doi:10.1145/1999995.1999998, 2011.

[3] Amma, C., Georgi, M., & Schultz, T. Airwriting: Hands-Free Mobile Text Input by Spotting and Continuous Recognition of 3d-Space Handwriting with Inertial Sensors. 2012 16th International Symposium on Wearable Computers. doi:10.1109/iswc.2012.21, 2012.

[4] Amma, C., Georgi, M., & Schultz, T. Airwriting: a wearable handwriting recognition system. Personal and Ubiquitous Computing, 18(1), 191–203. doi:10.1007/s00779-013-0637-3, 2013.

[5] Chen, Mingyu, Ghassan AlRegib, and Biing-Hwang Juang. "Air-writing recognition—Part I: Modeling and recognition of characters, words, and connecting motions." IEEE Transactions on Human-Machine Systems 46, no. 3: 403-413, 2015.

[6] "World Languages". [Online]. Available: https://www.mustgo.com/worldlanguages/tamil/ [Accessed: Oct. 3, 2023].

[7] "Tamil Script". [Online]. Available:https://en.wiktionary.org/wiki /Appendix:Tamil_script [Accessed: Oct. 3, 2023].

[8] Jayanthi, V. and Thenmalar, S. Recognition of handwritten words from digital writing pad using mmu-snet. Intelligent Automation & Soft Computing, 36(3), 3551-3564. https://doi.org/10.32604/iasc.2023.036599, 2023.

[9] Hsieh, C., Lo, Y., Chen, J., & Tang, S. Air-writing recognition based on deep convolutional neural networks. Ieee Access, 9, 142827-142836. https://doi.org/10.1109/access.2021.3121093, 2021.

[10] Xia, Q., Hong, F., Feng, Y., & Guo, Z. MotionHacker: Motion sensor based eavesdropping on handwriting via smartwatch. IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). doi:10.1109/infcomw.2018.8406879, 2018.

[11] Alam, M., Alam, M., Abbass, M., Imtiaz, S., & Kim, N. Trajectory-based air-writing recognition using deep neural network and depth sensor. Sensors, 20(2), 376. https://doi.org/10.3390/s20020376, 2020.

[12] Chu, Yen-Cheng, Yun-Jie Jhang, Tsung-Ming Tai, and Wen-Jyi Hwang. "Recognition of hand gesture sequences by accelerometers and gyroscopes." Applied Sciences 10, no. 18: 6507, 2020.

[13] Wang, J.-S., & Chuang, F.-C. An Accelerometer-Based Digital Pen With a Trajectory Recognition Algorithm for Handwritten Digit and Gesture Recognition. IEEE Transactions on Industrial Electronics, 59(7), 2998–3007. doi:10.1109/tie.2011.2167895, 2012.

[14] Chen, Y., Yin, P., Peng, Z., Lin, Q., Zhao, Z., Fan, Q., ... & Wei, Z. High-throughput recognition of tumor cells using label-free elemental characteristics based on interpretable deep learning. Analytical Chemistry, 94(7), 3158-3164. https://doi.org/10.1021/acs.analchem.1c04553, 2022.

[15] Lopez-Rodriguez, P.; Avina-Cervantes, J.G.; Contreras-Hernandez, J.L.; Correa, R.; Ruiz-Pinales, J. Handwriting Recognition Based on 3D Accelerometer Data by Deep Learning. Appl. Sci. 12, 6707. https://doi.org/10.3390/app12136707, 2022.

[16] Aakash, C., Kumar, P., Pilla, F., Skouloudis, A., Sabatino, S., Ratti, C., & Rickerby, D. End-user perspective of low-cost sensors for outdoor air pollution monitoring. The Science of the Total Environment, 607-608, 691-705. https://doi.org/10.1016/j.scitotenv.2017.06.266, 2017.

[17] Bastas, Grigoris, Kosmas Kritsis, and Vassilis Katsouros. "Air-writing recognition using deep convolutional and recurrent neural network architectures." 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 7-12. IEEE, 2020.

[18] "Human Activity Recognition". [Online]. Available: https://towardsdatascience.com/human-activity-recognition-har-tutorial-with-keras-and-core-ml-part-1-8c05e365dfa0 [Accessed: Oct. 3, 2023].

[19] Rala Cordeiro, J.; Raimundo, A.; Postolache, O.; Sebastião, P. Neural Architecture Search for 1D CNNs—Different Approaches Tests and Measurements. Sensors 2021, 21, 7990. https://doi.org/10.3390/s21237990, 2021.

[20] Barua, A.; Fuller, D.; Musa, S.; Jiang, X. Exploring Orientation Invariant Heuristic Features with Variant Window Length of 1D-CNN-LSTM in Human ActivityRecognition. Biosensors 2022, 12, 549. https://doi.org/10.3390/bios12070549, 2022.

[21] Kim, Y., Kyunglim, J., Jeong, H., & Lee, S. Wearable imu-based human activity recognition algorithm for clinical balance assessment using 1d-cnn and gru ensemble model. Sensors, 21(22), 7628, 2021.

[22] "Time Series Classification with Deep Learning". [Online]. Available:https://towardsdatascience.com/time-series-classification-with-deep-learning-d238f0147d6f [Accessed: Oct. 3, 2023].

[23] Khan, A., Sohail, A., Zahoora, U. et al. A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 53, 5455–5516, 2020. https://doi.org/10.1007/s10462-020-09825-6

[24] "Time Series Classification". [Online]. Available:https://towardsdatascience.com/how-to-use-convolutional-neural-networks-for-time-series-classification-56b1b0a07a57 [Accessed: Oct. 3, 2023].

[25] Bennasar, M.; Price, B.A.; Gooch, D.; Bandara, A.K.; Nuseibeh, B. Significant Features for Human Activity Recognition Using Tri-Axial Accelerometers. Sensors2022, 22, 7482. https://doi.org/10.3390/s22197482, 2022.

[26] Osman, Radwa Ahmed, Sherine Nagy Saleh, and Yasmine N. M. Saleh. "A Novel Interference Avoidance Based on a Distributed Deep Learning Model for 5G-Enabled IoT" Sensors 21, no. 19: 6555. https://doi.org/10.3390/s21196555, 2021

[27] Abir, F.A.; Siam, M.A.; Sayeed, A.; Hasan, M.A.M.; Shin, J. Deep Learning Based Air-Writing Recognition with the Choice of Proper Interpolation Technique. Sensors 2021, 21, 8407. https://doi.org/10.3390/s21248407, 2021.

[28] Li, D. and Li, L. Detection of water ph using visible near-infrared spectroscopy and one-dimensional convolutional neural network. Sensors, 22(15), 5809. https://doi.org/10.3390/s22155809, 2022.

[29] "Introduction to 1D convolutional neural networks in keras for time sequences". [Online]. Available:https://blog.goodaudience.com/introduction-to-1d-convolutional-neural-networks-in-keras-for-time-sequences-3a7ff801a2cf [Accessed: Oct. 3, 2023].

[30] Ismail Fawaz, Hassan, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. "Deep learning for time series classification: a review." Data mining and knowledge discovery 33, no. 4: 917-963, 2019.

[31] Mukherjee, Sohom, Sk Arif Ahmed, Debi Prosad Dogra, Samarjit Kar, and Partha Pratim Roy. "Fingertip detection and tracking for recognition of air-writing in videos." Expert Systems with Applications 136: 217-229, 2019.

[32] Abir, F., Siam, M., Sayeed, A., Hasan, M., & Shin, J. Deep learning based air-writing recognition with the choice of proper interpolation technique. Sensors, 21(24), 8407. https://doi.org/10.3390/s21248407, 2021.

[33] Mori, Y., Hasegawa, A., & Mori, J. The trends and developments of l2 japanese research in the 2010s. Language Teaching, 54(1), 90-127. https://doi.org/10.1017/s0261444820000336, 2020.

[34] Thomas, M. A role for "air writing" in second-language learners' acquisition of japanese in the age of the word processor. Journal of Japanese Linguistics, 30(1), 86-106. https://doi.org/10.1515/jjl-2014-0107, 2014.

[35] Roy, Prasun, Subhankar Ghosh, and Umapada Pal. "A CNN based framework for unistroke numeral recognition in air-writing." 16th international conference on frontiers in handwriting recognition (ICFHR), pp. 404-409. IEEE, 2018.

[36] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, Daniel J. Inman, 1D convolutional neural networks and applications: A survey, Mechanical Systems and Signal Processing, Volume 151, 107398, ISSN 0888-3270, https://doi.org/10.1016/j.ymssp.2020.107398, 2021.

[37] Diaz, M., Ferrer, M., Impedovo, D., Malik, M., Pirlo, G., & Plamondon, R. A perspective analysis of handwritten signature technology. Acm Computing Surveys, 51(6), 1-39. https://doi.org/10.1145/3274658, 2019.

Emotional Speech Transfer on Demand Based on Contextual Information and Generative Models: A Case Study

Andrea Veronica Porco
Dept. Information Engineering
University of the Ryukyus
Nishihara, Japan

Kang Dongshik
Dept. Information Engineering
University of the Ryukyus
Nishihara, Japan

***Abstract*—The automated generation of speech audio that closely resembles human emotional speech has garnered significant attention from the society and the engineering academia. This attention is due to its diverse applications, including audiobooks, podcasts, and the development of empathetic home assistants. In the scope of this study, it is introduced a novel approach to emotional speech transfer utilizing generative models and a selected emotional target desired for the output speech. The natural speech has been extended with contextual information data related with emotional speech cues. The generative models used for pursuing this task are a variational autoencoder model and a conditional generative adversarial network model. In this case study, an input voice audio, a desired utterance, and user-selected emotional cues, are used to produce emotionally expressive speech audio, transferring an ordinary speech audio with added contextual cues, into a happy emotional speech audio by a variational autoencoder model. The model try to reproduce in the ordinary speech, the emotion present in the emotional contextual cues used for training. The results show that, the proposed unsupervised VAE model with custom dataset for generating emotional data reach an MSE lower than 0.010 and an SSIM almost reaching the 0.70, while most of the values are greater than 0.60, respect to the input data and the generated data. CGAN and VAE models when generating new emotional data on demand, show a certain degree of success in the evaluation of an emotion classifier that determines the similarity with real emotional audios.**

***Keywords*—Emotion transfer; contextual information; speech processing; generative models; variational autoencoder; conditional generative adversarial networks; empathetic systems**

I. Introduction

The creation of empathetic systems, capable of understanding and responding to human emotions, marks a significant advancement in artificial intelligence. Empathetic systems hold the promise of transforming human-machine interactions, offering not just responses but genuine understanding. However, this task presents a profound challenge. During speech processing, the subtle nuances of human emotions often dissipate, making it a complex endeavor to imbue artificial intelligence with empathy. Preserving these emotional characteristics during speech processing remains an open issue in the field of AI.

It is expected in the near future to have home assistants capable of not only recognizing when people surrounded are feeling sad, happy, or anxious but also responding with appropriate empathy [1]–[3]. Such a system could offer invaluable support especially to the vulnerable population, providing comfort to the lonely, reassurance to the anxious, joy to the despondent and safety to children and seniors. The potential applications are vast, extending beyond homes to healthcare, customer service, and mental health support. Achieving this level of artificial empathy stands at the frontier of AI research, requiring innovative solutions to bridge the gap between raw data and the rich emotional tapestry of human speech.

There are several models proposed to achieve this target, including generative and non generative models.

Generative models have emerged as valuable tools for the synthesis of speech audios. These models offer the unique ability to create audio data that captures the nuanced patterns and complexities of human speech. Variational Auto-Encoders (VAEs) [4], [5] for instance, provide a structured approach to encoding and decoding data, allowing for the generation of diverse, high-quality audio samples. Conditional Generative Adversarial Networks (cGANs), on the other hand, introduce conditional factors, enabling the generation of speech data with specific attributes, such as different emotional states or gender-specific characteristics. These generative models excel in creating natural-sounding speech and have the potential to revolutionize applications like voice assistants, speech synthesis, and emotional speech generation.

Despite their promising capabilities, generative models also come with inherent challenges. One notable drawback is the risk of generating audio samples that, while coherent, may lack the nuanced emotional expressiveness present in natural human speech. The delicate interplay of pitch, rhythm, and intensity that defines emotional speech can be challenging to replicate accurately. Furthermore, generative models may struggle with gender-specific patterns, such as pitch variations and resonance differences between male and female voices. Additionally, ensuring the generated audio remains consistent with the intended emotional state or gender identity presents a significant challenge. These difficulties underscore the need for continued research and development in the field of generative speech modeling, particularly in the context of emotion, gender, and natural speech synthesis [6], [7].

There are several non-generative models used for speech synthesis, including Hidden Markov Models (HMMs) [8],

Long Short-Term Memory (LSTM) networks [9], and deep neural networks (DNNs) [10], [11]. While these models have proven effective in certain aspects of speech processing, they exhibit limitations that generative models, like VAEs and cGANs, can address.

One limitation of traditional non-generative models, such as HMMs and LSTMs, is their reliance on a fixed set of acoustic features or linguistic representations. These models often struggle to capture the rich nuances of natural speech, including emotional variations and gender-specific characteristics. The adaptability of non-generative models to generate highly expressive and contextually rich speech remains limited. Furthermore, these models may require extensive data preprocessing and manual feature engineering, making them less flexible and more labor-intensive in comparison to generative models.

Generative models, on the other hand, have the potential to overcome these limitations [12]–[14]. They can operate in an end-to-end fashion, learning complex patterns without the need for extensive feature engineering. By leveraging latent spaces, conditional information, and adversarial training, generative models can synthesize speech that better resembles natural human communication, including emotional variations and gender-specific traits. This adaptability and capacity to capture nuanced characteristics make generative models an attractive choice for applications where high-fidelity and emotionally expressive speech synthesis is crucial, such as the creation of diverse empathetic systems. Nonetheless, it is essential to recognize that both generative and non-generative models have their own strengths and limitations, and the choice between them depends on specific task requirements and constraints [15]–[21].

In this work, two generative models such as variational auto-encoder and conditional adversarial network were utilized, to train speech audio data, contextual audio data and an emotional selected target on demand, to generate an emotional speech audio with the target emotion. In this proposed case study, we also show how a neutral speech audio with a specific gender and a specific contextual data input (laughing by giggling, angry by shouting, crying sound, etc.) is converted into a happy speech audio automatically by the variational auto-encoder model. For the creation of the testing data, a TTS system is used by selecting a gender specific voice and an utterance close in pronunciation to the trained data.

To the best of our knowledge previous research works did not propose an emotional speech transfer on demand that train extended generative models such as VAE and CGAN with speech data and contextual related cues in gender and emotion. Furthermore, a TTS system is utilized to generate testing data for the proposed case study with a trained variational auto-encoder model and additional contextual cues.

The paper is structured as follows. Subsequent sections of the introduction section, sequentially detail the proposed approach and associated experiments, incorporating comprehensive information on data preprocessing, experimental methodologies, and results. Following this, the case study section is introduced to illustrate a practical application, featuring the integration of real Text-to-Speech (TTS) samples with the utilization of the proposed model and custom data. The following discussion section serves to expound on the evaluations and results pertaining to both models, delineating inherent limitations and identifying potential avenues for further research or enhancements. Lastly, the conclusion section encapsulates final remarks on the presented works and outlines prospective future endeavors.

II. Proposed Approach

In this proposed work, the functionality of a variational autoencoder model and a conditional adversarial network model were extended, for learning emotional patterns that have associated emotional contextual audio data, such as crying, shouting and laughing by giggling sounds. The sad emotion is associated with the crying sound, the angry emotion with the shouting sound, the happy emotion with the laughing by giggling sound and a normal emotion has a simple whisper sound of 1 second pattern.

The dataset used is an extended Ravdess dataset. The Ravdess database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent, “Kids are talking by the door” and “Dogs are sitting by the door” for speech and sing. Speech includes neutral, calm, happy, sad, angry, fearful, surprise, and disgust expressions. The selected data from Ravdess dataset is audio-only with 1 second recoding for 24 actors, equally gender balanced, with additional 24 equally balanced contextual audios that match by gender and by emotion.

The selection of contextual data associated is random, forming a total of a 2 second recording for input data. However, the model cannot distinguish this associated data, since is using it directly as a complete input data. The contextual audio data is firstly divided by gender, into female and male data. Even though the speech audio used correspond to a female or male specific voice, we did not associate the same exact voice by gender with the same pattern male/female voice. This is due to the generalisation of the voice we do have while laughing, crying or shouting and singing, where the voice is difficult to be perfectly recognised.

The generative models used benefit in different ways. The variational autoencoder model will keep training in an unsupervised manner, because after training, once we input a data with a specific associated contextual audio, and an additional gender specific voice speech input, the variational autoencoder will try to reconstruct the emotion present in similar audios. Therefore, trying to reconstruct the weak emotional speech part into a stronger emotional speech. For example, when people speech is happy, people use to laugh by giggling, and these two correlated actions we expect that our models will capture the essence and associate both action into one emotional concept. Certainly, other external noises from the context could be learnt by the models, however there will be not correlation with the emotional speech we emphasised.

The generative adversarial network in counterpart, will use the same input data but to feed the discriminator that during training passes information to the generator loss, and therefore to the generator itself. In this case, we do not expect the generator of CGAN model to be as good as the VAE encoder and decoder. This assumption is based on the CGAN model specifications, where after training, we ask the generator to

generate a happy audio but we cannot ensure which samples it will create in relation to gender and utterance given. This issue is improved when we train data by gender (male or female, but not both of them), or by specific utterance which is very restricted in terms of the usage of TTS normal emotion generated input audios. Therefore, for the case study we will show an example based on a variational auto-encoder model instead, when giving a TTS neutral audio voice as an input, setting the gender and the utterance that is expected to be reconstructed.

The VAE and CGAN models will try to transfer the emotion in a new audio file. It is important to mention the limitations we feat with this approach. The first limitation is the reconstruction of the input and the generation of the target data. The input data origin was initially separated, which causes that the output received by each model should also be treated separately at the end. This will cause the data to be more noisy and more difficult to reconstruct as an audio file.

Another important point is that these models produce noisy results with complex data, and speech data enter in that category. Times series data cannot be manipulated such as the image data because rotating, flipping, augmenting or shifting the data for images will not affect the final position or structure of the objects in an image, but it will completely corrupt our times series data. Therefore, we made a great effort while passing through these models to leave the data without manipulating it, whenever was possible during the emotional transfer process.

The details of the architecture of the proposed models are shown in Fig. 1, 2 and 3, respectively.

The input data of VAE and CGAN architectures, is a combined audio data between a 1 second speech audio from Ravdess dataset audios, with a connected 1 second associated emotional contextual information in a form of audio. Therefore, the two audios are concatenated into one audio with 2 seconds of total duration and the same sample frequency, 44100Hz. As previously mentioned, the relation between the speech data and the contextual information is the weakly emotional pattern present in Ravdess dataset, and specifically, the gender present in each speech audio data. As it was mentioned, it significantly differs from using any other unrelated noise that would not be useful in terms of our target, which is the emotional transfer. For our case study we extracted the log Mel spectrogram data and converted them into images.

Our proposed model architecture for VAE for training and testing can be observed in Fig. 1 and 2, respectively. The basic architecture of the VAE model has two associated networks, called encoder and decoder, which are connected by the latent space representing the probabilistic part of the model.

The proposed CGAN model's architecture can be observed in Fig. 3. In this figure we can identify two main parts, that serve as networks, the generator and the discriminator. The discriminator will output the fake or real classification affecting the discriminator loss or the generator loss being mutually exclusive results. The generator cannot see directly the input images which make it more difficult to learn the emotional patterns in the input audios, but while receiving the generator loss information it will learn to generate them as closer as possible to the original input data. In this case we assumed

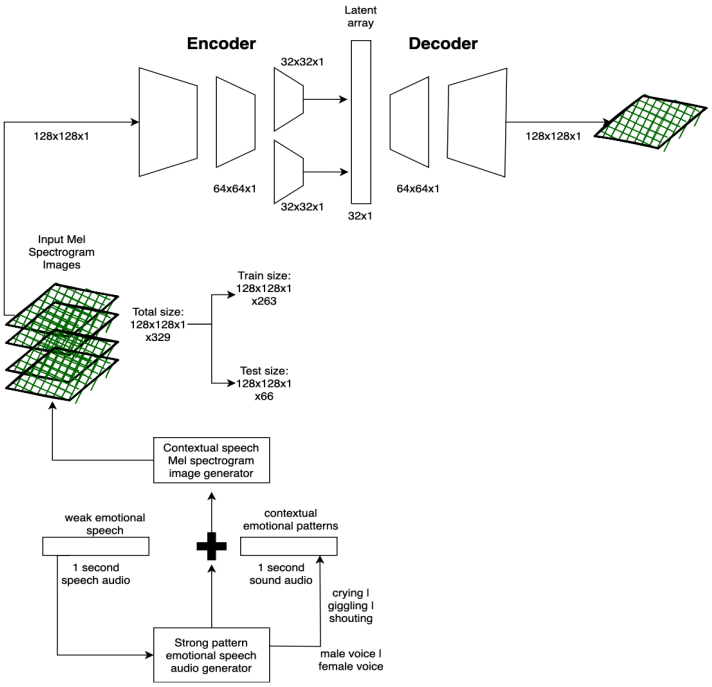


Fig. 1. Proposed architecture of the VAE model for training.

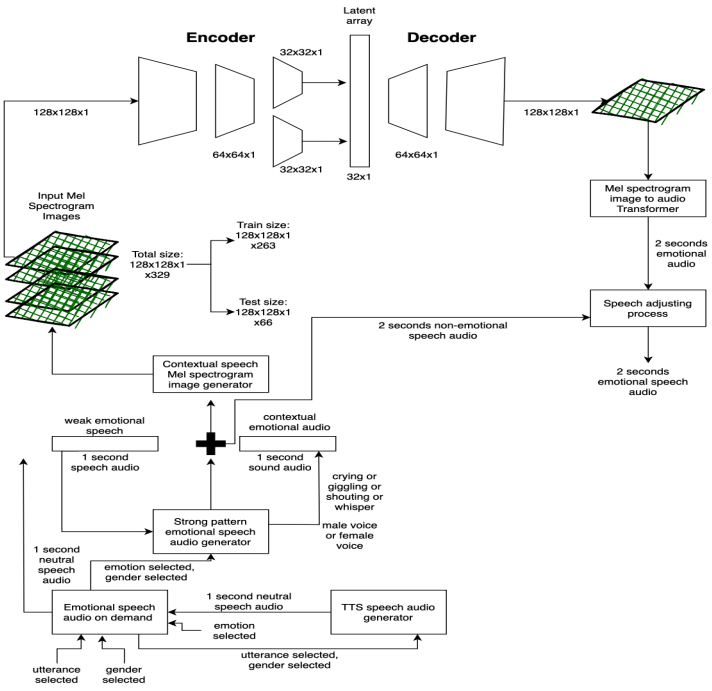


Fig. 2. Proposed architecture of the VAE model for testing.

some restrictions in the gender training to be male or female and not both of them, because CGAN is not good in terms of managing multiple conditions at the same time. In the testing of the CGAN model, the generator of the model is directly asked to produce the emotional data by emotion selected label.

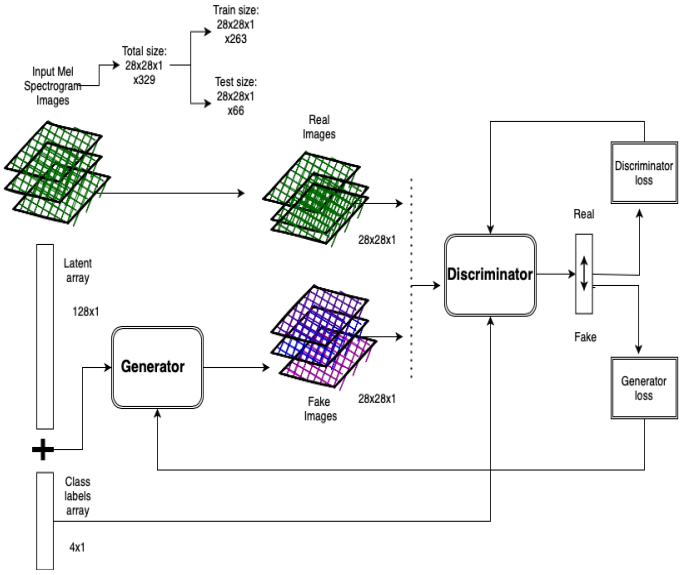


Fig. 3. Proposed architecture of the CGAN model for training.

In each model, the loss function is affected by the phonetic regularization with the extraction of the formants F1 to F5, presents in the Mel spectrogram data.

III. Experiments

A. Data Preprocessing

The input audio data before transforming into log Mel spectrograms can be observed in Fig. 4. This data were used for training, for the VAE and CGAN model to recognizse the happy speech and the contextual data that are usually present in a normal conversation. We limited our research to a one side speaker to be analysed.

The first audio A represents the speech of a male speaker from the Ravdess dataset that has an original duration of 3 seconds. The utterance in this happy speech is saying "Dogs are sitting by the door". The identification number in the original Ravdess dataset is "03-01-03-01-02-01-09" and is open to the public. This happy audio does not sound happy to our ears, since it is part of a weak emotional dataset, which is one of the reasons why the emotional transfer is complex to achieve in time series data. As can be observed in the image, the audio A contains zero data in the beginning, and at the end of the speech. With the aim to eliminate the non-useful data, the audio was reduced into a 1 second audio, while remaining the speech content.

The second audio B remains in 1 second, since originally each contextual data has 1 second of duration. This represents naturally what happens with human beings, since usually our laughing takes about the same time in being produced. The combination results in a total duration of 2 seconds for the concatenated audio C. These preprocessing tasks provide the desired adjustment, while maintaining a consistent audio quality.

The contextual information related to emotion is added as follows. We selected 6 audios per gender and per weakly emotion to add to any weakly emotional audio that matches by gender and emotion class, and chosen randomly. This is possible, given that, while singing or making emotional noises, we can not perfectly distinguishing these pattern belonging gender and voice. The duration of each audio is 1 second, its sampling rate is 44100Hz and its format extension is WAV. For happy emotion, we selected 6 female giggling patterns and 6 male giggling patterns, in total 12 audios for constructing the emotional extended audios. In the case of sad emotion, we selected 6 female crying pattern audios and 6 male crying pattern audios. When selecting the angry audios, we picked 6 female and 6 male shouting audios. However for neutral pattern audios we selected 6 female and 6 male whisper and natural English speaking pauses, such as, "Emmmm....", "AHA....", "Ahhhh....", "cause.....", and so on, where the utterance is almost not listened.

All the training and testing audios were denoised, trimmed and adjusted by volume, especially the audio B, since many artefacts were present. This is because it is custom data, downloaded freely from the Freesound site. Generally, custom data has many artefacts content present in the audios.

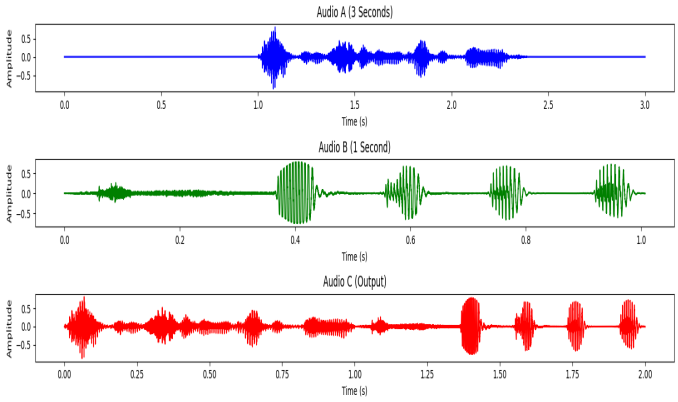


Fig. 4. Original audios A and B representing a happy speech, preprocessed and combined into one audio output file C.

The following step after gathering the clean concatenated audio is to transform the audio into frequency domain. The log Mel spectrogram image data generated for testing, as a case study can be observed in Fig. 5. The sampling rate is 44100Hz, the number of FFTs is 2048, with a hop length of 512. For the VAE model the target size is a 128 by 128 image. For the CGAN we selected the target 28 by 28, because CGAN generates better results with smaller sizes of images.

Before training, the pixel values were normalized to a (-1, 1) interval instead of using RGB values ranging from 0 to 255. Firstly, larger input values may slow down or disrupt the learning process of neural networks and setting them to smaller values is good practice [21]. Secondly, this normalization was required since the generator outputs tanh activations within the same interval.

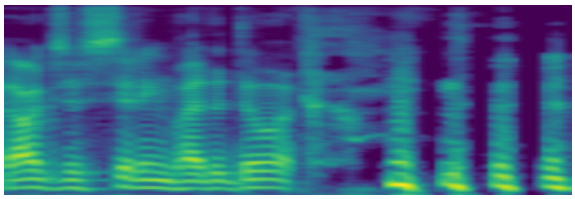


Fig. 5. Input log Mel spectrogram data example for the VAE and CGAN model. Happy emotion male speaker with original Ravdess dataset extended with male giggling sound.

B. Experimental Details

Two distinct methodologies were introduced, involving the VAE and CGAN models. The initial approach utilizes the VAE model in an entirely unsupervised manner, whereas the second approach employs the CGAN model in a supervised capacity.

Through the experiments, it was demonstrated that enhancing the input dataset with contextual attributes yields superior outcomes compared to explicit labeling or enforced supervision. Despite CGAN being a supervised model, it also benefits the core task of emotional transfer. This advantage is attributed to the data complexity being reduced when well-patterned contextual data is linked with the original input. Such simplification is only achievable with contextual information related to emotions and speaker gender.

In CGAN's labeled input-guided framework, we must assign appropriate labels to guide the generation of new samples. The model needs to be trained with labels that are familiar to it during the training phase. In contrast, the VAE utilizes extended input data without the explicit management of labels; instead, it separates classes by observing the data's intrinsic patterns during training.

To generate data from a specific class using the aforementioned CGAN model, several structured steps need to be followed. Firstly, a one-hot encoded label vector representing the desired class is created. This label must align with the format of labels in our training dataset. Secondly, random latent vectors are generated as input for the generator, sampled from a normal distribution. Thirdly, the one-hot encoded label vector is concatenated with the random latent vectors to form the conditional input for the generator. In the final step, the generator is employed to create data samples based on the prepared conditional input.

C. Experimental Results

The training dataset comprises 263 samples, while the test dataset consists of 66 samples. The distribution of training data is as follows: 37 samples for neutral audios, 72 for happy audios, 75 for sad audios, and 79 for angry audios. In the testing set, there are 11 neutral audios, 20 happy audios, 18 sad audios, and 17 angry audios. The limited number of "neutral" samples is due to the smaller quantity available in the Ravdess dataset for each actor, requiring additional steps to balance the dataset, such as cloning voices and creating neutral audios with different utterances and intensities. This task, although valuable, falls beyond the scope of our current research and could be explored in future studies.

The computational environment used for testing included a RAM occupancy of 5.38GB out of 51GB and a disk space usage of 26.83GB out of 166.77GB. The experiments utilized a NVIDIA T4 GPU provided by Google Compute Engine. The execution time for the CGAN model training over 10,000 iterations was 20 minutes. For the VAE models, the training process involved 10,000 epochs and took approximately 20 minutes. The programming language employed for these tasks was Python version 3.

To assess the VAE model, an unsupervised approach lacking predefined target classes, two metrics were utilized, mean square error (MSE) and structural similarity index (SSIM) between the original and generated data. In MSE, lower values indicate higher similarity between images, with 0 representing a perfect match, although realistically, a small value signifies good similarity. SSIM values range from -1 to 1, where 1 signifies a perfect match. Values closer to 1 indicate strong similarity, and a value above 0.9 is generally considered a robust match. It is essential to consider that ideal values can vary based on the specific domain and image quality.

VAE model results of training data after 10000 epochs, for original Ravdess dataset can be observed in Fig. 6, 7, and 8. The training and validation loss of the variational autoencoder trained with the original Ravdess dataset for 10000 epochs can be observed in Fig. 6. The training loss starts in a high value 10^5 from epoch 1 and the validation loss starts in 10 in the epoch 1. The values are steadily decreasing over the epochs as expected. The data generated by VAE model in 10000 epochs

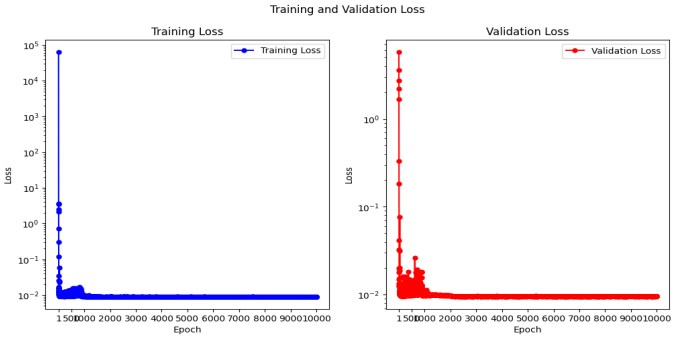


Fig. 6. Training and validation loss of the variational autoencoder after 10000 epochs of training with Ravdess dataset.

of training can be observed in Fig. 7. The generated results

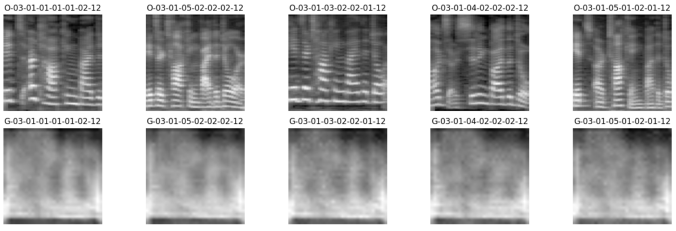


Fig. 7. VAE model random generation of different classes of data after 10000 epochs of training with Ravdess dataset. "O" stands for original, "G" for generated, and the emotion class is the 3rd number from left to right reading (01:Neutral, 03:Happy, 04:Sad, 05:Angry).

comparison with Ravdess dataset can be seen in Fig. 8.

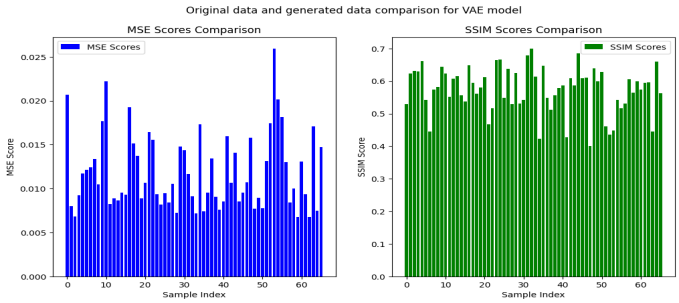


Fig. 8. VAE results comparison with Ravdess dataset. MSE and SSIM measured data results.

CGAN model results of training data after 10000 epochs, for original Ravdess dataset can be observed in Fig. 9, 10, 11, 12 and 13. CGAN loss for training data over 10000 epochs can be observed in Fig. 9. The angry, sad, happy and

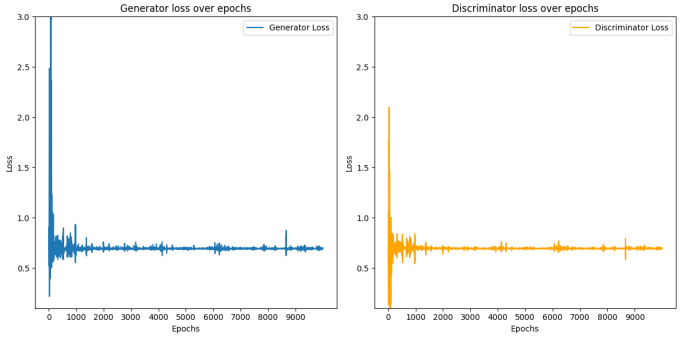


Fig. 9. CGAN loss over 10000 epochs of training.

neutral emotional data generated by CGAN model after 10000 epochs of training can be observed in Fig. 10, 11, 12 and 13, respectively.

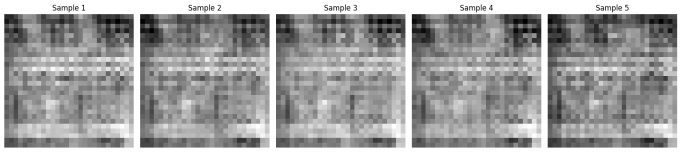


Fig. 10. CGAN model generation of data class 1 (neutral emotion) after 10000 epochs of training.

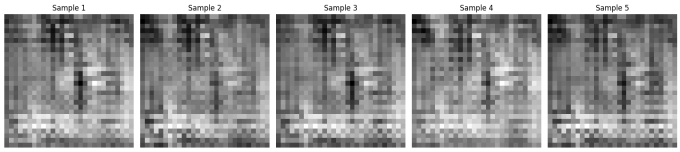


Fig. 11. CGAN model generation of data class 3 (happy emotion) after 10000 epochs of training.

VAE proposed model results after 10000 epochs of training, with Ravdess dataset extended with contextual information data, can be observed in Fig. 14, 15 and 16. The training

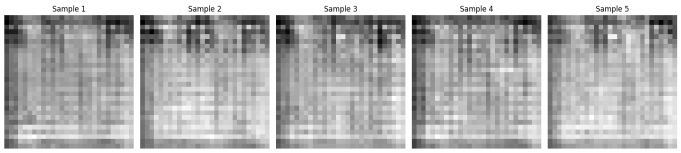


Fig. 12. CGAN model generation of data class 4 (sad emotion) after 10000 epochs of training.

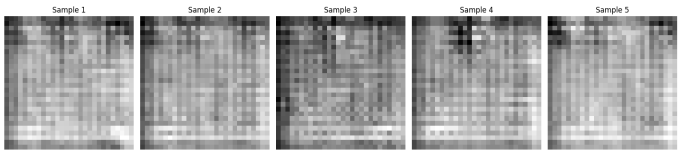


Fig. 13. CGAN model generation of data class 5 (angry emotion) after 10000 epochs of training.

and validation loss of the variational autoencoder trained with Ravdess dataset extended with contextual information data for 10000 epochs can be observed in Fig. 14. The training loss starts in a lower value above 10^1 from epoch 1 and the validation loss starts also in a lower value of 10^0 in the epoch 1, in comparison with the training and validation loss of original Ravdess dataset in Fig. 6. The values are steadily decreasing over the epochs as expected. The data generated

Fig. 14. Training and validation loss of the variational autoencoder after 10000 epochs of training with the Ravdess dataset extended with contextual information data.

by VAE model in 10000 epochs of training can be observed in Fig. 15. The generated data improved the previous original Ravdess training with VAE, since the quality of each image increased significantly while remaining the same training and testing conditions. The generated results comparison after

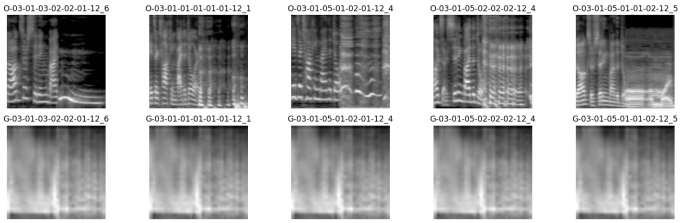


Fig. 15. VAE model random generation of different classes of data after 10000 epochs of training with contextual information extended data. “O” stands for original, “G” for generated, and the emotion class is the 3rd number from left to right reading (01:Neutral, 03:Happy, 04:Sad, 05:Angry).

10000 epochs of training with Ravdess dataset extended with contextual information data, can be seen in Fig. 16. CGAN

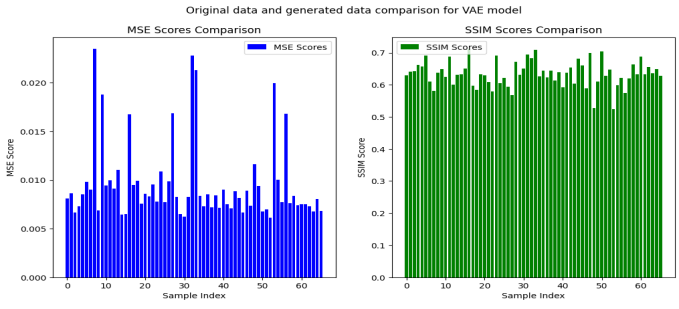


Fig. 16. VAE results comparison with the proposed custom dataset. MSE and SSIM measured data results.

proposed model results after 10000 epochs of training, for Ravdess dataset extended with contextual information data can be observed in Fig. 17, 18, 19, 20 and 21. CGAN loss for training data over 10000 epochs can be observed in Fig. 17. It shows that the loss starts with a higher value and decreased accordingly, as it is expected for loss functions in both sides, with no strange jumpings or increasing values from both sides. It also shows a convergence and a estabilization point. The

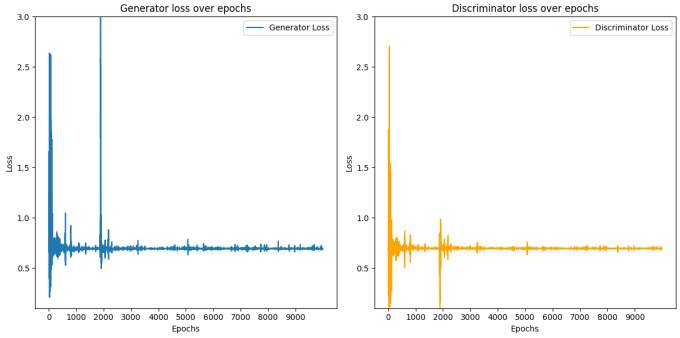


Fig. 17. CGAN loss over 10000 epochs of training.

angry, sad, happy and neutral emotional data generated by CGAN model after 10000 epochs of training can be observed in Fig. 18, 19, 20 and 21, respectively. The generated data improved the previous original Ravdess training with CGAN, since the quality of each image increased significantly.

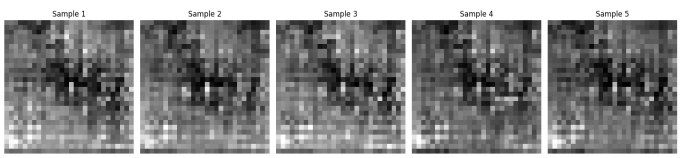


Fig. 18. CGAN model generation of data class 1 (neutral emotion) after 10000 epochs of training.

To validate the accuracy of the generated data, we developed an emotional classifier program, trained on the same input data utilized in both the VAE and CGAN models. The classifier's accuracy indicates the models' ability to generate emotional data that can be correctly classified into specific

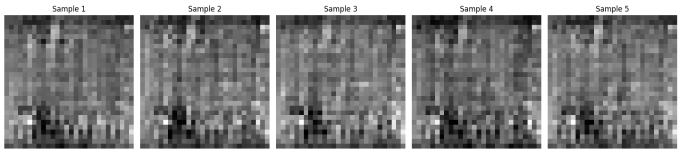


Fig. 19. CGAN model generation of data class 3 (happy emotion) after 10000 epochs of training.

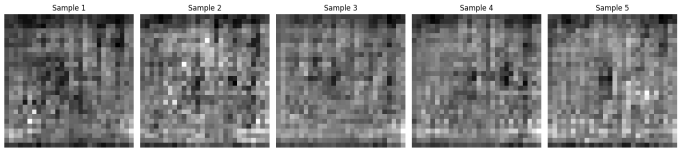


Fig. 20. CGAN model generation of data class 4 (sad emotion) after 10000 epochs of training.

emotion classes. It is important to remark that this measure can not obtain a 100 percent of accuracy due to the lack of train and validation over this fresh generated data. Therefore, the generated data is non-trained and unseen data for the classifier, however it will give us a notion of how much we should adjust the classifier and the models to improve the results. Most importantly, it will result in all classes zero classified, if it is not able to detect any emotional generated data.

When testing both models with the original Ravdess input data, it is anticipated that the generated results to be weakly classified by the classifier. This expectation arises because the Ravdess data, as previously discussed, is inherently weakly emotional. Moreover, it is important to acknowledge the inherent noise in the outcomes produced by generative models, particularly concerning audio data features such as log Mel spectrogram values.

The VAE generated data was sent to the classifier for testing generation accuracy. The results can be observed in Fig. 22. The results of the classifier after training with original Ravdess data shows a 20.51 percent of correct classification of emotions in a total of 39 tested generated data. Even though the result is low in comparison with a perfect accuracy, as explained above, it is expected that the classifier cannot easily extends its classification with these non-trained unseen data. This is the evidence that our CGAN model in Fig. 23, is trying to reconstruct the data getting a better emotional generated data's result, in comparison with VAE model. Additionally, the classifier needs to be adjusted for future works to measure more precisely our data.

The CGAN generated data was sent to the classifier for testing generation accuracy. The results can be observed in

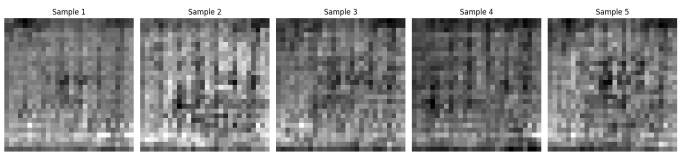


Fig. 21. CGAN model generation of data class 5 (angry emotion) after 10000 epochs of training.

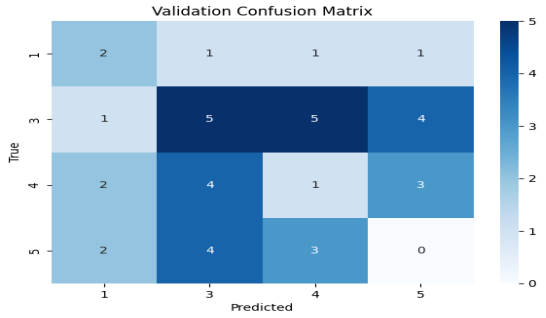


Fig. 22. Validation confusion matrix for VAE model generated data after classified by an emotional classifier.

Fig. 23. The results of the classifier after training CGAN with extended data shows a 25.641 percent of correct classification of emotions in a total of 39 tested generated data. Even though the result is low in comparison with a perfect accuracy, as discussed above, it is expected that the classifier cannot easily extends its classification with these non-trained unseen data. This is the evidence that our model is trying to gradually reconstruct the data and the classifier needs to be adjusted for future works to measure more precisely our generated data.

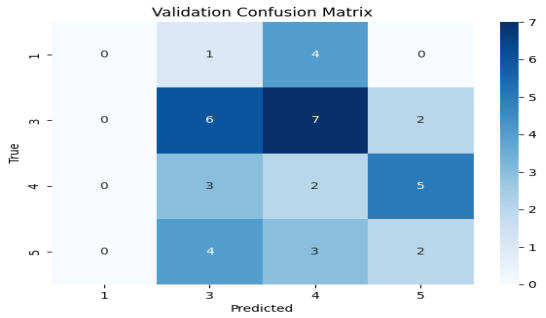


Fig. 23. Validation confusion matrix for CGAN model generated data after classified by an emotional classifier.

During this evaluation, we anticipate the input data to be classified in closer proximity to the generated data. However, we cannot definitively confirm how the generative models interpret the essential features in the presence of additional contextual information. Nevertheless, the generated data is expected to align more closely with its input data due to the utilization of paired information. For instance, the weak speech emotion in the Ravdess dataset, when associated with the extended contextual information like a female voice's giggling pattern linked to a happy emotion, should be more accurately classified as happy. In comparison with the vanilla VAE model and standard CGAN model trained with original Ravdess dataset, the results are clearly improved. The results shows also that more efforts are required for future works in order to better represent the reconstructed emotional audios.

D. Case Study

As an extension of the emotional data generation capabilities with the proposed VAE and CGAN models, in this case study we showcase the specific emotion transfer with additional conditions such as specific gender, specific voice, and specific utterance, with an additional change in emotions, from neutral to happy, sad, angry respectively. Since the variational auto-encoder model is more flexible in terms of receiving new input data to regenerate, we reutilized the pre-trained proposed variational auto-encoder model.

Furthermore, the utterances in this study were generated using a Text-to-Speech (TTS) system, which inherently produces a "Neutral" emotion speech audio since it lacks emotional variation. Converting "Neutral" audio to another "Neutral" audio is not necessary for our evaluation; it would not yield any change and only indicates the models' understanding of elements like whispering or speaking pauses. Although our training data includes neutral sounds, such as whispers or slight delays in speech, they are not utilized in our TTS systems for testing purposes.

For specific test scenarios, assuming the model is trained, we paired our TTS-generated voice with additional emotional audio, creating samples as follows:

1) 'Bob is by the door' with female giggling.
2) 'Bob is by the door' with male giggling.
3) 'Bob is by the door' with female shouting.
4) 'Bob is by the door' with male shouting.
5) 'Bob is by the door' with female crying.
6) 'Bob is by the door' with male crying.

To demonstrate the practicality of this approach, a "Neutral" emotional speech is converted into a "Happy" emotional speech on demand through a TTS system, by using the proposed variational auto-encoder model with custom data for the reconstruction. A selected raw audio data generated for testing can be seen in Fig. 24. The TTS system utilizes a neutral MAC PC male voice "Alex", uttering "Bob is by the door". The duration was condensed to one second for consistency across speech data. The TTS voice is neutral without any emotional inflection. For this use case, the utterance was varied from the trained sets u_1 ("Kids are talking by the door") to u_2 ("Dogs are sitting by the door"). This variation illustrates the model's ability to transfer emotion even with slight differences in the trained words.

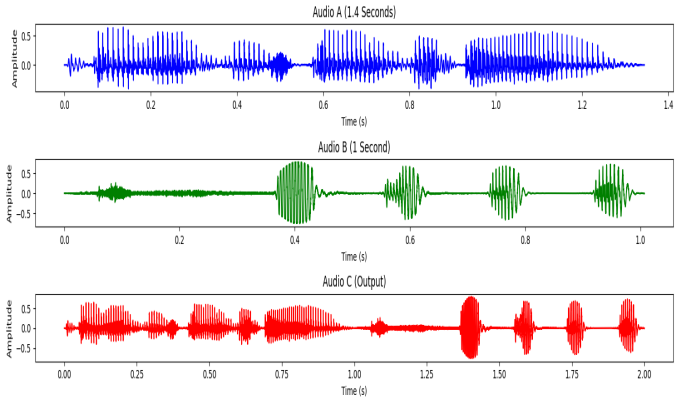


Fig. 24. Raw audio A is a Neutral speech audio generated by a TTS MAC OS male voice system. Original audio B representing a happy emotion by giggling. Preprocessed audios were combined into one audio output file C.

The following step after gathering the clean concatenated audio is to transform the audio into frequency domain. The log Mel spectrogram image data generated for testing, as a case study can be observed in Fig. 25. The sampling rate is 44100Hz, the number of FFTs is 2048, with a hop length of 512. For the VAE model the target size is a 128 by 128 image.

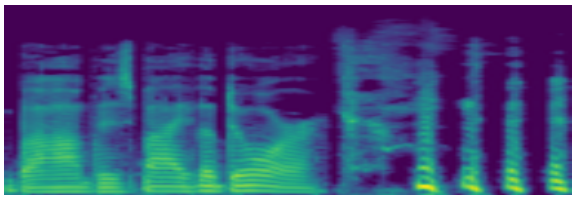


Fig. 25. Input log Mel spectrogram example for the VAE. Neutral speech with additional happy contextual data.

The result of the VAE model emotion transfer can be observed in Fig. 26, when given the happy bob spectrogram generated, the proposed VAE model try to reconstruct it with happy emotion.

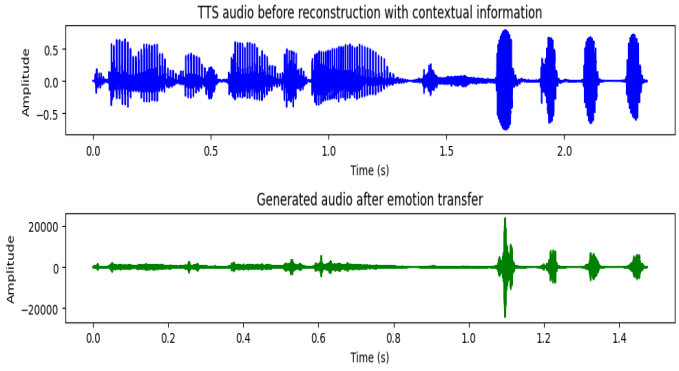


Fig. 26. Result of the VAE model emotion transfer from neutral to happy speech sample.

In this implementation, an unsupervised model was used to infer the emotion that should be reconstructed by contextual information. This task was developed with the proposed VAE model. The TTS speech was created with 1 second plus a related audio extension of 1 second, depending of the gender and the emotion desired. These samples show how the emotion is reconstructed among the trained utterances and among similar utterances with words composition variability. The utterance selected for this testing is as follows. u_3 : "Bob is by the door".

The total data size used are 329 audios, the train data size is 263 audios, transformed into log Mel spectrogram images. The test data size is 66. This case study showcase one sample among 6 samples presented. The size of each image is 128 by 128. The initial VAE training is performed with grayscale images. Each audio has a sampling rate of 44100Hz, which makes it more difficult to recover the audio after passing to Mel spectrogram.

Some pending tasks are required such as volume regularization and denoising tasks, however it can be seen how the contextual information such as giggling in a male speaker is greatly represented and trying to recover the left side of the speech with the same accent and pauses that the male speaker represents.

IV. Discussion

To assess the VAE model, an unsupervised approach lacking predefined target classes, two metrics were utilized, mean square error (MSE) and structural similarity index (SSIM) between the original and generated data. In 66 reconstructed audios for the proposed VAE model, when testing the custom data, the MSE is lower than 0.010 and the SSIM is almost reaching the 0.70, while most of the values are greater than 0.60. For the Ravdess dataset, the MSE of the reconstruction is lower than 0.020 and the SSIM is almost reaching the 0.70, while most of the values are higher than 0.50. This signify that the proposed data along with the proposed model is getting better similarity in the reconstruction of tested data, while the mean squared error is lower which emphasizes the improvement in the reconstruction by using this proposed unsupervised model.

An external tool was created to measure the level of reconstruction in terms of emotions. This tool is an emotional classifier, trained with the same emotion classes and the same input data size and characteristics as the proposed models. Certainly the reconstructed data from CGAN by setting a specified emotion label and the TTS generated data used as input for the VAE make them different than the trained data of the created tool. However, this tool can state if the generated data by both models is not considered emotional at all by the classifier showing near zero values in the accuracy. This tool help to test the research results while avoiding the subjective measures by human resources. The VAE and the CGAN generated data were sent to the classifier for testing generation accuracy. The results of the classifier after training with custom data shows a 20.51 percent of correct classification of emotions in a total of 39 tested generated data for VAE. The results of the classifier after training CGAN with custom data shows a 25.641 percent of correct classification of emotions in a total of 39 tested generated data. Even though the result is low in comparison with a perfect accuracy, as explained above, it is expected that the classifier cannot easily extends its classification with these non-trained unseen data.

The results of the classifier also show that there is a need of a better vocoder function in future improvements of this work, and that the output should be better adjusted by a volume regularizer and an extra denoise function. This is due to the weak points present in generative models, where a better vocoder could probably increase the accuracy in the emotional classifier results, after generating new unseen and non-trained data. The new generated data is still identified as different for the emotional classifier, because the data just created posses different characteristics such as moved characteristics from the original input data that the classifier usually consumes.

The treatment of the utterances could be improved in future works, such as injecting the words spoken and its characteristics right after the generative models resampling. This is because the generative models tend to loose the accents of each spoken word during the transition from original data

to resampled data. However, a strong point of this research to mention is the extension of the trained utterances to similar utterances that have some words in common but differ in others. As it is known, with models such as HMM, RNN and LSTM, the utterances are totally fixed without any possibility of extension. This also make the models fail in the presence of utterances that contain different words, even when the general pronunciation of the words is similar and the sentence has the same duration. In comparison with other datasets, we used the Toronto emotional speech set (TESS) with two female speakers and a variation in utterances such as "Say the word book", "Say the word bought", or other variations from 200 target words for the emotions happy, sad, angry and neutral. SSIM and MSE results with our proposed model and extension with contextual cues show, for female speaker's emotions, similarity to the evaluated Ravdess dataset with 0.7 and 0.020, respectively. Therefore, demonstrating the ability to be robust and scalable to other utterances and voice characteristics. For future works, the evaluation with other datasets that include the male counterpart audios is also required for better determine the level of scalability by gender.

V. Conclusion

In this research, a method for transferring emotional speech utilizing generative models and specific emotional targets for the output was presented. The generative models employed in this task include a variational autoencoder model and a conditional generative adversarial network model. Although further refinement is needed to enhance the accuracy of generated emotional speech, both proposed models have demonstrated the ability to reconstruct emotional speech with a certain degree of quality.

In the presented case study, it was utilized an input voice audio, a desired utterance, and user-selected emotional cues to automatically transform ordinary speech into emotionally expressive speech audio using a variational autoencoder model. Remarkably, the proposed VAE model achieved this task without requiring specific labels to control the generated output, highlighting the efficiency of this approach with unsupervised learning. The model attempts to replicate the emotion inherent in the emotional contextual cues used for training directly into the ordinary speech. The findings reveal that the suggested unsupervised VAE model for generating emotional data achieves an MSE below 0.010 and an SSIM nearly reaching 0.70. The majority of values surpass 0.60 in comparison to both the input and generated data. When generating new emotional data on demand, CGAN and VAE models demonstrate a discernible level of success in the evaluation of an emotion classifier, determining its similarity to real emotional audios used for training.

In future works, the primary focus will be on refining the generative models further and implementing additional strategies to achieve a better balance between real contextual information and emotionally rich speech. This ongoing effort aims to bridge the gap between human understanding and artificial agents' comprehension in the essence of a desirable authentic speech.

References

[1] Z. Du, B. Sisman, K. Zhou, and H. Li, "Disentanglement of Emotional Style and Speaker Identity for Expressive Voice Conversion," in Interspeech, Sep 2022.

[2] HSU, Jia-Hao, et al. Empathetic Response Generation based on Plug-and-Play Mechanism with Empathy Perturbation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.

[3] PAIVA, Ana, et al. Caring for agents and agents that care: Building empathic relations with synthetic agents. In Autonomous Agents and Multiagent Systems, International Joint Conference on IEEE Computer Society, 2004. p. 194-201.

[4] R. Shankar, H.-W. Hsieh, N. Charon, and A. Venkataraman, "Multi-speaker emotion conversion via latent variable regularization and a chained encoder-decoder-predictor network," in Proc. Annu. Conf. Int. Speech Commun. Assoc., pp. 3391– 3395, 2020.

[5] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2020, pp. 6284–6288.

[6] K.Zhou, B. Sisman, R. Rana, B.W.Schuller and H.Li, "Emotion intensity and its control for emotional voice conversion," IEEE Tran. on Affective Computing, 2023.

[7] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," in Proc. Odyssey Speaker Lang. Recognit. Workshop, 2020, pp. 230–237.

[8] DENG, Jun, et al. Recognizing emotions from whispered speech based on acoustic feature transfer learning. IEEE Access, 2017, vol. 5, p. 5235-5246.

[9] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion," in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2016, pp. 2453–2457.

[10] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in Proc. IEEE Int. Conf. Acoust., Speech Signal Pro- cess., 2017, pp. 4905–4909.

[11] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2015, pp. 4475–4479.

[12] R. Shankar, J. Sager, and A. Venkataraman, "Non-parallel emotion conversion using a deep-generative hybrid network and an adversarial pair discriminator," 2020, arXiv:2007.12932.

[13] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2020, pp. 3502–3506.

[14] K. Zhou, B. Sisman, and H. Li, "Vaw-GAN for disentanglement and recomposition of emotional elements in speech," in Proc. IEEE Spoken Lang. Technol. Workshop, 2021, pp. 415–422.

[15] D. P.. Kingma, M. Welling, "Auto-encoding variational Bayes", in Proc. 2nd International Conference on Learning Representations, 2014.

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. June 10, 2014. arXiv: 1406.2661 [cs, stat]. URL: http://arxiv.org/abs/ 1406.2661

[17] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. Nov. 6, 2014. arXiv: 1411.1784 [cs, stat]. URL: http://arxiv.org/abs/1411.1784

[18] Ali Borji. Pros and Cons of GAN Evaluation Measures. Oct. 23, 2018. arXiv: 1802.03446 [cs]. URL: http://arxiv.org/abs/1802.03446

[19] Pegah Salehi, Abdolah Chalechale, and Maryam Taghizadeh. Generative Adversarial Networks (GANs): An Overview of Theoretical Model, Evaluation Metrics, and Recent Developments. May 27, 2020. arXiv: 2005.13178 [cs, eess]. URL: http://arxiv.org/ abs/2005.13178

[20] Sudarshan Adiga, Mohamed Adel Attia, Wei-Ting Chang, and Ravi Tandon. "On the tradeoff between mode collapse and sample quality in generative adversarial networks". In: 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP). Anaheim, CA, USA: IEEE, Nov. 2018, pp. 1184–1188. I SBN: 978-1-72811-295-4. DOI: 10.1109/GlobalSIP.2018.8646478. URL: https:// ieeexplore.ieee.org/document/8646478/

[21] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. June 10, 2016. arXiv: 1606.03498 [cs]. URL: http://arxiv.org/abs/1606.03498

Imbalance Node Classification with Graph Neural Networks (GNN): A Study on a Twitter Dataset

Alda Kika, Arbër Ceni, Denada Çollaku, Emiranda Loka, Ledia Bozo, Klesti Hoxha
University of Tirana, Faculty of Natural Sciences, Department of Informatics
Bulevardi Zogu I, Tiranë 1001, Albania

***Abstract*—Social networks produce a large volume of information, a part of which is fake. Social media platforms do a good job in moderating content and banning fake news spreaders, but a proactive solution is more desirable especially during global threats like COVID-19 pandemic and war. A proactive solution would be to ban users who spread fake news before they become important spreaders. In this paper we propose to model user's interactions in a social media platform as a graph and then evaluate state of the art (SOTA) graph neural networks (GNN) that can classify users' (nodes) profiles as being suspended or not. As with other real world data, we are faced with the imbalanced data problem and we evaluate different algorithms that try to fix this issue. Data used for this study were collected from X (Twitter) by using Twitter API 1.1 from November 2021 to July 2022 with the focus to collect information spread through tweets about vaccines. The aim of this paper is to evaluate if current models can deal with real world imbalanced data.**

***Keywords*—GNN; imbalanced data; Twitter; social networks; GCN; GraphSage; GAT; GraphSMOTE; ReNode**

I. Introduction

Social media has changed the way that the news is created and spread worldwide. According to Statista [1] in 2022 over 4.59 billion people were using social media, a number which is estimated to increase to almost six billion in 2027. In March 2020 World Health Organization declared Coronavirus disease (COVID-19) pandemic. Very soon a lot of people started to use social media to spread information about this new disease resulting in a massive infodemic. Infodemic[1] refers to false information related to a disease. This phenomenon is amplified through social networks spreading farther and faster like a virus [2]. World Health Organization (WHO) Director-General, Tedros Adhanom Ghebreyesus, in [3] emphasized the importance of fighting not only the pandemic but also the infodemic that was spreading. During difficult time for all the countries worldwide it is very important to spread the correct information rather than false and fake news that undermines the global response and jeopardizes measures to control the pandemic [4].

There are various definitions of fake news. Allcott and Gentzkow [5] defined fake news as "news articles that are intentionally and verifiably false and could mislead readers". Other studies have defined it as "a news article or message published and propagated through media, carrying false information regardless of the means and motives behind it" [6], [7], [8]. Much fake news about COVID-19 on social media has been circulating but the spread of fake news and disinformation about vaccines had a significant influence on vaccine acceptance, with many people opting not to get vaccinated posing a threat to individual and collective health. In order to prevent the spreading of fake news it is important to develop tools that will predict this.

Tweets associated with COVID-19 have been collected for a period of nine months from November 2021 until July 2022 for the purpose of applying social network analysis techniques on these posts. In [9], the authors focused their analysis on three main terms: pfizer, moderna and AstraZeneca. For these terms and for the nine month period, 27 graphs were built and an analysis of these graphs was performed. It was pointed out that most of the time the most influential users (based on beetweenness centrality) in the network were engaged in a form of information disorder and their accounts were suspended at the time of the study. Unfortunately the real reason behind the account suspension is not declared by X (Twitter), but one can safely presume that the main reason an account can be suspended is if a user does not comply with the TOS and engaging in mis/dis information is not in compliance with the platforms' TOS [2], [3]. We value that it is of great interest to be able to predict a user's status based on his position in the network as well as his attributes (date of joining the platform, number of posts, number of followers etc.). To this purpose, we applied different graph neural networks (GCN [10], GraphSage [11] and GAT [12]) on our dataset for the downstream task of classifying nodes. Given the fact that the dataset is imbalanced (there are more active users than inactive users), models which claim to fix this problem [13], [14], [15], [16], [17] were also studied.

The paper is organized as follows: Section II reviews recent studies on deep learning models on node classification in graph neural networks about fake news and infodemic prediction. The data that is used in this study, as well as the methodology that is applied for node classification and evaluation, is presented in Section III. In Section IV, the results are discussed. The conclusion and a concise summary conclude this paper.

II. Related Work

Graph neural networks are the most reliable and effective way to automate the process of fake news detection [18]. Feng et al. in [19] propose a bot detection framework that encode multi-modal user information such as user account description, the content of the post and numerical and categorical features that they can assign to an account available from the

[1]https://www.who.int/health-topics/infodemic

[2]https://help.twitter.com/en/resources/addressing-misleading-info

[3]https://help.twitter.com/en/rules-and-policies/crisis-misinformation

Twitter API. Users are treated as nodes of the heterogeneous graph. Relational Graph Convolutional Network is applied to the graph to tackle the challenges of bot disguise and bot communities.

In another work [20] for bot detection the authors propose a transductive model that combines symmetrically BERT and GCN. They constructed a heterogenous graph composed of unique words and documents in a collection as nodes. Word occurrences in a document or word co-occurrence are weighted using TD-IFD and PMI and stored as information on the edges of the graph. The experiment shows that a better performance can also be achieved by the proposed framework on a wide range of social robot detection datasets.

Li and Goldwasser [21] constructed a social information graph by embedding the items that are shared by Twitter users, the users who can be political users and other users who spread content. The edges represent the follower relations between political users and other users that follow them. They used two-layer Graph Convolutional Networks to capture the documents' social context and to detect news items as fake or not by node classification.

In [22] the dataset of CheckThat-2022 Task 3 dataset [23] is used as the primary dataset to analyze a corpus with unknown topics through multiclass classification, encompassing true, false, partially false, and other categories. They have explored three BERT-based models—SBERT, RoBERTa, and mBERT. The problem of imbalanced dataset is tackled by enhancing results via ChatGPT-generated artificial data for class balance and improving the results for the true, partially false, and other classes witnessed improvements of 5%, 9%, and 3%, respectively, while the results for false news experienced a decline of 9%.
Data augmentation has been applied by adopting synthetic minority oversampling in Zhao et al. [16]. They proposed a novel framework, GraphSMOTE, based on SMOTE [24] approach which addresses the imbalance problem by generating new samples, performing interpolation between samples in minority classes and their nearest neighbors. In [17] the authors have followed the approach of the imbalance of topological structure on the graph for handling the imbalance problem in graph-structured data. They proposed a ReNode framework for solving the problem of edge graph structure by re-weighting the influence of labeled nodes adaptively based on their relative positions to class boundaries.

III. Methodology and Data

In the following section we describe the dataset creation process as well as a description of different GNN models.

A. The Dataset

The data used in this paper is described in [9]. The data consist of 27 graphs. The size of each graph is displayed in Table I. To be able to run any graph neural network in any of these graphs, first we need to convert these graphs into a dataset that can be fed to the GNN model. To this purpose, we decided to use PyTorch[4] and PyG[5] libraries.

TABLE I. Size of the Created Graphs

	Pfizer		Moderna		AstraZeneca	
Month	**Nodes**	**Edges**	**Nodes**	**Edges**	**Nodes**	**Edges**
November 2021	367294	731875	358701	602031	216976	390200
December 2021	278266	543490	267581	430523	167942	288595
January 2022	198258	451691	247518	413592	110906	192169
February 2022	229197	673992	298231	600610	117994	213708
March 2022	244193	717114	323143	634862	93629	176314
April 2022	203168	608647	275389	569098	66175	201440
May 2022	210783	626985	273315	629752	79209	222855
June 2022	172249	526720	238628	494884	56726	130264
July 2022	40060	75511	56759	79718	23721	39747

We used the procedure[6] described in the PyG documentation to create our dataset. The advantage of converting the data into a dataset is that the data can then be fed into different neural networks without having to adapt the data each time. Another advantage is that we can make use of the many functions available in PyG to correctly split the dataset into development, test and validation sets. Creating a dataset from the available graphs, involves choosing node and edge features from the possible attributes. All node and edge features need to be of numeric type which limits our choice. In Tables III and IV node and edge features are shown. A transformation was required for some of the features as not all of them are numeric. Boolean features were transformed to 0 or 1. Date/time features were transformed to UNIX timestamp and category features such as "Relationship" were transformed to a number from 1 to 5 (Tweet = 1, Retweet = 2, Mentions = 3, Replies to = 4 and MentionsInRetweet = 5). The active status of a user was obtained by automatically making an HTTP request to the X (Twitter) web site since the API is no longer available. The active status for every user (node) in all the 27 constructed graphs was obtained in August 2023. As we expect in every real world dataset, we can notice by looking at Table II that the data is imbalanced for this dataset (i.e. there are fewer inactive users than active users). This will affect the prediction results of the standard models making them biased towards the majority class. Special models which account for the imbalance of the data need to be tested in this case.

B. GNN Models

Graph Neural Networks (GNNs) are a type of deep learning model designed for handling data represented in graph structures. In recent years, the field of graph neural network research has witnessed significant advancements [25], with a notable expansion in the range of GNN designs [25], [26]. Among these various designs in this paper we use Graph Convolutional Network (GCN) [10], GraphSAGE [11], and Graph Attention Networks (GAT) [12]. GCNs [10] are adapted to handle irregular and non-Euclidean data. For each node in the graph, GCN aggregates the features of all the neighbors of the node and the node itself. Different functions can be used for feature aggregation. The aggregated values are passed to the neural network which returns the feature vector resulting from the model. The GNC model can use several GCN layers on top of each other where the output of one layer will

[4]https://pytorch.org/
[5]https://pyg.org/
[6]https://pytorch-geometric.readthedocs.io/en/latest/tutorial/create_dataset.html

TABLE II. DATASET STATISTICS

	Pfizer			Moderna			AstraZeneca		
Month	**Inactive**	**Active**	**IR**	**Inactive**	**Active**	**IR**	**Inactive**	**Active**	**IR**
November 2021	20216	347078	0.058	16968	341733	0.049	10015	206961	0.048
December 2021	15515	262751	0.059	13278	254303	0.052	7843	160099	0.048
January 2022	11623	186635	0.062	12174	235344	0.051	5415	105491	0.051
February 2022	13402	215795	0.062	17089	281142	0.060	5611	112383	0.049
March 2022	14087	230106	0.061	17384	305759	0.056	4313	89316	0.048
April 2022	11678	191490	0.060	14217	261172	0.054	3638	62537	0.058
May 2022	12492	198291	0.062	14813	258502	0.057	4117	75092	0.054
June 2022	10128	162121	0.062	12207	226421	0.053	2853	53873	0.052
July 2022	2181	37879	0.057	2056	54703	0.037	1139	22582	0.050

TABLE III. NODE FEATURES

Feature Name
ID
Active
Degree
In-Degree
Out-Degree
Betweenness Centrality
Closeness Centrality
Eigenvector Centrality
PageRank
Clustering Coefficient
Reciprocated Vertex Pair Ratio
User ID
Followed
Followers
Tweets
Favorites
Joined Twitter Date (UTC)
Listed Count
Verified
Tweeted Search Term?
Vertex Group

TABLE IV. EDGE FEATURES

Feature Name
Relationship
Relationship Date (UTC)
Imported ID
In-Reply-To Tweet ID
Favorited
Favorite Count
In-Reply-To User ID
Is Quote Status
Retweet ID
Unified Twitter ID
Vertex 1 Group
Vertex 2 Group

be the input for the next layer. GraphSage [11] generates node embeddings by sampling and aggregating information from their neighbors. This model is suitable for large graphs because it is not necessary to process the entire graph at once, thus avoiding the limitations derived from the processing power and memory capacity available to us. The input of the GraphSage model is the graph which consists of nodes and edges. Each node in the graph has some features. To find node embeddings, the model selects for each node a fixed-size subset of its neighbors. The features of the selected neighbors are aggregated by means of an activation strategy thus creating an aggregated representation for each node. This aggregation reveals information about the local neighborhood of the node. These aggregated representations are passed as input to an activation function. To discover more complex relationships, updated representations can be passed through several layers of the GraphSage model. The last layer of the network is determined by the task to be performed. This model can be used to classify a node, predict a link. GAT [12] discovers dependencies and relationships in graph-structured data. Unlike GCN where all neighbors of a node have the same importance, GAT uses an "attention" mechanism to weight the importance of each neighbor during the aggregation process. In this way the network works only with the information of the nodes that are important. Weights are learned during the training process based on the importance each neighbor has on the specific task. Despite the progress in learning from graphs with GNN, existing work mainly focuses on balanced datasets [13]. In real world applications, we are faced with imbalanced data and GNNs fail to accurately predict the samples that belong to the minority class. Imbalanced class problems can be solved by modifying the GNN model to bias toward minority class, or by resampling that consists in altering the dataset by adjusting the number of instances for each class to achieve a balanced distribution [16], [14]. The most used approach is resampling because it can be integrated with any classifier [16], [14]. Resampling can be achieved either by undersampling or oversampling. Undersampling methods remove instances that belong to the majority class, but this can result in a loss of valuable information [16], [14]. Oversampling methods increase the number of minority classes which can lead to overfitting [16]. The results of applying these approaches to graphs are suboptimal because they consider each sample as independent and don't take into account the relation that exists in the graph data [16]. In this paper we first apply the three standard SOTA models (GCN, GraphSage and GAT) to classify nodes as being active or inactive. Given the fact that our data is imbalanced we also use GraphSMOTE [16] and ReNode [17] to deal with imbalance data. GraphSMOTE uses a feature extractor to learn node representations. Node representation should reflect the similarities and dissimilarities between samples considering node attributes, node labels, and

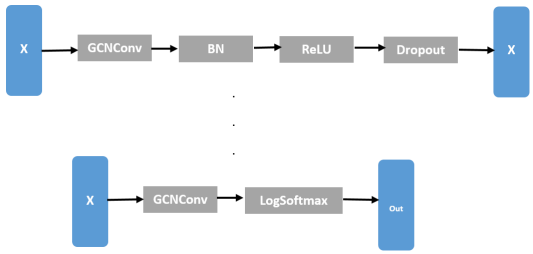


Fig. 1. The GCN model.

local graph structures [16]. After obtaining node representation for each node, GraphSMOTE uses the SMOTE algorithm to generate synthetic minority nodes in the latent space [16]. In order to include the new nodes in the graph, GraphSMOTE simulates connections between synthetic nodes by performing a training on existing nodes and edges. The GNN classifier can then perform node classification based on the augmented graph created by GraphSMOTE. Another framework that we have used to handle the topology imbalance issue of our network is ReNode. ReNode framework re-weight the influence of labeled nodes according to their position. Their structural position are located by using the conflict detection-based Topology Relative Location (Totoro) metric to leverage the interaction among them. The training weights of nodes with small conflict are increased and vice versa [17].

IV. Experimental Results

In this section, we present the results of our experiments in our imbalanced dataset. It will try to answer the following research questions:

- RQ1: Do standard GNN models yield acceptable results in case of imbalanced data?
- RQ2: Do special GNN models designed for imbalanced data help in our case?

A. Experimental Setup

The models for GCN, GraphSage and GAT are implemented in Pytorch and PyG by following a message passing paradigm. All the training is done in python 3.8 in a machine with 1 Titan RTX with 24GB, AMD Ryzen Threadripper PRO 5995WX 64-Cores CPU with 451 GB of RAM.

1) GCN implementation: The GCN model consists of three layers as shown in Fig. 1. Each layer consists of a GCNConv module, a batch normalization module, a ReLu activation function and a dropout. The last layer of the model contains only a GCNConv module and a LogSoftmax which yields the output. The input dimensionality is the number of features we have selected (21). Each intermediate (hidden) layer has 256 dimensions and the output layer has a dimensionality equal to the number of classes in the dataset (2). The learning rate is set to 0.01, dropout to 0.5 and the training was set to run for 500 epochs.

2) GraphSage implementation: The GraphSage implementation follows the architecture described in [11]. It consists of two message passing layers with 21 input dimensions, 32 hidden dimensions and 2 output dimensions. The dataset was divided into batches with a batch size of 32. Other parameters include: Adam optimizer, learning rate is set to 0.01, dropout probability equal to 0.5 and the training was set to run for 500 epochs.

3) GAT implementation: Our GAT implementation follows the architecture presented in [12]. It consists of 2 message passing layers with 21 input dimensions (equal to the number of node features), 32 hidden dimensions and 2 output dimensions. The number of attentions heads was set equal to 2. Similar to GraphSage, the dataset was divided into batches with batch size of 32, Adam optimizer was used, learning rate was set to 0.01, dropout equal to 0.5 and number of epochs equal to 500.

4) GraphSMOTE implementation: We have used the GraphSMOTE implementation from [16] available from the official GitHub page of the authors[7]. This model implementation expects the graph to be represented with an adjacency matrix while in a PyG dataset (ours as well) the graph is represented as a COO edge list. In order to run GraphSMOTE in our dataset we had to first convert the edge list to an adjacency matrix which was performed using Pytorch's built-in functions. The following parameter values have been used: GraphSage as embedding model, batch size was set to 40, learning rate 0.001, dropout probability equal to 0.1, epochs equal to 2000.

5) ReNode implementation: The implementation of the ReNode [17] algorithm used in this paper was adopted from the author's Github page[8]. Similar to the GraphSMOTE case also here we needed to adapt our dataset to the ReNode implementation especially in the case of the inductive settings. The main change was related to converting the edge list into a CSR adjacency matrix. Some of the parameter values used for this algorithm are: learning rate equal to 0.005, hidden layer dimensionality is set to 32, number of layers equal to 2, personalized page rank teleportation is set to 0.15.

6) Evaluation metrics: For every prediction problem, the correct evaluation of the results is very important. In the specific case of imbalanced data, choosing the appropriate evaluation metric is of critical importance. Our dataset has imbalanced data as can be seen from Imbalance Ratio, IR value from able II. Two classes Active versus Inactive users have IR values in the range from 0.037-0.062. In order to be considered balanced the dataset should have an IR equal to 1. For the above mentioned models we have used Macro F1 score and Macro recall to evaluate their performance in our dataset. **Recall** is one of the most used evaluation metrics. It gives accurate measurements regarding the detection of samples from the minority class. We can distinguish $recall^+$, also known as *sensitivity* and $recall^-$ also known as *specificity*. These two metrics can be calculated using the following formulas

$$recall^+ = \frac{TP}{TP + FN}$$

$$recall^- = \frac{TN}{TN + FP}$$

where TP is the number of true positives and FN is the number of false negatives. Macro recall is the arithmetic mean of

[7]https://github.com/TianxiangZhao/GraphSmote
[8]https://github.com/victorchen96/ReNode

recalls for different classes without considering the importance of different classes. **F1 score** is another very important metric used to evaluate the performance of machine learning models. It is defined as

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

where TP = True Positive, FP = False Positive and FN = False Negative. Macro F1 score is the arithmetic mean of F1 score of each class.

B. Results of Standard GNN Models (RQ1)

Given the extreme imbalance ratio that our data has, it is to be expected that the standard GNN models will have problems to correctly classify nodes. Our implementations of GCN, GraphSage and GAT are really good at learning really fast how to cheat so they can achieve a high accuracy. After only five iterations the model reports a high accuracy of around 95%. Looking closely at the results one can notice that the model only predicts "TRUE". By always predicting "TRUE" (the user is active) the model is certain to achieve around 95% accuracy because 95% of the users in the data is active. The reported Recall and F1 score for all these models is the same for our dataset. Sensitivity = 1 and Specificity = 0. These two metrics can be combined into Macro recall which in this case is equal to 0.5. The F1 score is equal to 0.97. After we were presented with these results, we noticed the need to test other models which are designed to handle imbalanced data.

C. Results of Specific GNN Models (RQ2)

The first specific GNN model that we tested on our data is GraphSMOTE [16]. Given the fact that GraphSMOTE is based on SMOTE [24] which in turn is one of the most used techniques in case of imbalanced non-graph data, this method is expected to perform better than other methods. After running the algorithm on our dataset, we noticed that this algorithm did classify some users as being inactive, in contrast to the standard GNN models. However, not all the predictions were correct. Calculating Sensitivity, Specificity and F1 for this method yields the following results: Sensitivity = 0.95, Specificity = 0.04, Macro recall = 0.49 and F1 score = 0.96. We notice a decrease in both these measures which gives the impression that this algorithm performs worse than the standard GNN models. One can argue that an algorithm which detects from both classes, even though not always correct, is better than an algorithm which is biased towards the majority class. Other evaluation metrics may capture this fact better than F1 score. The second algorithm that claims to handle imbalanced data and that we tested was ReNode [17]. This model classified more samples to be part of the minority class than what was expected. We ran both the inductive and transductive settings and also played with TINL and QINL settings, however the results were poor. This algorithm scored a Sensitivity of 0.37, Specificity of 0.67 and F1 score = 0.54 (see Table V).

TABLE V. EVALUATION METRICS FOR THE IMPLEMNTED MODELS

	Sensitivity	Specificity	Macro F1
GCN	1	0	0.97
GraphSage	1	0	0.97
GAT	1	0	0.97
GraphSMOTE	0.95	0.04	0.96
ReNode	0.37	0.67	0.54

V. CONCLUSIONS

In this paper we tackled the problem of fake news detection and spreading by providing a proactive solution: to detect and ban fake news spreaders before they become important spreaders. We created a dataset from November 2021 to July 2022 that contains information about the users features and relationships that spread news through tweets about vaccines. The problem of classification in imbalanced data is raised since our real data dataset was deeply imbalanced for all the periods that were taken into consideration. Three graph neural network models were trained in our datasets for node classification: Graph Convolutional Network, GraphSAGE, and Graph Attention Networks. They achieve high accuracy after few iterations because they learn really fast by predicting always True having Sensitivity = 1, Specificity = 0 and F1 = 0.97. In order to reach our goal to find and prevent the potential spreaders, increasing true negative which will yield increase in specificity becomes very important. We have tested two other frameworks to overcome the problem of imbalanced data: GraphSMOTE and ReNode. Both of these frameworks claim to give better results in imbalanced data. We found that GraphSMOTE does a better job in increasing the Specificity, but this comes at the cost of increasing the false positives rate. For this model we report Sensitivity = 0.95, Specificity = 0.04 and F1 score = 0.96. We found ReNode to not be as good as the authors claims in our case. For our dataset this technique scores Sensitivity = 0.37, Specificity = 0.67 and F1 score = 0.54.

ACKNOWLEDGMENT

This publication, is made possible with the financial support of National Agency for Scientific Research and Innovation(AKKSHI). Its content is the responsibility of the author, the opinion expressed in it is not necessarily the opinion of AKKSHI.

REFERENCES

[1] S. J. Dixon. (2023, Aug.) Number of social media users worldwide from 2017 to 2027. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[2] J. Zarocostas, "How to fight an infodemic," *The Lancet*, vol. 395, no. 10225, p. 676, Feb. 2020.

[3] R. Datta, A. Yadav, A. Singh, K. Datta, and A. Bansal, "The infodemics of covid-19 amongst healthcare professionals in india," *Medical Journal Armed Forces India*, vol. 76, no. 3, pp. 276–283, Jul. 2020.

[4] W. H. Organization. (2020, Sep.) Managing the covid-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation. [Online]. Available: https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation

[5] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, May 2017.

[6] N. Kshetri and J. Voas, "The economics of "fake news"," *IT Professional*, vol. 19, no. 6, pp. 8–12, Nov. 2017.

[7] A. Kucharski, "Study epidemiology of fake news," *Nature*, vol. 540, no. 7634, pp. 525–525, Dec. 2016.

[8] J. Golbeck, M. Mauriello, B. Auxier, K. H. Bhanushali, C. Bonk, M. A. Bouzaghrane, C. Buntain, R. Chanduka, P. Cheakalos, J. B. Everett, W. Falak, C. Gieringer, J. Graney, K. M. Hoffman, L. Huth, Z. Ma, M. Jha, M. Khan, V. Kori, E. Lewis, G. Mirano, W. T. Mohn IV, S. Mussenden, T. M. Nelson, S. Mcwillie, A. Pant, P. Shetye, R. Shrestha, A. Steinheimer, A. Subramanian, and G. Visnansky, "Fake news vs satire: A dataset and analysis," in *Proceedings of the 10th ACM Conference on Web Science*, ser. WebSci '18. ACM, May 2018.

[9] A. Ceni, A. Kika, and D. X. Çollaku, "A social network analysis of COVID-19 vaccines tweets," in *Proceedings of the 5th International Conference on Recent Trends and Applications in Computer Science and Information Technology, Tirana, Albania, April 26-27, 2023*, ser. CEUR Workshop Proceedings, E. Xhina and K. Hoxha, Eds., vol. 3402. CEUR-WS.org, 2023, pp. 1–12. [Online]. Available: https://ceur-ws.org/Vol-3402/paper01.pdf

[10] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[11] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 1024–1034. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html

[12] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[13] Z. Huang, Y. Tang, and Y. Chen, "A graph neural network-based node classification model on class-imbalanced graph data," *Knowledge-Based Systems*, vol. 244, p. 108538, May 2022.

[14] Y. Liu, Z. Zhang, Y. Liu, and Y. Zhu, "Gatsmote: Improving imbalanced node classification on graphs via attention and homophily," *Mathematics*, vol. 10, no. 11, p. 1799, May 2022.

[15] Z. Zhu, H. Xing, and Y. Xu, "Balanced neighbor exploration for semi-supervised node classification on imbalanced graph data," *Information Sciences*, vol. 631, pp. 31–44, Jun. 2023.

[16] T. Zhao, X. Zhang, and S. Wang, "Graphsmote: Imbalanced node classification on graphs with graph neural networks," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, ser. WSDM '21. ACM, Mar. 2021.

[17] D. Chen, Y. Lin, G. Zhao, X. Ren, P. Li, J. Zhou, and X. Sun, "Topology-imbalance learning for semi-supervised node classification," Oct. 2021.

[18] I. A. Pilkevych, D. L. Fedorchuk, M. P. Romanchuk, and O. M. Naumchak, "Approach to the fake news detection using the graph neural networks," *Journal of Edge Computing*, vol. 2, no. 1, pp. 24–36, May 2023.

[19] S. Feng, H. Wan, N. Wang, and M. Luo, "Botrgcn: Twitter bot detection with relational graph convolutional networks," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '21. ACM, Nov. 2021.

[20] Q. Guo, H. Xie, Y. Li, W. Ma, and C. Zhang, "Social bots detection via fusing bert and graph convolutional networks," *Symmetry*, vol. 14, no. 1, p. 30, Dec. 2021.

[21] C. Li and D. Goldwasser, "Encoding social information with graph convolutional networks forpolitical perspective detection in news media," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019.

[22] E. Shushkevich, M. Alexandrov, and J. Cardiff, "Improving multiclass classification of fake news using bert-based models and chatgpt-augmented data," *Inventions*, vol. 8, no. 5, p. 112, Sep. 2023.

[23] J. Köhler, G. K. Shahi, J. M. Struß, M. Wiegand, M. Siegel, T. Mandl, and M. Schütz, "Overview of the CLEF-2022 checkthat! lab: Task 3 on fake news detection," in *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, ser. CEUR Workshop Proceedings, G. Faggioli, N. Ferro, A. Hanbury, and M. Potthast, Eds., vol. 3180. CEUR-WS.org, 2022, pp. 404–421. [Online]. Available: https://ceur-ws.org/Vol-3180/paper-30.pdf

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002.

[25] A. M. Mahmoud, A. S. Desuky, H. F. Eid, and H. A. Ali, "Node classification with graph neural network based centrality measures and feature selection," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 2, p. 2114, Apr. 2023.

[26] H. T. Phan, N. T. Nguyen, and D. Hwang, "Fake news detection: A survey of graph neural network methods," *Applied Soft Computing*, vol. 139, p. 110235, May 2023.

Preventing Cyberbullying on Social Networks with Spanish Parental Control NLP System

Gabriel A. León-Paredes[1], Omar G. Bravo-Quezada[2], Pedro P. Bermeo-Aguaysa[3]
María J. Peláez-Currillo[4] and Ledys L. Jiménez-González[5]
Universidad Politécnica Salesiana, Cuenca, Ecuador[1,2,3,4]
Universidad Bolivariana del Ecuador, Durán, Ecuador[5]

***Abstract*—The boom in social networks and digital communication has given place to innovative forms of social interaction. However, it has also made possible new forms of harassment of others anonymously and without repercussions. Such is the case of cyberbullying, an increasingly common problem, especially among young people. Its effects on individuals can be devastating, ranging from anxiety and depression to social isolation and low self-esteem. Furthermore, there is a wide variety of applications called parental control, which allow parents to show, the pages the child or adolescent has accessed, know how often the child or adolescent accesses them, and control the time spent on social networks or other entertainment platforms. Therefore, the present research aimed to analyze, design, and implement an intelligent application based on data mining algorithms and the Latent Semantic Analysis (LSA) method for the presumed detection of cyberbullying in social networks in adolescents. The methodological process of the study was carried out following the fundamentals of applied research with a qualitative-quantitative descriptive, and cross-sectional approach. As a result, a multi-platform application was obtained that alerts about suspected bullying to parents or guardians. For the validation of the application, the technique of expert judgment was applied. Also, the process of obtaining negative and positive text similarity was performed based on cosine similarity. In the analysis of Twitter accounts, values of 46% with negative texts and 6.71% with positive texts are obtained, which allows inferring that this is a presumed case of cyberbullying in this account.**

***Keywords*—Cyberbullying; control parental system; natural language processing; Spanish cyberbullying prevention system**

I. Introduction

Social networks have transformed the way people communicate and interact with each other, and have made it possible to instantly connect with people all over the world. Undoubtedly, the era of digital transformation has created new opportunities in various areas of society; for education, commerce, and culture, and has allowed people to connect and share ideas in ways never before imagined. In this sense, it is valid to say that there are many satisfactions experienced by users of digital media with the advancement of technologies, although positive experiences do not always occur given the appearance of malicious users that cause negative effects on people. Among these negative effects is cyberbullying on social networks; an increasingly common problem. It refers to bullying, harassment, intimidation, and other forms of aggressive behavior that take place online through social networks and other digital platforms. The effects of cyberbullying can be devastating, from anxiety and depression to social isolation and low self-esteem [1].

Today, everyone can be a victim of cyberbullying, but young people are more vulnerable [2] for example, adolescents are more likely than adults to spend extended time online and be more involved in social networks. This means they have greater opportunities to interact with other users, including people who may have bad intentions. Also, young people often lack the emotional maturity and experience to deal with cyberbullying situations [3]. Often, they may feel embarrassed or scared to talk to someone about bullying, which can make the situation worse. In addition, they are more likely to make impulsive or risky decisions in online situations, which may increase their vulnerability to victimization in digital media.

Another factor contributing to young people's vulnerability to cyberbullying is the lack of adult supervision. Parents often do not have a full understanding of online platforms and the social interactions their children have with them, which can make it more difficult for them to detect signs of cyberbullying or to intervene before situations become serious [4]. Also, it should be considered that many of these young people use these platforms as their primary means of communication and socialization. However, they are unprepared to deal with cyberbullying, feeling alone and isolated if they do not have the support of their parents and friends. Therefore, cyberbullying can have a very negative impact on the social and emotional lives of adolescents.

As a serious problem that can lead to significant emotional and psychological consequences, parents should rely on technological advances to limit the exposure of adolescents to inappropriate content. It is important to note that these consequences are not exhaustive, as the experience of each victim of cyberbullying may be different. However, their impact underscores the importance of creating strategies to prevent and address bullying in any space. In this sense, the study carried out by the Ministry of Education of Ecuador with World Vision and UNICEF, in which 5,511 students from 126 institutions were consulted, states that in Ecuador 6 out of 10 students have been victims of bullying and the use of social networks. The students surveyed state that they have been victims of the dissemination of messages; in some cases to reveal private conversations and in other cases to generate threats by anonymous users [5].

School violence has been silenced, not because it has been solved, but because it now occurs in digital scenarios, which has generated new terms to be taken into account by adults, such as the well-known cyberbullying, a form of harassment based on the use of Information and Communication Tech-

nologies (ICTs) [6], [7]. In addition to social networks, it occurs in message-sharing platforms, video games, and blogs, among others [8]. Therefore, the diversification of online environments makes it difficult for adults to monitor them. This happens, first, because most parents are not tech-savvy. Secondly, because parents or guardians do not have enough time to carry out a constant review of the new media; and, thirdly, in many cases young people keep this type of situation secret. By the above, adults have been forced to adapt digitally in a fast way to the new generations to carry out adequate supervision [9].

In this sense, technological progress has allowed the development of intervention alternatives that facilitate parental control processes. Parental control systems are nowadays a viable alternative for parental intervention in the protection of adolescents from bullying on social networks. Likewise, it can help identify and prevent bullying situations, providing parents with the ability to monitor and control the use of digital environments and prevent cyberbullying [10]. However, these apps generally offer features such as monitoring browsing history, blocking inappropriate websites, monitoring messages and phone calls, and setting time limits for device use [10]. Hence, there is limited research related to parental performance in the mechanisms of prevention and intervention in the face of cyberbullying [11].

Among the limitations of existing applications is that they do not focus on analyzing the activities that adolescents carry out on their social networks to determine whether any action taken against an adolescent is alleged cyberbullying [12]. In this regard, [6] argues that the use of Natural Language Processing (NLP), as a branch of Artificial Intelligence (AI), in applications for parental control, is part of the technological innovation of recent times in the field of online protection. Among its potentialities is the ability to understand and process human language with high levels of similarity to that of people. Over the last few years, it has been applied in social network analysis to identify topics, sentiments, and other characteristics of natural text [13]. In addition, it contains a variety of tools and techniques to extract valuable information in large volumes of texts, identify trends, and determine opinion patterns, preferences, and user behaviors [14]. These techniques find functionality through the creation of algorithms that allow the analysis of specific information and obtaining accurate results in real-time [15]. One such tool is the Latent Semantic Analysis (LSA) method, which uses mathematical and statistical techniques to create a numerical representation of words and documents in a corpus. This numerical representation allows patterns to be identified in the way words are used in different documents and how they are related to each other [16], [17].

Despite its innumerable advantages, the documentary review evidences a limited availability of NLP applications developed to detect cyberbullying in the Spanish language. Likewise, NLP uses the LSA method for the identification of semantic patterns and relationships in sets of texts. Although the usefulness of LSA has been demonstrated in different NLP applications and has significant contributions in language pattern identification, topic detection, sentiment analysis, and language variant detection; its use in cyberbullying detection applications in social networks is also limited. According to [18], LSA is useful in a wide variety of applications, because of its ability for information retrieval, automatic document classification, topic clustering, and content recommendation. In addition, it is used in the construction of search engines to identify relevant information in large text datasets.

Attention to such, this paper presents the results of the research developed to analyze, design, and implement an application based on NLP for the prevention of presumptive cyberbullying in the social networks of Ecuadorian adolescents. As a contribution, a multiplatform application for parental control based on NLP in Spanish and using the Latent Semantic Analysis method is presented. The proposal responds to the need to protect Spanish-speaking adolescents from online cyberbullying and the scarce production of solutions based on these technologies in Spanish. Taking advantage of the potential of NLP and semantic analysis made it possible to develop an effective tool to detect and prevent cyberbullying in the linguistic and cultural context of the Spanish language, giving parents and guardians the ability to protect adolescents in digital environments.

To effectively communicate the findings and contributions of the study on the multiplatform application of parental control through Natural Language Processing in the Spanish language. This paper is organized as follows. In Section II, we present some studies related to the use of information and communication technologies for the detection and prevention of cyberbullying, highlighting the limitations of the most recent studies. Section III, outlines the methodological components of the research, as well as the phases guiding the development of the Intelligent Parental Control System. In the Results Section IV, the data obtained in the implementation of the mobile and web applications are detailed. In the Discussion Section V, the results are interpreted in a broad context of cyberbullying in social networks, comparing the results of related works with the challenges for the practical implementation of the designed application. Finally, in Section VI, conclusions derived from the study are presented, highlighting the contributions of the NLP-based application in Spanish for the prevention of presumptive cyberbullying in the social networks of adolescents.

II. Related Works

The following are the results of the main research developed to prevent cyberbullying using advances in Information and Communication Technologies.

Researchers from the National University of the USA developed the SafeGuard Web application, intending to detect dangerous situations in educational institutions based on social networks [19]. By constantly monitoring publications on social networks, SafeGuard determines the relationship of the language users use with pre-established keywords such as suicide, death, violence, and so on. When these terms are identified, the application sends alert messages to the system administration, thus protecting potential victims from any emotional or physical harm. SafeGuard's configuration employs web technologies with hosting in the cloud environment, allowing login from any location. In terms of operation, SafeGuard employs IP address-based access restrictions for client login over the network and uses a JSON Web token with expiration times. To respond to client requests, the application uses the HTTPS protocol. The results of its implementation show the application's ability to

detect cyberbullying, threats, or distress situations through the configuration of monitors or keywords. However, it requires adding weighting factors to the monitors that favor the search for particular keywords by the weighting assigned to them.

The authors of the study [20] analyzed the increase in cyberbullying over the last 15 years, and came up with BullyScan. It is a novel framework based on natural language processing and machine learning with the ability to identify online bullying automatically with high accuracy and efficiency. The application employs a logistic regression algorithm developed from the validation and testing of five different machine-learning models with the combination of three datasets of cyberbullying and/or hate speech. The tests conducted demonstrated 92% accuracy in detecting cyberbullying in real time, evidencing the ability to significantly reduce cyberbullying rates and increase positivity on social networks, as well as prevent the consequences of cyberbullying.

On the other hand, [21] proposes a system that employs PLN and machine learning to detect bullying related to messages in digital environments in the English language. For the research, 1651 tweets were collected and applied to a PLN approach to identify the most offensive terms related to cyberbullying. Using the obtained dataset, Random Forest (RF) and Support Vector Machine (SVM) algorithms were trained. The former outperforms the latter with an accuracy of 98.5%. The analysis of the results was performed using the Root Mean Square Error (RMSE) and the Mean Square Error (MSE). The RF algorithm scored better than the SVM. The results show the existence of cyberbullying and the need to address it immediately.

Similarly, the authors [22], agree that cyberbullying is an extremely prevalent issue at the moment; as access to social platforms increases, messages of hate, toxicity, and cyberbullying. In this sense, it is fundamental to generate mechanisms that guarantee the security of social networks and that any form of violence or hate crime can be automatically detected. Based on this, they proposed the analysis of cyberbullying through natural language processing employing the compression of the use of slang in social networks. As a result of their research, they achieved greater accuracy in identifying online harassment situations through experiments with multiple models such as Bi-LSTM, GloVe, and BERT and the application of the unique processing technique for the incorporation of an abusive corpus of slang. The model demonstrates greater effectiveness than models that do not contain slang preprocessing.

In a more comprehensive perspective, [23] developed an NLP tool that uses the social network Twitter as a basis for the extraction of information related to cyberbullying. The methodological procedure followed consisted of analyzing a set of tweets with the SARNA technique and classifying them based on their content as cyberbullying or neutral. The authors created a labeling system for people to classify the tweets using a reliability scale from 1 to 4, where ratings 1 and 2 indicate that they are non-cyberbullying tweets, while 3 and 4 refer to tweets with cyberbullying content. For the classification process, the BERT model was used, which was trained to identify aggressive, toxic, or threatening comments with label 1 and neutral comments with label 0. The results obtained demonstrate the importance of providing an adequate knowledge base, training the supervised learning model, and conducting case studies to accurately detect cyberbullying using NLP techniques; nevertheless, the authors do not detail the data obtained during the process.

In the same context, the Salesian Polytechnic University has promoted the development of applications that contribute to the reduction of the effects of this problem. The authors in [6], which some authors are also authors of this paper have deployed a Cyberbullying Prevention System (CPS) in Spanish based on Natural Language Processing and the use of Machine Learning techniques such as Naive Bayes, Support Vector Machine, and Logistic Regression. As in the previously mentioned research, the social network Twitter was used for the extraction of the database or corpus. As for the training process, it consisted of the use of precision metrics, and corpus sizes with variability. The level of accuracy of the SPC system was validated with the application in three case studies, obtaining a 93% of reliability.

Furthermore, according to [24] most of the research developed to detect bullying on social network platforms is based on machine learning models that use datasets extracted from individual social networks. Therefore, they have proposed a cross-platform data system that uses text collected from posts made on seven social networks. They propose an annotation system composed of a series of stages and techniques to identify posts and hashtags through crowdsourcing, and then identify posts that require annotation through machine learning methods. The advantage of the presented model lies in the possibility of cyberbullying cases and the limitation of the particular characteristics of the publications, unlike traditional methods based on post-selection and tagging. The training process of the models on the diverse dataset evidences a good performance and allows for an increase in the number of positive examples with the same amount of resources and the applicability of the models in different media.

Thus, although research related to cyberbullying has advanced in recent years, there are still great challenges to be faced. Among these challenges is the need to address the issue in a standardized manner [25] and the generation of applications that consider the linguistic and cultural context in the automatic detection of situations related to cyberbullying. In addition, the aforementioned related works differ from our proposal, since to our knowledge and according to the analysis of the state of the art, there are no applications proposed at the international and even more at a local level that use the Latent Semantic Analysis method for the detection and prevention of cyberbullying in social networks. In Table I, we present a comparative analysis of the cyberbullying detection studies presented before and we have included the present research.

III. Parental Control Multiplatform System for the Prevention of Cyberbullying

The methodological framework followed for this research was based on the principles of applied research whose purpose is to contribute to the solution of society's problems through the application of knowledge and tools of a specific scientific discipline. In this sense, the research is focused on the (a) design and (b) implementation of a multiplatform system based on Natural Language Processing for the prevention of pre-

TABLE I. COMPARATIVE ANALYSIS OF CYBERBULLYING DETECTION STUDIES HIGHLIGHTING METHODOLOGIES, KEY FEATURES, OUTCOMES, AND UNIQUE CONTRIBUTIONS OF EACH STUDY, INCLUDING THE PRESENT RESEARCH

Study	Methodology	Key Features	Outcomes/Accuracy	Unique Contributions
SafeGuard (Wyne, 2021)	Web application, keyword monitoring	Cloud hosting, IP address restrictions, HTTPS protocol	Effective in detecting cyberbullying via keyword monitoring	Focus on educational institutions, requires keyword weighting
BullyScan (Shrimali, 2022)	NLP & ML, Logistic Regression	Combination of three datasets, real-time detection	92% accuracy	High accuracy and efficiency in real-time detection
Afrifa et al. (2022)	NLP & ML, RF and SVM algorithms	Analysis of 1651 tweets, RMSE and MSE evaluation	98.5% accuracy with RF algorithm	High accuracy, focus on English language messages
Bhatia et al. (2022)	NLP, Bi-LSTM, GloVe, BERT	Slang preprocessing, abusive corpus analysis	High accuracy in identifying online harassment	Effectiveness in slang and abusive language detection
Soto et al. (2022)	NLP, BERT model	SARNA technique, labeling system for tweets	Not detailed	Importance of adequate knowledge base and training
CPS (Leon Paredes, 2019)	NLP & ML, Naive Bayes, SVM, Logistic Regression	Twitter data extraction, precision metrics	93% reliability	Focus on Spanish language, high reliability
Van et al. (2020)	Cross-platform data system, ML	Posts from seven social networks, crowdsourcing annotation	Good performance	Cross-platform applicability, diverse dataset
Present study	NLP & LSA, cosine similarity analysis	Focus on Spanish language, multiplatform application, analysis of Facebook, Ask.fm, and Twitter	46% with negative texts and 6.71% with positive texts	Unique use of LSA for cyberbullying detection in Spanish

sumptive cyberbullying in the social networks of adolescents focused in the city of Cuenca - Ecuador.

On the one hand, we detail some important points related to the (a) design of the Parental Control Multiplatform System for the Prevention of Cyberbullying based on NLP. To implement an accurate design, we need to know how young people (students between 18 and 24 years old) of the Salesian Polytechnic University use digital media to determine the main problems of cyberbullying in social networks in our local context. Hence, a structured survey called "Survey of Safety and Cyberbullying in Social Networks" was applied. The survey consisted of 20 items with dichotomous and polytomous response options. The instrument with the questions was applied to 133 students enrolled in the Computer Science program. The size of the probabilistic sample was defined considering Eq. (1) for a finite population. The total population was 5,576 on-campus students from all UPS courses. The procedure for the selection of the sample is indicated below,

$$n = \frac{N\sigma^2 Z^2}{(N-1)e^2 + \sigma^2 Z^2} \tag{1}$$

where, *n*, is equal to the size of the sample; *N*, is equal to the size of the population; σ, is equal to the standard deviation of the population, when a value is not available the constant value of 0.5 is used; *Z*, is obtained by confidence levels, in case of not having a value 95% of confidence is placed and this is equivalent to 1.96 (it is the one commonly used); *e*, is equal to the acceptable limit of the sampling error, when a value is not available a range from 1% (0.01) and 9% (0.09) is used, this value depends on the interviewer.

The process of tabulating the results obtained through the survey was crucial within the methodology proposed for the design of the Parental Control Multiplatform System, so we highlight the most important results below. The analysis of the results shows that 88% of the participants are between 18 and 24 years of age, 86% are male, and 100% use social networks. Regarding the use of social networks by the participants, 9% stated that they use social networks to meet new people, while 37% use them for entertainment, 16% to communicate with people who live in different places, 36% to communicate with friends, family, acquaintances, among others, and only 2% responded that they use the networks for other situations such as business. The data obtained show that most of the young students surveyed use social networks to communicate with other people.

Furthermore, when participants were asked about the time they spend using social networks, 2% use social networks for less than one hour, 31% use them for one to two hours a day, 41% use them for three to four hours a day, 17% use them for five to six hours a day, 4% use them for seven to eight hours a day, and 5% use social networks for more than nine hours a day. In this sense, it is evident that most of the participants use social networks for several hours during the day. Regarding the most used social networks, 23% of the participants stated that they use WhatsApp, 22% use Facebook, 8% use Twitter, 1% use Tumblr, 23% use YouTube, 20% use Instagram, and 3% use Pinterest. Thus, the majority of the surveyed students mostly use social networks such as WhatsApp, Facebook, YouTube, and Instagram.

An important fact to be considered is that 46% of participants confirm having received cyberbullying messages. Regarding the social networks in which they have sent, known, or received messages of cyberbullying, 15% indicate WhatsApp, 38% indicate Facebook, 4% indicate Twitter, 11% indicate Instagram, 1% indicate that Snapchat, Tinder, and Badoo are social networks where this type of problem happens, 2% indicate that YouTube is the social network where messages of cyberbullying are sent, known or received, 1% indicate that in no social network this problem happens, and 27% did not respond. Thus, most of the participants who send, know, or receive cyberbullying messages have received it at least once on any of the social networks WhatsApp, Facebook, Instagram, Twitter, Snapchat, Tinder, and YouTube. Finally, it is important to highlight that 95% of participants say that if they had a computer tool to prevent and detect cyberbullying, they would use it.

Hence, based on this diagnosis, we have determined the features that the Parental Control Multiplatform System proposed in this paper should have. First, we have designed it to be used by parents, guardians, teachers, and psychologists,

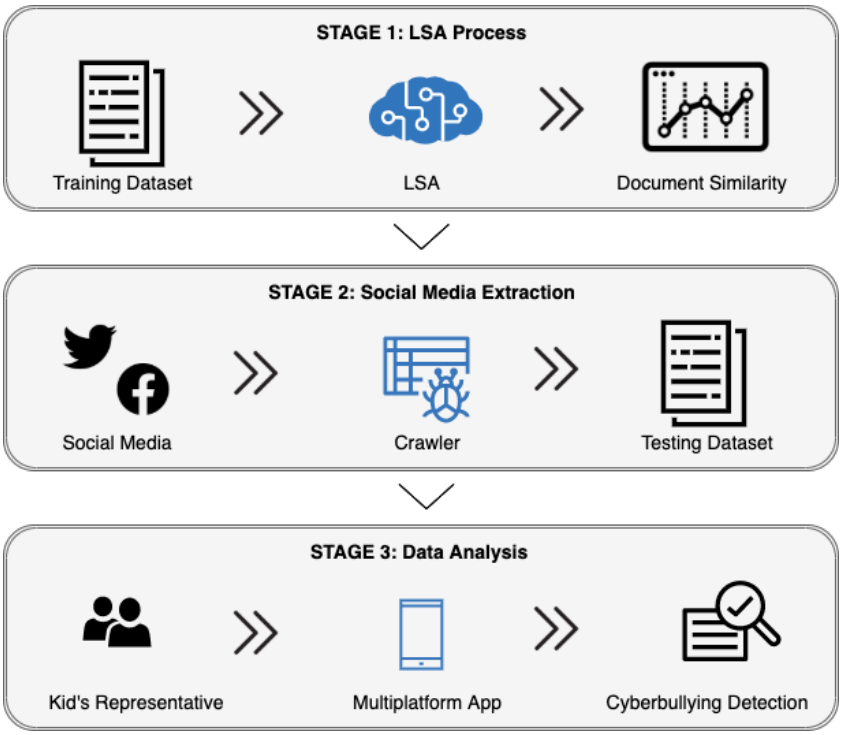


Fig. 1. Stages followed for the implementation of the parental control multiplatform system.

among others, and to be aware of the type of information that adolescents receive through social networks. We have focused on young people who have initiated using these platforms and do not have the knowledge to deal with bullying situations. Also, the Parental Control System was designed to work both in web environments and mobile devices, through the creation of a multiplatform system. It is necessary to note that, for web environments, the application has features that differ from the version for mobile devices, mainly due to the incorporation of extra functionalities.

On the other hand, we detail the (b) implementation of the Parental Control Multiplatform System based on Natural Language Processing through the Latent Semantic Analysis (LSA) method for preventing presumptive cases of cyberbullying in social networks. The system has been divided into three main stages, as shown in Fig. 1.

For the first stage of system implementation, we utilize the Latent Semantic Analysis (LSA) method, which allowed us to obtain a base training model for detecting cases of presumptive cyberbullying in several social networks. At this point, it is important to indicate that this paper is oriented to the Natural Language Processing of the Spanish language with an emphasis on social network users from Ecuador. So, we started by obtaining a Spanish dataset related to cases of cyberbullying in Ecuador. We need this specific type of dataset due to the slang used among adolescents. Thus, we worked with a dataset referred to a previous paper published by some authors of this work, as shown in the research document "*Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language*".

As mentioned in study [6], the dataset is compounded with a total of 960,578 tweets, of which 416,567 correspond to presumptive cyberbullying. Then, all the tweet's text data was cleaned up by using text processing techniques such as the removal of stopwords, lemmatizing, and stemming of the remaining words. After the processing of the tweet's text data, we constructed the Term-by-Document matrix, and then we reduced this matrix to 300 k dimensions by applying the Singular Value Decomposition (SVD) truncated method [26].

At this point, it is important to clarify how we used the Latent Semantic Analysis method for the presumptive detection of cyberbullying. As mentioned, we employed a dataset of 416,567 tweets with plausible cyberbullying, and 544,011 tweets with no presumptive cyberbullying to train our knowledge base. Then, each comment of a post made on the social network of the adolescent is compared against the trained dataset (cyberbullying, and no cyberbullying). Therefore, a similarity value is obtained for each trained document from the dataset versus the comment post issued on the social network. Next, these values are ordered from highest to lowest. We left with the top 10 values, to subsequently obtain a general similarity value of the comment posted on the social network by applying the following equation,

$$SimPos = \frac{\sum_{i=1}^{10} SimDoc_i}{10} \quad (2)$$

where, *SimPos* is equal to the general similarity of the comment posted on the social network, *SimDoc* is equal to the similarity obtained between each document of the trained dataset and the comment posted on the social network. Finally, the value of *SimPos* is then evaluated between the general

similarity obtained of the trained dataset with presumptive cyberbullying and the trained dataset with no presumptive cyberbullying, which allows us to determine whether the comment post contains presumptive cyberbullying or not. For our case studies, if the general similarity of the trained dataset with presumptive cyberbullying is greater than the general similarity of the no presumptive cyberbullying dataset, then it can be inferred that the comment contains presumptive cyberbullying.

In the second stage of the implementation of the Parental Control Multiplatform System, we crawled the text data from social networks, in this case, Twitter, Facebook, and AskFM. First, the system needs the credentials of the Facebook and AskFM social network user, to extract the text interaction of its account. Thus, for these cases, we utilized frameworks, that can automate the web browser navigation throughout the execution of programming scripts. However, for the case of Twitter, we used the official API, which can permit us to extract the text tweet data of a specific user when its account is public. Hence, for this case, we don't need the user credentials, only a key API consumer of a Twitter developer account.

Moreover, the data crawled from the social network has to be extracted constantly to detect them as soon as possible the presumptive cyberbullying. Consequently, we proposed and developed automated tasks on the server side to extract the social network text data using the crawlers specified in the previous paragraph. These automated tasks can be executed hourly, daily, weekly, or monthly depending on the configuration each user makes on the system.

In the last stage, we developed the multiplatform (progressive web and mobile) system. Some of the functionalities created have been focused on registering new users, authenticating and authorizing user accounts, recovering account passwords, registering and authorizing the credentials of the adolescent's social network accounts, crawling the text data from the registered and authorized social network accounts, scheduling when the text data is extracted, and presenting the analysis results of the presumptive or not cyberbullying, as shown in Fig. 2, and 4. In addition, as part of raising awareness of the cyberbullying problem, we added some relevant documentation (articles, videos, and psychologist experts' contacts), as shown in Fig. 3.

Finally, taking into account these three phases in the implementation of the Parental Control Multiplatform System, its "normal" functionality is explained in greater detail below. Then, as a first step, the parent or guardian responsible for the child or adolescent must register as a user within the system, for which personal information such as identification number, full name, address, telephone, email, and password are requested. Next, they must go through a process of verification of their personal data, for which an email is sent to the guardian indicating the steps for activating their account in the system.

With these previous steps completed, the new user will be able to log in and then have the option of registering different social network accounts of his/her tutored. At this point, it is important to indicate that, when registering Facebook and ASK FM social network accounts, the username and password of the adolescent's account must be entered so that this information is sent to the crawlers and the text information can be extracted.

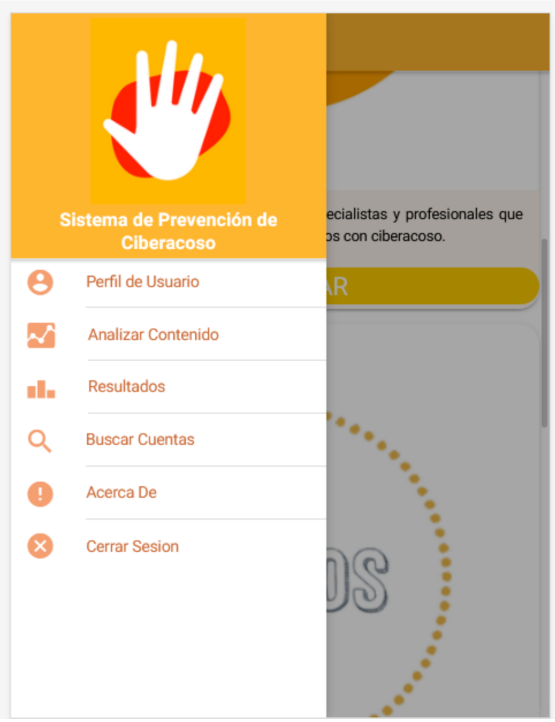


Fig. 2. Graphical user interface of the functionalities that the user of the parental control multiplatform system has access to.

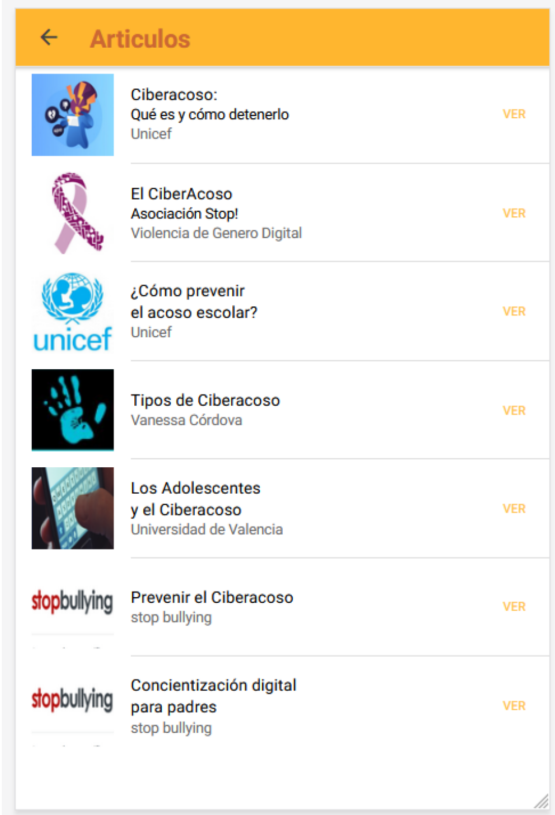


Fig. 3. Graphical user interface of the relevant documentation related to cyberbullying presented in the parental control multiplatform system.

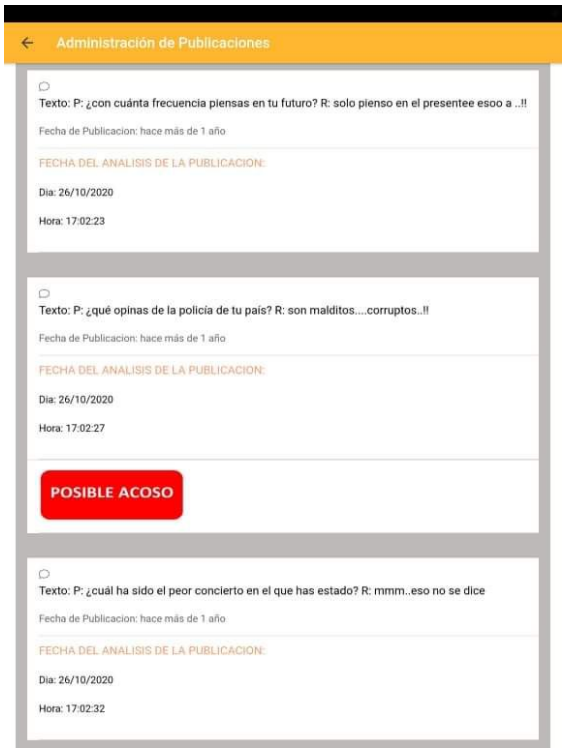


Fig. 4. Graphical user interface of the list of publications from the social network adolescents' accounts in the parental control multiplatform system.

Once the corresponding user and account registrations have been made, the frequency of analysis to be performed on the accounts can be configured. This analysis can be hourly, daily, weekly, or monthly and is done through automated tasks on the server side. However, there is the possibility that the user from the system can run a scan at any time. This allows the web services to be consumed to extract the information from the social networks, and subsequently analyze them. Regardless of which of the two analysis options is selected, once the system completes the analysis, a message is sent to the user's email indicating that the analysis is complete.

IV. Results

Below, we present the results of the experiments carried out with the Parental Control Multiplatform System for the prevention of presumed cases of cyberbullying.

A. Results of the Analysis of Accounts from Different Social Networks using the Latent Semantic Analysis Method

This sub-section presents the results of the analysis of social network accounts that have been specifically selected for testing and validation. The objective of these experiments is to test the performance of the LSA method. So, these selected accounts have published content considered as alleged harassment, even some accounts belonging to a well-known case in Cuenca - Ecuador, where the implicated user faced legal problems and is currently serving a sentence. On the other hand, accounts of public figures that publish messages of peace and love considered as "without harassment" have been used. The validation procedure consisted of first extracting the information from the social network accounts and then analyzing this information using the LSA method. As a final phase, the results were presented with the percentage of similarity, both in texts with presumptive cyberbullying (negative) and without presumptive cyberbullying (positive) as can be seen in Fig. 4, which shows an example of how the analysis results are presented in the Parental Control Multiplatform System. Then, for this experiment, two Twitter accounts were analyzed.

For this experiment, two Twitter accounts were analyzed. In the first account, we analyzed posts issued by a known user in the city of Cuenca - Ecuador who committed serious crimes against adolescent women, and his main communication media were social networks, where he hooked his victims, for which he faced legal problems and is currently serving a sentence. In this first account analyzed, although the total value of the similarity of "positive" texts prevailed over "negative" texts, the algorithm evidences the existence of texts with assumed cyberbullying as some examples can be seen in Table II. The analysis reveals the ability of the Parental Control Multiplatform System designed to identify terms that suggest harassment, manifested through swearing, insults, words with high levels of aggressiveness or derogatory phrases, and even threats. An example of a threat identified by the algorithm is the phrase *"a cada PUERCA le llega su carnival"*, which means in English "every PIG has its own carnival". It should be noted that in the local context, a carnival is a local holiday where families come together to eat a pig. Therefore, in this post, the user refers to his victims as animals. In this sense, the documentary review argues that it is completely normal for a social network stalker to use phrases with insinuations to threaten his victim [25]. This is because, cyberbullying can take different forms, including spreading rumors, posting humiliating comments, or direct threats, among others. The insinuations can be used as a form of psychological harassment [27], which seeks to destabilize the victim and make them feel uncomfortable or insecure.

Likewise, the phrase "When you are cold look for me", detected by the algorithm as a form of cyberbullying, might seem like an innocent offer, but it is designed to make the victim feel observed or surveilled. This is explained by the fact that in some cases, the cyberbully may use subtle or indirect language to avoid detection by security filters or by the victim's parents and guardians [28]. These insinuations may seem harmless or even flattering at first glance. Therefore, the difference between the percentage of negative and positive text is minimal. Nevertheless, the system determines that it is a threatening or intimidating message that should alert parents or guardians.

The second Twitter account analyzed was the account of Pope Francis, Supreme Pontiff of the Catholic Church; to establish a comparison of the percentages of similarity between the analyzed account with presumed cyberbullying publications and the @Pontifex_es account. As shown in Table III, the percentage of similarity of "positive" text is well above the "negative" text. This indicates that there is a minimal percentage of text linked to cyberbullying. The Parental Control Multiplatform System developed in this proposal has obtained alarms from this account on posts where terms such as "war", "weapons", "pain", and "hurt" appear. In this sense, although the main function of the LSA method is to identify patterns

TABLE II. SAMPLES OF THE RESULTS OBTAINED BY THE ALGORITHM OF THE PARENTAL CONTROL MULTIPLATFORM SYSTEM WHEN ANALYZING THE FIRST TWITTER ACCOUNT OF THE USER WHO HAS BEEN SENTENCED IN THE CITY OF CUENCA - ECUADOR FOR CRIMES WHERE SOCIAL MEDIA WAS HIS MAIN COMMUNICATION MEDIA

QUERY	% SIMILARITY NEGATIVE TEXT	% SIMILARITY POSITIVE TEXT	CYBERBULLYING?
Ustedes no son nada mas que P"t's Del Cabaret,	35.45	11.07	Yes
Prefiero que sean p"t's pero no mentirosas	46.59	06.71	Yes
Una zorra se merece tu ver..@,,, Una dama tu corazón!	19.17	11.30	Yes
En el cielo se guardan nuestros secretos,	29.12	51.20	No
Hasme el amor pero de tu vida,	09.31	17.25	No
Aprecia lo que la vida te da, porque no te da dos veces,,,	17.14	18.04	No
Cuando tengas frío búscame,	27.55	25.26	Yes
Aprecia lo que la vida te da, porque no te da dos veces,,,	17.14	18.04	No
Si te estorba la virginidad yo soy experto en curar esos males y estorbos,	11.16	10.79	Yes
A cada PUERCA le llega su Carnaval	14.97	8.11	Yes
Lo que aprendí de Mario Bross es que mientras mas moneditas tengas, es más fácil llegar a la princesa,	13.00	17.53	No
No confío en tus palabras, que se las lleve el viento muy lejos del lugar en el que yo me encuentre,	07.18	09.94	No
Tienes que saber diferenciar para que te quería, para amarte o solo para TIRARTE, Y siempre fuiste lo SEGUNDO,,,	30.55	42.61	No

and semantic relationships between words and documents in a text corpus, it is important to keep in mind that latent semantic analysis is not always able to capture the full context and emotional connotations of a text [16], [17]. Therefore, it becomes necessary to critically read the analysis and consider the full context before drawing conclusions.

Another experiment conducted in this research was with the Ask.fm social network, a question-and-answer-based social platform that allows users to interact anonymously or in an identity-based mode. This platform was launched in 2010 and has become very popular among teenagers and young adults as a way to ask funny, curious, or personal questions to friends and strangers. To validate the algorithm of the Parental Control Multiplatform System, two Ask.fm accounts were created. The first account created has username *mapesystems2*, and is an account that has been created to disseminate messages with pretended harassment, as can be seen in Fig. 5. In this account, the total similarity value of "negative" texts was well above the similarity of "positive" texts, which evidences a large number of texts with presumptive cyberbullying.

The second analyzed account of the Ask.fm social network belonged to the user *mapesystems4344*, which was also created to validate the algorithm of the Parental Control Multiplatform System, and unlike the previous account, its objective consisted of posting messages without "much" presumed harassing content, as can be seen in Fig. 6. Unlike the first analyzed account, in this account, the total value of "positive" text similarity prevails over "negative" text similarity. This indicates that in this account there is a lower amount of texts with cyberbullying.

B. Expert Validation of the Parental Control Multiplatform System

It is important to mention that, for the validation process, the collaboration of psychologists, and specialists in the area of cyberbullying, as well as students and parents or guardians was requested. For the validation of the Web application, first, an explanation of the system's functionalities was given user registration; social network account creation, validation; analysis, and detection of presumptive cyberbullying; among other

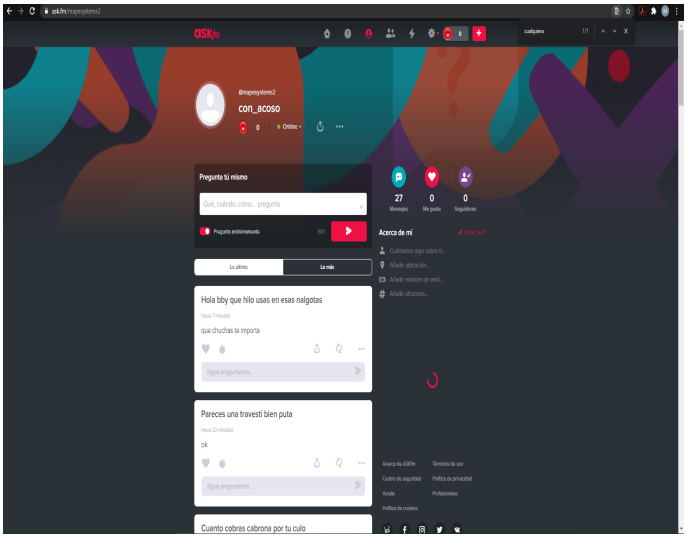


Fig. 5. Sample messages posted on the Ask.fm social network by the user *mapesystems2*.

functionalities. Once the whole process was explained, the following link https://cloudcomputing.ups.edu.ec/controlParental/ was sent to them so that they could interact with the system and perform the cyberbullying analysis on different social network accounts.

Once the presentation and validation of the web application were finished, the functionalities available in the mobile application were explained, such as viewing the results of the analyzed accounts, checking account information, and updating user data. After the demonstration, the psychologist and the guardians downloaded[1] the mobile application from the Google Play Store. To assess the effectiveness and adequacy of our mobile and web applications in detecting suspected cyberbullying cases, a group of experts conducted testing. Following this, we distributed a questionnaire to a diverse

[1]At the moment of writing this paper and after being published for several months, the application has been unpublished due to Google's policies.

TABLE III. SAMPLES OF THE RESULTS OBTAINED BY THE ALGORITHM OF THE PARENTAL CONTROL MULTIPLATFORM SYSTEM WHEN ANALYZING THE @PONTIFEX_ES TWITTER ACCOUNT

QUERY	% SIMILARITY NEGATIVE TEXT	% SIMILARITY POSITIVE TEXT	CYBERBULLYING?
"La oración es el centro de la vida, Si hay oración, también el hermano, la hermana, se vuelve importante, Quien adora a Dios, ama a sus hijos, Quien respeta a Dios, respeta a los seres humanos," $\#AudienciaGeneral@Pontifex_es$, 2012,	18.02	24.26	No
"La pertenencia a Cristo y el estilo de vida que se deriva de ella no aíslan al creyente del mundo; por el contrario, lo hacen protagonista de un servicio de amor en favor del bien común," $@Pontifex_es$, 2012	25.12	36.93	No
"¿Una decisión valiente? Destinar el dinero utilizado para las armas a un "Fondo mundial" para acabar con el hambre, Esto evitaría muchas guerras y la emigración de muchos hermanos y hermanas nuestras de los países más pobres, #JornadaMundialAlimentación" (@Pontifex_es, 2012)	14.45	13.73	Yes
"Es tiempo de suscribir un pacto educativo global por y con las jóvenes generaciones, un pacto que comprometa a familias, comunidades, escuelas, universidades, religiones, instituciones, gobernantes, a la humanidad entera, para formar personas maduras, #GlobalCompactOnEducation" @pontifex_2012 ,	14.98	22.37	No
"María, la madre que cuidó a Jesús, también cuida con afecto y dolor materno este mundo herido," (@Pontifex_es, 2012)	20.41	14.88	Yes
"El mundo es algo más que un problema a resolver, es un misterio gozoso que contemplamos con jubilosa alabanza, #TiempoDeLaCreación #LaudatoSì" (@Pontifex_es, 2012)	05.75	07.87	No

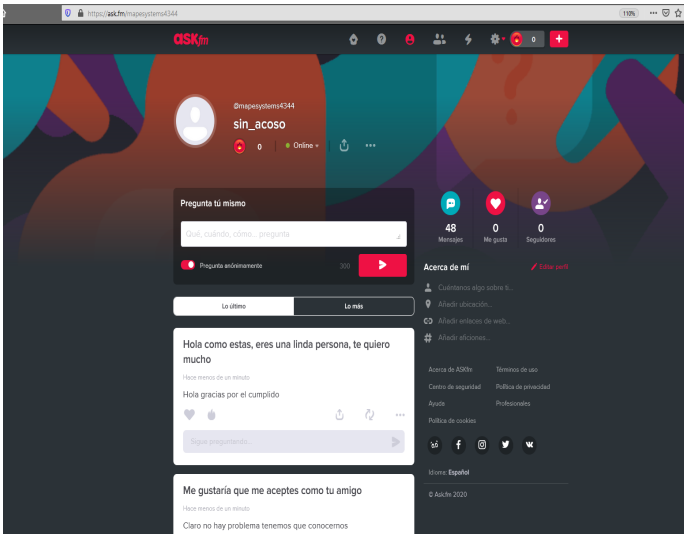


Fig. 6. Sample messages posted on the Ask.fm social network by the user *mapesystems4344*.

group of 10 users, including parents, guardians, students, and psychologists. The questionnaire aimed to gather feedback on the application's features and determine its overall effectiveness. The most relevant results of the survey are presented below.

As evidenced in Table IV, 50% of the surveyed users consider that how the results of the analyses are presented in the web application is understandable and 50% consider that how the results are presented in the application is easy to understand. This shows that all participants understood and comprehended how the information is displayed.

Furthermore, Table V shows that 60% of the surveyed

TABLE IV. DOES THE WEB APPLICATION EFFECTIVELY PRESENT THE ANALYSIS RESULTS IN A CLEAR AND COMPREHENSIBLE MANNER?

	Users	Answer
Absolutely understandable	5	50%
Easy to understand	5	50%
Absolutely difficult to understand	0	0%

users consider that the mobile application has an excellent appearance, and 50% consider that it has a good appearance in terms of colors, images, icons, and visibility. This shows that all respondents consider that the mobile application has an adequate appearance and meets their needs.

TABLE V. WHAT DID YOU THINK OF THE APPEARANCE (COLORS, IMAGES, ICONS, VISIBILITY) OF THE MOBILE APPLICATION?

	Users	Answer
Excellent appearance	6	60%
Good appearance	4	40%
Bad appearance	0	0%

In that order, users were asked whether they consider how the results of the analyses are presented in the mobile application to be understandable. According to the results presented in Table VI, 60% of the surveyed users consider it to be understandable and 40% consider it to be easy to understand. This means that all participants understand and can comprehend the presentation of the results.

Table VII, shows that 88.9% of surveyed users consider the security of the application to be excellent and 11.1% consider it to be regular. This means that most of the participants consider that both the web and mobile applications have good security in terms of the information handled, since it is confidential between each user.

TABLE VI. DOES THE MOBILE APPLICATION EFFECTIVELY COMMUNICATE THE ANALYSIS RESULTS UNDERSTANDABLY?

	Users	Answer
Absolutely understandable	6	60%
Easy to understand	4	40%
Absolutely difficult to understand	0	0%

TABLE VII. HOW SATISFIED ARE YOU WITH THE SECURITY OF THE APPLICATION?

	Users	Answer
Excellent	9	88.9%
Regular	1	11.1%
Bad	0	0%

Regarding the ease of use of the application to detect presumptive cases of cyberbullying, Table VIII, evidences that 70% consider that both the mobile application and the web are very easy to use and 30% consider that are easy to use. This means that the majority of participants find the applications easy to use.

TABLE VIII. HOW SATISFIED ARE YOU WITH THE EASE OF USE OF THESE APPLICATIONS?

	Users	Answer
Very easy to use	7	70%
Easy to use	3	30%
Absolutely difficult to use	0	0%

Moreover, when users were asked, how satisfied are they with the reliability of this application? As Table IX shows, 80% consider that it is very reliable to use the applications, while 10% consider that it is not very reliable and the remaining 10% consider that it is not reliable to use the applications. This means that a minimum percentage of surveyed users are not sure about the reliability of the application.

TABLE IX. HOW SATISFIED ARE YOU WITH THE RELIABILITY OF THIS APPLICATION?

	Users	Answer
Very reliable	8	80%
Unreliable	1	10%
Not reliable	1	10%

Finally, users were asked if they would use this application daily to keep control over the social networks of the adolescents or young people they represent to prevent them from suffering presumptive cyberbullying. Table X, shows that in response to this inquiry, 80% consider that it does meet the objective of helping with cyberbullying issues and 20% consider that it meets the objective of helping with cyberbullying issues regularly.

TABLE X. WOULD YOU USE THIS APPLICATION DAILY TO MAINTAIN CONTROL OVER YOUR CONSTITUENTS' SOCIAL NETWORKS AND PREVENT THEM FROM PRESUMPTIVE CYBERBULLYING?

	Users	Answer
Yes	9	90%
No	1	10%

V. DISCUSSIONS

The review of the literature on the use of new information and communication technologies to prevent and detect cases of cyberbullying in young people and adolescents demonstrates the current concern to address the problem. The total of the research consulted is mainly focused on finding effective solutions for the detection and prevention of cyberbullying in digital environments, especially in social networks. NLP is one of the most novel technologies for analyzing and understanding contextual content generated online and identifying patterns of cyberbullying, offensive language or abusive content. However, the consideration of the contextual aspect is a very little addressed topic. Therefore, the contributions of the present research demonstrate that understanding the context is a relevant element to accurately detect cases of cyberbullying on digital platforms.

Tests carried out with the use of two Twitter accounts demonstrated the efficiency of the Parental Control Multiplatform System to extract information from the social network accounts of adolescents, analyze it using the LSA method, and generate results with the percentage of similarity, both in "negative" and "positive" texts, and alert parents or guardians whether or not there is a presumptive case of cyberbullying.

Unlike the applications developed by [19], [20], [21], [22]. The author in [23], the present research focused on the use of LSA in the Spanish language to achieve greater precision in the analysis of the similarities of texts related to cyberbullying in the context of Ecuadorian adolescents. The research analyzed demonstrates the efforts made by different countries to protect adolescents from online cyberbullying and the scarcity of solutions in Spanish. Thus, it is possible to demonstrate that the use of LSA allows obtaining an appropriate percentage of similarity on alleged cases of cyberbullying in social networks using a knowledge base extracted semi-automatically from social networks such as Twitter.

In summary, the use of Latent Semantic Analysis for detecting cyberbullying in Ecuadorian social media has demonstrated its valuable capabilities, alongside certain inherent limitations in capturing emotional connotations. LSA's strength in identifying explicit aggressive language, as evidenced in high-profile cyberbullying cases, highlights its utility as a tool in initial screening processes. However, our experiments also brought to light the method's challenges in interpreting more nuanced emotional contexts. For instance, the misclassification of contextually complex terms in Pope Francis's account illustrates the need for a deeper understanding of emotional subtleties beyond LSA's word co-occurrence framework. Also, the minimal disparities in negative and positive text percentages in certain cases further underscore the importance of integrating LSA with more context-sensitive methods. Embracing these limitations as opportunities for improvement, LSA can be effectively complemented with advanced, emotionally intelligent algorithms, paving the way for more nuanced and culturally aware cyberbullying detection systems.

The validation, functioning, and operability of the Parental Control Multiplatform System were positively assessed by representatives and experts in the field of psychology. When consulted, they stated that the multiplatform application is a novel tool for identifying possible cases of bullying on social networks. The multiplatform application to detect suspected cyberbullying differs from other existing applications because although they provide parents or guardians with tools to monitor their children's online activity [29], they are not focused on analyzing the online activities of adolescents to detect possible cases of cyberbullying through specific information analysis techniques and the identification of bullying patterns [10].

In terms of functional requirements, parental control applications generally offer monitoring of browsing history, blocking of inappropriate websites, monitoring of messages and phone calls, and setting time limits for the use of devices [19]. Unlike these, our designed application is based on Natural Language Processing using the LSA method, which allows for establishing search patterns, information retrieval, grouping of topics, generating alerts and content recommendations [17], [18].

Regarding the security of the proposed system, it could be demonstrated that it is a secure platform. It is designed to protect users' personal information and comply with privacy and data security standards. Nevertheless, a very small number of users stated that they would not use the application frequently to monitor their children's social network usage. The tests conducted showed that the application does not present potential risks in terms of privacy and data security. One of the major limitations of the apps available for parents and guardians is that they rely on excessive handling of personal information, which becomes a risk factor for the safety of families.

VI. Conclusions

Nowadays, despite the evolution of information and communication technology, very little importance has been given to the negative situations that these advances entail. The most significant issue, and one that is occurring worldwide, has to do with cyberbullying. Although its effects can be devastating, especially for young people and adolescents, there are not many tools or applications aimed at analyzing the activities that adolescents perform on their social networks and determining cases of possible harassment on digital platforms given certain publications, comments, or messages that are inappropriate.

The Parental Control Multiplatform System developed through Natural Language Processing in Spanish had the objective of detecting presumptive cases of cyberbullying in the social networks of the accounts registered in the applications, based on the extraction of information. With the use of data mining, it was possible to generate scripts called *crawlers* to obtain the information, with the respective authorization of the person who owns the account.

The content analysis is performed using the Latent Semantic Analysis method. The creation of a semantic space made it possible to determine the existing similarities between a harassment dataset, which includes positive and negative texts, and the information of each social network, with the purpose of evidencing whether there is alleged cyberbullying.

The Parental Control Multiplatform System provides the necessary information to the user. In addition to the analysis, the user is also provided with multimedia content that will allow him/her to have more information about cyberbullying. This is because, for many representatives, this is still a new and unknown topic. The first option is to contact specialists on the subject, such as psychologists, who can provide guidance and have the appropriate knowledge and the necessary tools to cope with this type of situation that could happen to the person they represent. Another option is to view a list of videos on the subject and the last option is to review articles on the Internet.

Cyberbullying is a constantly evolving phenomenon and new technologies and forms of online communication present new challenges for its investigation and prevention. Therefore, there is still much to explore and discover about cyberbullying, especially in relation to newer forms of cyberbullying, such as cyberbullying through online gaming platforms and next-generation social networks. Similarly, it is important to note that cyberbullying has significant impacts on the mental health and well-being of young people, so more research and efforts in prevention and treatment are needed.

Acknowledgment

The authors would like to thank the Cloud Computing, Smart Cities & High Performance Computing Research Group (GIHP4C), and the Computer Science students of the *Universidad Politécnica Salesiana* (UPS) for their collaboration in the development of the study. In addition, they are grateful for the support of the psychology specialists who helped in the validation of the Parental Control Multiplatform System for the prevention of cyberbullying. The parents, guardians, and representatives are thanked for their observations and suggestions for improving the use of the system.

References

[1] G. W. Giumetti and R. M. Kowalski, "Cyberbullying via social media and well-being," Current Opinion in Psychology, p. 101314, 2022.

[2] K. Subaramaniam, R. Kolandaisamy, A. B. Jalil, and I. Kolandaisamy, "Cyberbullying challenges on society: a review," Journal of positive school psychology, vol. 6, no. 2, pp. 2174–2184, 2022.

[3] D. J. Meter, R. Budziszewski, A. Phillips, and T. E. Beckert, "A qualitative exploration of college students' perceptions of cyberbullying," TechTrends, vol. 65, pp. 464–472, 2021.

[4] S.-M. Bae, "The relationship between exposure to risky online content, cyber victimization, perception of cyberbullying, and cyberbullying offending in korean adolescents," Children and youth services review, vol. 123, p. 105946, 2021.

[5] Unicef et al., "Una mirada en profundidad al acoso escolar en el ecuador," Violencia entre pares en el sistema educativo. Obtenido de: https://www. unicef. org/ecuador/acoso_escolar. pdf, 2015.

[6] G. A. León-Paredes, W. F. Palomeque-León, P. L. Gallegos-Segovia, P. E. Vintimilla-Tapia, J. F. Bravo-Torres, L. I. Barbosa-Santillán, and M. M. Paredes-Pinos, "Presumptive detection of cyberbullying on twitter through natural language processing and machine learning in the spanish language," in 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), Nov 2019, pp. 1–7.

[7] D. A. Andrade-Segarra, G. A. Le et al., "Deep learning-based natural language processing methods comparison for presumptive detection of cyberbullying in social networks," International Journal of Advanced Computer Science and Applications, vol. 12, no. 5, 2021.

[8] N. B. Alotaibi, "Cyber bullying and the expected consequences on the students' academic achievement," IEEE Access, vol. 7, pp. 153 417–153 431, 2019.

[9] ——, "Cyber bullying and the expected consequences on the students' academic achievement," IEEE Access, vol. 7, pp. 153 417–153 431, 2019.

[10] G. Wang, J. Zhao, M. Van Kleek, and N. Shadbolt, "Protection or punishment? relating the design space of parental control apps and perceptions about them to support parenting for online safety," Proceedings of the ACM on Human-Computer Interaction, vol. 5, no. CSCW2, pp. 1–26, 2021.

[11] E. L. Helfrich, J. L. Doty, Y.-W. Su, J. L. Yourell, and J. Gabrielli, "Parental views on preventing and minimizing negative effects of cyberbullying," Children and Youth Services Review, vol. 118, p. 105377, 2020.

[12] S. Ali, M. Elgharabawy, Q. Duchaussoy, M. Mannan, and A. Youssef, "Betrayed by the guardian: Security and privacy risks of parental control solutions," in Annual Computer Security Applications Conference, 2020, pp. 69–83.

[13] M. S. M. Suhaimin, M. H. A. Hijazi, R. Alfred, and F. Coenen, "Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts," in 2017 8th International conference on information technology (ICIT). IEEE, 2017, pp. 703–709.

[14] T. Kanan, O. Sadaqa, A. Aldajeh, H. Alshwabka, S. AlZu'bi, M. Elbes, B. Hawashin, M. A. Alia et al., "A review of natural language processing and machine learning tools used to analyze arabic social media," in 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT). IEEE, 2019, pp. 622–628.

[15] J. Minango, A. Flores, M. Zambrano, W. Paredes Parada, and C. Tasiguano, "Radar probability of detection in multipath environments," Trends in Artificial Intelligence and Computer Engineering: Proceedings of ICAETT 2022, pp. 91–103, 2023.

[16] D. Jurafsky and J. Martin, "Computational linguistics and speech recognition, 2000," 2000.

[17] J. Daniel, M. James H et al., Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. prentice hall, 2007.

[18] G. A. León-Paredes, L. I. Barbosa-Santillán, and J. J. Sánchez-Escobar, "A Heterogeneous System Based on Latent Semantic Analysis Using GPU and Multi-CPU," Scientific Programming, vol. 2017, p. 19, 2017.

[19] M. F. Wyne, J. Sood, C. Kempton, and T. Dao, "Safeguard: A web-based application to guard against cyberbullying." Journal of Education and Learning, vol. 10, no. 4, pp. 63–69, 2021.

[20] S. Shrimali, "A natural language processing and machine learning-based framework to automatically identify cyberbullying and hate speech in real-time," in 2022 IEEE MIT Undergraduate Research Technology Conference (URTC). IEEE, 2022, pp. 1–5.

[21] S. Afrifa and V. Varadarajan, "Cyberbullying detection on twitter using natural language processing and machine learning techniques," International Journal of Innovative Technology and Interdisciplinary Sciences, vol. 5, no. 4, pp. 1069–1080, 2022.

[22] B. Bhatia, A. Verma, Anjum, and R. Katarya, "Analysing cyberbullying using natural language processing by understanding jargon in social media," in Sustainable Advanced Computing: Select Proceedings of ICSAC 2021. Springer, 2022, pp. 397–406.

[23] J. M. A. Soto, H. Á. Gonzales, and V. B. Saines, "Uso de una herramienta de nlp aplicada a la detección del ciberacoso en twitter," Innovación y Software, vol. 3, no. 2, pp. 81–90, 2022.

[24] D. Van Bruwaene, Q. Huang, and D. Inkpen, "A multi-platform dataset for detecting cyberbullying in social media," Language Resources and Evaluation, vol. 54, pp. 851–874, 2020.

[25] R. Slonje, P. K. Smith, and A. Frisén, "The nature of cyberbullying, and strategies for prevention," Computers in human behavior, vol. 29, no. 1, pp. 26–32, 2013.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[27] M. Eyuboglu, D. Eyuboglu, S. C. Pala, D. Oktar, Z. Demirtas, D. Arslantas, and A. Unsal, "Traditional school bullying and cyberbullying: Prevalence, the effect on mental health problems and self-harm behavior," Psychiatry research, vol. 297, p. 113730, 2021.

[28] S. Hinduja and J. W. Patchin, Bullying beyond the schoolyard: Preventing and responding to cyberbullying. Corwin press, 2014.

[29] M. Anderson, "Parents, teens and digital monitoring," 2016.

Mukh-Oboyob: Stable Diffusion and BanglaBERT Enhanced Bangla Text-to-Face Synthesis

Aloke Kumar Saha, Noor Mairukh Khan Arnob*, Nakiba Nuren Rahman, Maria Haque,
Shah Murtaza Rashid Al Masud, Rashik Rahman*
Department of CSE, UAP, Dhaka, Bangladesh

***Abstract*—Facial image generation from textual generation is one of the most complicated tasks within the broader topic of Text-to-Image (TTI) synthesis. It is relevant in several fields of scientific research, cartoon and animation development, online marketing, game development, etc. There have been extensive studies on Text-to-Face (TTF) synthesis in the English language. However, the amount of relevant existing work in Bangla is limited and not comprehensive. As the TTF field is not vastly prospected for Bangla language, the objective of this study sets forth to explore the possibilities in the field of Bangla Natural Language Processing and Computer Vision. In this paper, a novel system for generating highly detailed facial images from textual descriptions in the Bangla language is proposed. The proposed system named Mukh-Oboyob consists of two essential components: a pre-trained language model, BanglaBERT, and Stable Diffusion. BanglaBERT, a transformer-based pre-trained text encoder, is a language model used to transform Bangla sentences into vector representations. Stable Diffusion is used by Mukh-Oboyob to generate facial images utilizing the text embedding of the Bangla sentences. Moreover, the work utilizes CelebA Bangla, a modified version of the CelebA dataset consisting of face images, Bangla facial attributes, and Bangla text descriptions to develop and train the proposed system. This paper establishes a system for image synthesis with excellent performance and detailed image outcomes, as evidenced by a comprehensive analysis incorporating both qualitative and quantitative measures, leading to the system under consideration achieving an impressive FID score of 34.6828 and an LPIPS score of 0.4541.**

Keywords*—Bangla text-to-face synthesis; Natural Language Processing (NLP); Bangla NLP; Computer Vision (CV); Generative Model; stable diffusion; BanglaBERT

I. Introduction

Diffusion models have emerged to be a useful tool for generating realistic images in a variety of domains, such as human faces and natural landscapes. The ability to generate high-quality images from textual representations has attracted a lot of attention because of its possible applications in the development of content, augmented reality, and customized advertising.

Text-to-image generation is an approach to generating a picture from a given textual input. TTF is a subsection of TTI generation in which a human face description is provided and a facial image is generated based on the description. Compared to text-to-image generation, creating images of faces is a more difficult piece of work considering the complexity of facial features. TTF synthesis has plenty of applications that could be used, including internet marketing, animation, development of games, forensic science, and the metaverse. A significant amount of literature has been penned about the creation of faces and images from text in recent years, as this field of study has grown in prominence. Interestingly, a significant proportion of academics have focused on image creation in the English language [1].

While substantial progress has been made in the area of text-to-image synthesis based on the English language, there is a lack of advanced and enhanced research concerning non-English languages, particularly Bangla. The Bangla language presents unique challenges to the synthesis of TTF due to its unique linguistic and cultural nuances. Facial image generation from Bangla text requires an extensive knowledge of the phonological, syntactic, and semantic structures of the language. To create authentic and culturally relevant facial portrayals, it is essential to accurately capture the visual diversity and unique facial features of Bangla-speaking people.

GAN (Generative Adversarial Network)-based models used for text-to-image synthesis face unstable training, mode collapse and non-convergence intrinsically due to adversarial training [2]. Diffusion models [3] are more capable of synthesizing realistic images compared to GANs as they seldom fall into such issues, thanks to a more stable training process. Vector quantized diffusion models produce better results compared to GAN-based models using diffusion strategy to avoid error assembling for image synthesis. Moreover, It achieves improved image generation speed while maintaining excellent image quality [4].

In the field of text to face synthesis, there is a lack of extensive research for Bangla language. Therefore, this paper proposes a novel Diffusion-based system, Mukh-Oboyob, specifically to generate face images from Bangla textual input to progress TTF generation for the Bangla language. The objective of the proposed system, Mukh-Oboyob, is to mitigate an existing void in the domain of TTF generation by addressing the difficulties associated with generating images with varying structures, that differ in appearance, and level of detail while upholding the realism of the images generated from Bangla descriptions that will significantly contribute to the advancement of the field of Bangla natural language processing.

The suggested system, Mukh-Oboyob, consists of two major parts: a pre-trained language model and a latent diffusion-based model, namely, BanglaBERT and Stable Diffusion respectively. BanglaBERT [5] is utilized to learn bi-directional contexts from Bangla sentences and extract semantic information essential to the text by encoding Bangla descriptions into vector representations and performing transformations over them in order to extract contextual information. Following

that, a Stable Diffusion [3] model which is used to synthesize images from text descriptions. The model is trained and evaluated following a modified version of the CelebA dataset [6] called CelebA Bangla [7]. This dataset incorporates 40 facial attributes derived from a semantically accurate Bangla vocabulary and includes a collection of facial images aligning with the 40 corresponding facial attributes. The CelebA Bangla dataset follows a novel algorithm [7] to create Bangla textual descriptions of facial images. The evaluation of the quality, diversity, and accuracy of the generated facial images is carried out by comprehensive testing and the application of both quantitative and qualitative evaluation metrics. The system under consideration achieved an FID (Fréchet Inception Distance) score of 34.6828 and an LPIPS (Learned Perceptual Image Patch Similarity) score of 0.4541.

The following sections of the paper are constructed as follows: Section II contains the literature review. The dataset is described in Section III. Section IV discusses the methodology followed by the system. Result Analysis is elaborated in the Section V. Section VI takes through the discussion whereas Section VII states the limitations of this work. FInally, Section VIII draws the conclusion and the references are added at the end of the paper.

II. Literature Review

In this section, significant studies utilizing generative models in the field of TTI and TTF synthesis are presented.

A. Text to Image Generation

This section provides a concise overview of a few methods that are notable and have come across to achieve impressive results for text-to-image synthesis.

Reed *et al.* [8] suggested a method of translating single-sentenced text descriptions directly into image pixels by introducing a deep convolutional GAN based on text description embedding compressed using a fully connected layer and leaky-ReLU activation and text features used to perform feed-forward inference by the generator and discriminator network fundamentally demonstrating enhanced text to image synthesis. In Paper [9], they proposed the use of the Bangla Attentional Generative Adversarial Network (AttnGAN) to generate high-quality images from texts through multi-staged processing and incorporation of specific details in distinct parts of images, achieving an enhanced inception score on the CUB dataset.

Naveen *et al.* [10] examined the combination of various Transformer models and the Attentional GAN (AttnGAN) to create the AttnGANTRANS architecture to generate images from texts and validates the effectiveness of the Transformer models by assessing the performance of generated images using the Frechet Inception Distance and Inception Score evaluation metrics. In another work [11], the Cross-Modal Contrastive GAN (XMC-GAN) for text-to-image synthesis was proposed which generates images that are well-aligned with the texts and accomplish significant improvements in image quality utilizing multiple contrastive losses. Tao *et al.* [12] proposed a novel Deep Fusion GAN that generates high-quality images through a one-stage architecture and uses a deep fusion block which helps to integrate text and visual characteristics entirely. In another work by Siddharth *et al.* [13], a combination of a GAN-based model and pre-trained text encoder, Attentional GAN and ROBERTa, respectively, were used, which resulted in a significant decrease in the FID score. In paper [14], they used diffusion models for image synthesis contextual to natural language descriptions and compared CLIP and classifier-free guidance as guidance strategies. Saharia *et al.* [15], presented a text-to-image diffusion model named Imagen that achieved a high level of photorealism and text and image alignment, leading towards deep language understanding by using diffusion models along with large transformer language models for text understanding.

B. Text to Face Generation

This section presents a comprehensive summary of various methods that have garnered attention and demonstrated remarkable outcomes in the field of TTF synthesis.

Deorukhkar *et al.* [16] employed three GAN-based architectures, DCGAN, DFGAN, and SAGAN for TTF synthesis. In this paper they used the CelebA dataset [6] consisting of celebrity images, Sentence BERT for sentence embeddings to encode the textual descriptions of the images from the dataset and compared the results of the three models using IS and FID evaluation metrics. In another work [17], a GAN-based two-stream architecture was introduced to generate images with great quality and diversity. They extracted features from the images through a Contrastive Language-Image Pre-training encoder and used Cross-Modal Distillation to align the image and text features. Xia *et al.* [18] suggested a novel GAN architecture, TediGAN, which generates multi-modal images from text descriptions and proposed a Multi-Modal CelebA-HQ dataset to evaluate the result and achieve the FID score. Ayanthi *et al.* [19], used StyleGAN2 to generate visually impressive facial images with accurate rendering of the facial features and utilized BERT for text embedding to generate high-quality images that align with the text description. Paper [20] proposed a generative architecture, OpenFaceGAN that creates facial images from natural language descriptions with improved inference speed, image quality and efficiency in text and image alignment. In another paper [21], a novel network called PixelFace was proposed for TTF synthesis which utilizes a dynamic parameter-generating method to transform text features into embeddings for predicting continuous values of pixels. They validated the experimental results on the MM-CelebA dataset [18]. Nair *et al.* [22], suggested the utilization of diffusion-based models for multi-modal image synthesis that demonstrated impressive results compared to uni-modal network.

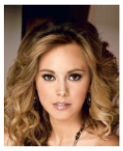
There had been research work on TTF synthesis for Bangla language but the images generated by the models were not of high quality, and the FID score was comparatively poor. Previous Bangla TTF works were unable to depict some Bangla facial attributes in synthetic images accurately. Low-quality image generation and limited consistency between Bangla text and generated images in the domain of TTF synthesis has emerged to be the existing gap for Bangla language. Therefore, the purpose of our paper is to elevate the image quality, TTI consistency and betterment of FID score for Bangla TTF synthesis.

III. Dataset

In this paper, we have used the modified version of the CelebA dataset [6] (containing celebrity images and English captions), called CelebA Bangla [7]. The novel algorithm utilized by the CelebA Bangla dataset to generate Bangla textual descriptions of facial images was originally introduced by [7]. The dataset comprises forty facial attributes that have been extracted from Bangla vocabulary that ensures semantic accuracy. Additionally, it contains around 202,599 facial images of size 128×128 of celebrities that correspond to the forty aforementioned facial attributes. The CelebA Bangla dataset[1] is available publicly on Kaggle.

Some of the samples of the CelebA Bangla dataset and text descriptions from CelebA Bangla are shown in Tables I and II.

TABLE I. Samples from the CelebA Bangla Dataset

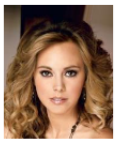
image_name	হালকা_দাড়ি (light_beard)	কুচকানো_ভ্রু (arched eyebrows)	আকর্ষণীয় (attractive)	.....	অল্পবয়স্ক (young)
		.....	.....		
000012.jpg 	-1	-1	1	.....	1
000013.jpg 	-1	-1	-1	.....	1
		.....	.....		
202378.jpg 	-1	1	1	.....	1
		.....	.....		

IV. Methodology

A. BanglaBERT

BanglaBERT [5] is an ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) transformer language model. Mukh-Oboyob uses BanglaBERT for obtaining accurate text embeddings from Bangla textual descriptions. Of all the available pre-trained Bangla text encoders, BanglaBERT was chosen for this study due to its superior performance on a plethora of NLP tasks. BanglaBERT has its own tokenizer written for the Bangla language, with a rich vocabulary and customized tokenization process. Due to this tokenizer, Bangla text is tokenized properly without loss of valuable information present in subtle parts of the text. As shown in Fig. 1, BanglaBERT turns the input text into tokens $Tok_1, Tok_2, ..., ToK_N$. These tokens are given to the ELECTRA model, which comprises of two main components: Electra embedding layers and 12 Electra layers. An Electra embedding layer consists of word embedding, position embedding, token type embedding, layer norm and dropout layers. For the purpose of capturing the semantic meaning of the tokens, the Electra layers transform the tokens into continuous vector representations. These representations are fed into 12 Electra layers. Each Electra layer consists of three components: Electra Self Attention, Electra Intermediate, and Electra Output layers. Electra layers help the model to capture contextual information from the input sequence. Electra Self Attention has two constituents: Electra Attention and Electra Self output. For capturing dependencies between tokens, the Electra layer assigns weights by using a Self-Attention mechanism. The Electra Intermediate layer consists of a dense layer and a GELU (Gaussian Error Linear Unit) activation to present contextual information. The Electra Output layer assists the model in grasping thematic insight efficiently. Finally, the output of the 12 Electra layers is a text embedding of dimensions $[num\_prompts \times max\_length \times embedding\_dim]$. Here, $num\_prompts$ is the number of prompts/input Bangla textual descriptions given to the text encoder. max_length is the max-

TABLE II. Sample Text Descriptions from CelebA Bangla Dataset

image_name	text_description
	....
000012.jpg 	ছেলেটির চোখের নিচে কালি ছিল। ছেলেটির কালো চুল ছিল। ছেলেটির সোনালী চুল ছিল। ছেলেটির উঁচু গালের হাড় ছিল। ছেলেটির মুখ কিছুটা খোলা ছিল। ছেলেটির দাড়ি নেই। ছেলেটির চেহারা ডিম্বাকৃতির। ছেলেটির মুখে ছিল হাসি। ছেলেটির সোজা চুল ছিল। (The male has pretty high cheekbones and an oval face. He has black and straight hair. He has bushy eyebrows and a slightly open mouth. The male is smiling, seems young and attractive.)
000012.jpg 	ছেলেটির সোনালী চুল ছিল। ছেলেটির উঁচু গালের হাড় ছিল। ছেলেটির মুখ কিছুটা খোলা ছিল। ছেলেটির দাড়ি নেই। ছেলেটির চেহারা ডিম্বাকৃতির। ছেলেটির মুখে ছিল হাসি। ছেলেটির সোজা চুল ছিল। (The man has high cheekbones and an oval face. His hair is blond and straight. He has a slightly open mouth. He seems young and is smiling.)
	....
202378.jpg 	মেয়েটির ভ্রু কুচকানো ছিল। মেয়েটির বড় ঠোঁট ছিল। মেয়েটির সোনালী চুল ছিল। মেয়েটির মুখে ভারী মেকাপ ছিল। মেয়েটির মুখ কিছুটা খোলা ছিল। মেয়েটির চেহারা ডিম্বাকৃতির। মেয়েটির চোখা নাক ছিল। মেয়েটির ঢেউ খেলানো চুল ছিল। মেয়েটির কানে দুল পরা ছিল। মেয়েটির লিপস্টিক পরা ছিল। মেয়েটির নেকলেস পরা ছিল। (The woman has an oval face. She has brown and wavy hair. She has arched eyebrows, big lips, a slightly open mouth and a pointy nose. She looks young, attractive and has heavy makeup. She is wearing earrings, lipstick and a necklace.)
	....

[1]https://www.kaggle.com/datasets/rashikrahmanpritom/celeba-bangla-dataset

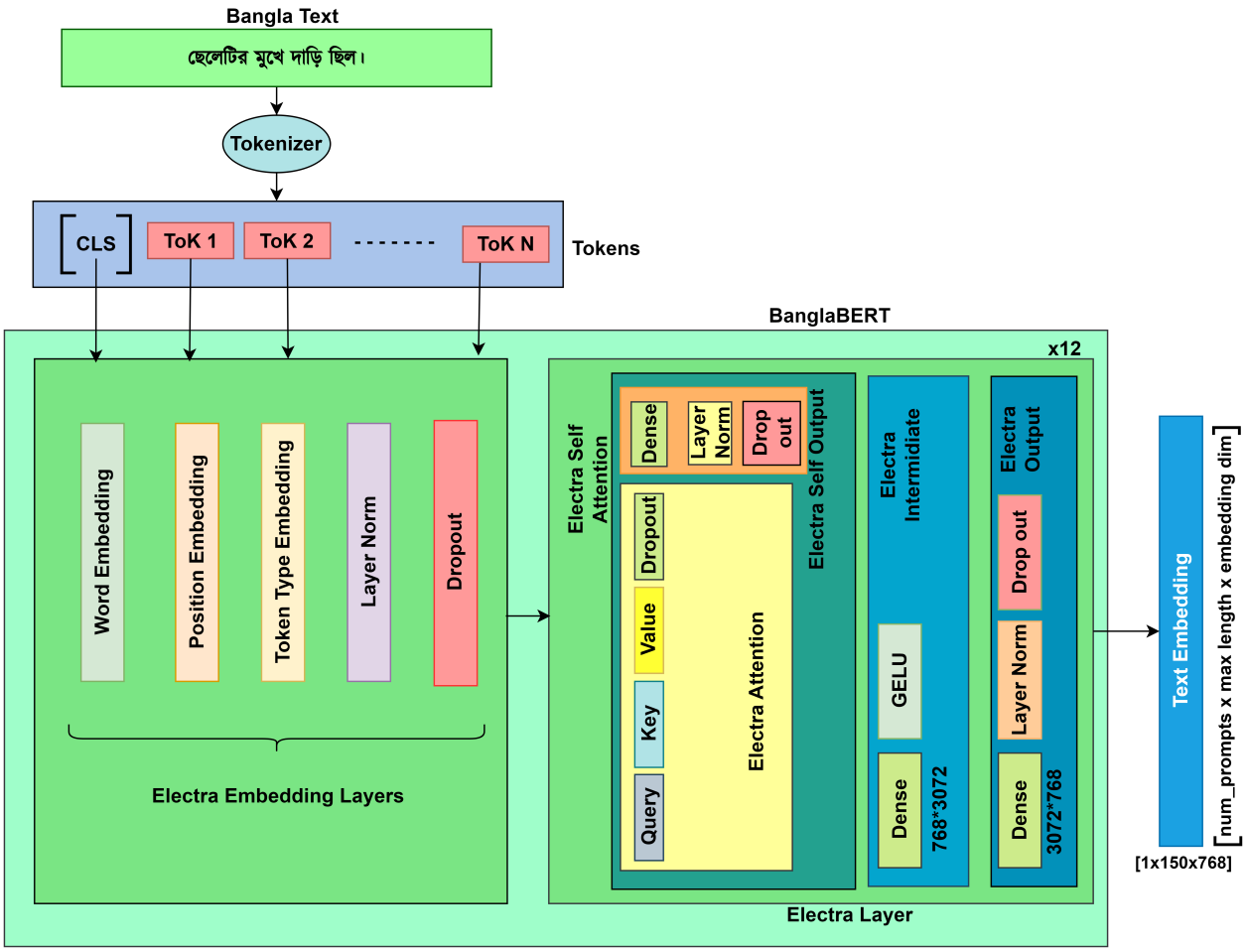


Fig. 1. Internal layered architecture of BanglaBERT.

imum number of tokens allowed to be given by the tokenizer. $embedding\_dim$ represents the embedding dimension of the embedding layers.

B. Stable Diffusion

Stable Diffusion [3] is based on the latent diffusion model that produces synthetic images from text input. Stable diffusion mainly comprises a Variational Autoencoder(VAE), some Schedulers, a Text Encoder, a U-Net Model, and Classifier Free Guidance. The training and image generation process is explained briefly below.

Training: The training process of Stable Diffusion is outlined in Fig. 2. During the training of Stable diffusion, an input facial image is passed through the VAE encoder to obtain a latent vector. As shown in Eq. 1, the latent vector is scaled by a scaling factor defined in the configuration of the VAE. Scaling the latent vector allows Mukh-Oboyob to control the amount of randomness of the probability distribution of synthetic facial images.

$$latent\_vector_{scaled} = scaling\_factor \times vae\_encoder(image_{face}) \quad (1)$$

It is shown in Eq. 2 that the scaled latent vector, a random noise vector, and timestep are passed to a noise scheduler for timesteps $t = 1....T$. This is the Forward Diffusion process. In this process, the noise scheduler gradually adds noise to the latent image, thus obtaining a noisy latent vector.

$$latent\_vector_{noisy} = noise\_scheduler(latent\_vector_{scaled}, noise_{random}, timesteps) \quad (2)$$

A Bangla textual description is passed through BanglaBERT to obtain a text embedding. The Text embedding, Random noise, and noisy latent vector are fed into the U-Net for the purpose of predicting a noise vector; as demonstrated in Eq. 3. The U-Net utilizes its contracting path and expansive path to better predict a noise vector close to the random noise previously used in the forward diffusion process.

$$Noise_{predicted} = U\text{-}Net(Text\_Embedding, timesteps, latent\_vector_{noisy}) \quad (3)$$

The predicted noise and random noise are compared using Mean Squared Error Loss as defined in Eq. 4. Here, N signifies

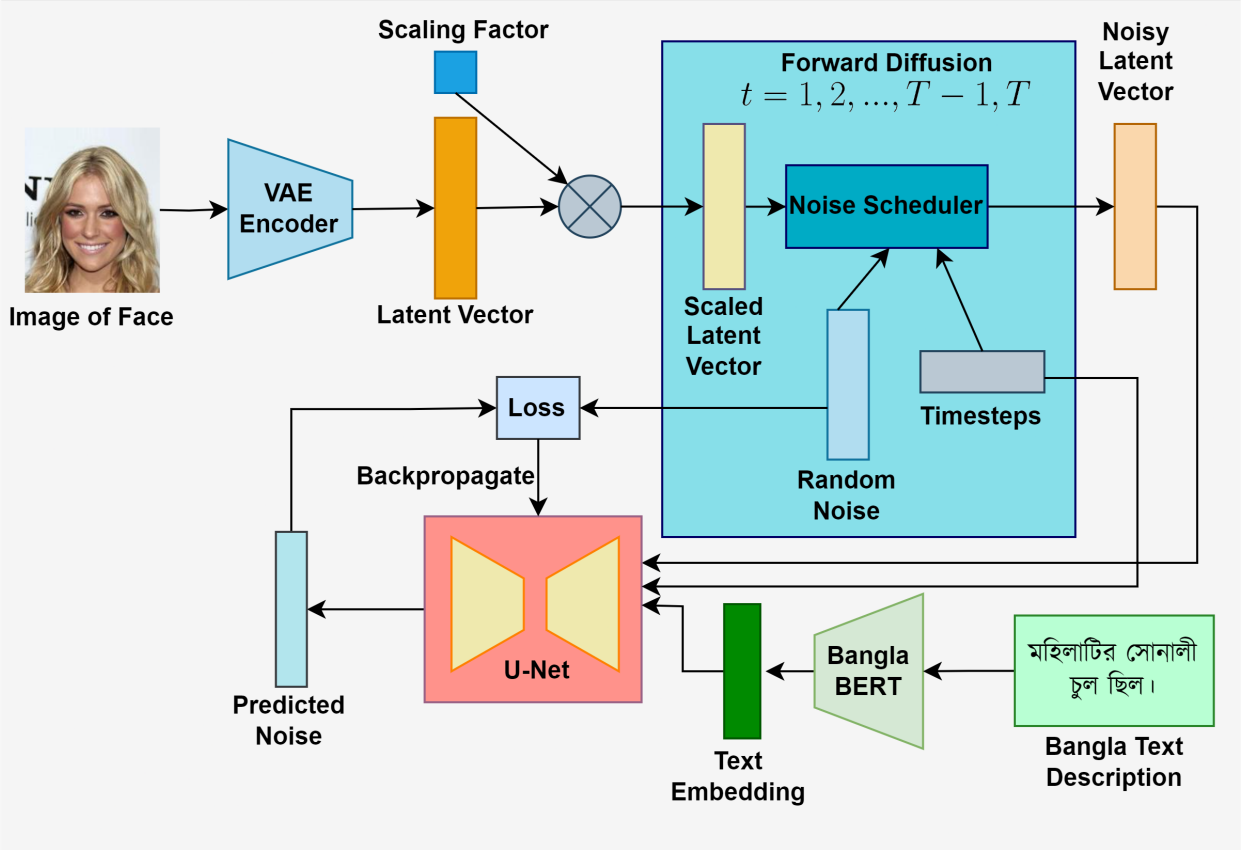


Fig. 2. Training procedure of the stable diffusion model used by Mukh-Oboyob.

the number of rows in the noise vectors. This Loss is back-propagated through the U-Net to help the U-Net predict better noise vectors in future iterations.

$$Loss_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(Noise_{random} - Noise_{predicted})^2 \quad (4)$$

Image Generation: The Image Generation Phase as depicted in Fig. 3 generates synthetic images from Bangla Text. During the Image Generation or Prompting phase, a Bangla Text prompt is sent to BanglaBERT to obtain a prompt embedding. A random latent vector is scaled by following Eq. 5. Here σ is used to control how much noise is added to the latent text representation.

$$latent\_model\_input = \frac{random\_latent}{\sqrt{\sigma^2 + 1}} \quad (5)$$

The latent model input, prompt embedding and timestep are given to the U-Net for predicting Noise in Eq. 6. This Noise vector is the U-Net's attempt to produce a latent representation of the text; which can later be decoded into an image.

$$Noise_{predicted} = U\text{-}Net(latent\_model\_input, timestep, Text\_Embedding) \quad (6)$$

However, This predicted noise is not satisfactory at timestep $t = T - 1$. Therefore, as shown in Eq. 7, the predicted noise and timesteps are iteratively passed on to the scheduler which produces another latent vector for timesteps $t = T-1, T-2, T-3, ..., 3, 2, 1$. This is the Reverse Diffusion process.

$$latent\_vector = Scheduler(Noise_{predicted}, timesteps) \quad (7)$$

Finally, in Eq. 8, the latent vector achieved at timestep $t = 1$ is scaled by a scaling factor defined in the variational autoencoder's configuration. This scaling is done to ensure that the latent vector is normalized and has values in a specific range, thus helping to improve consistency across a multitude of samples.

$$latent\_vector_{scaled} = \frac{latent\_vector}{scaling\_factor} \quad (8)$$

The scaled latent vector is now passed to the Decoder of the Variational Autoencoder used in Mukh-Oboyob. As depicted in Eq. 9, The Decoder produces an image of a face in accordance to the textual prompt given to the Text Encoder earlier.

$$image_{face} = Decoder_{VAE}(latent\_vector_{scaled}) \quad (9)$$

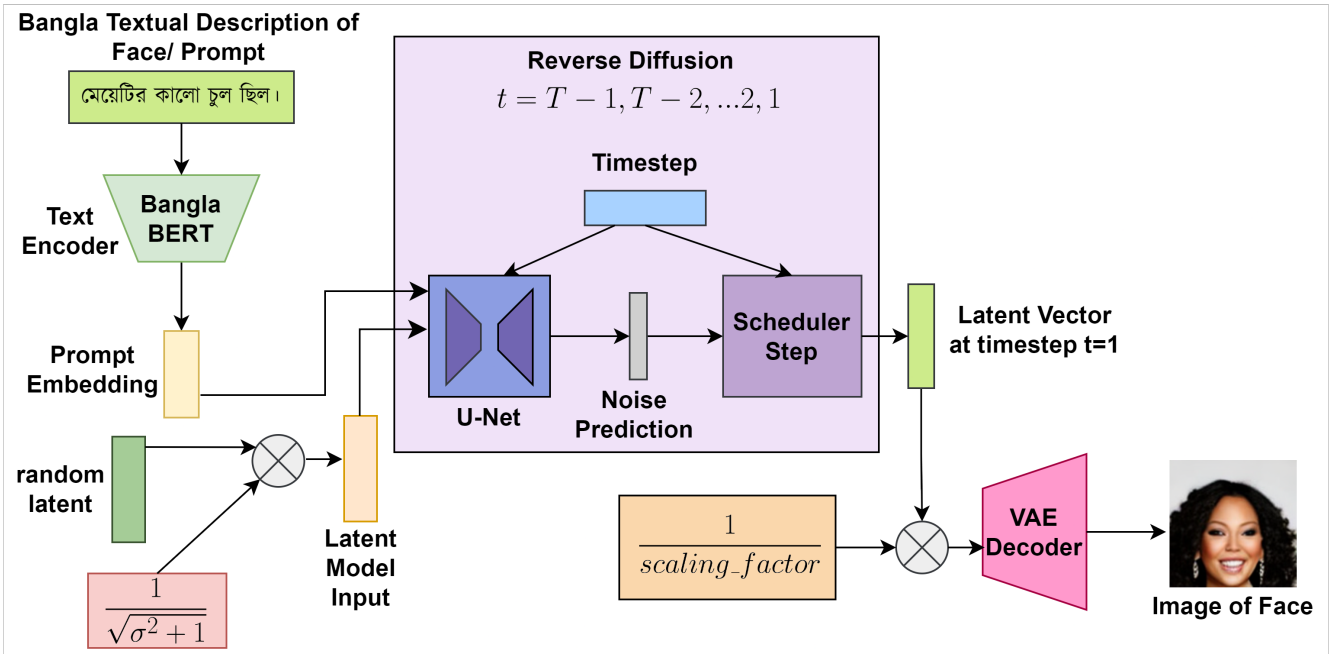


Fig. 3. Image generation procedure used in stable diffusion of Mukh-Oboyob.

C. LoRA

Fine-tuning the entire Stable Diffusion model can be a hardware and time-consuming task; often unfeasible in limited hardware and electric power support. Therefore Mukh-Oboyob uses LoRA (Low-Rank Adaptation) [23] to fine-tune the cross-attention layers in the U-Net model part of Stable diffusion. Let the weight matrix of a cross-attention layer be C_0. LoRA will selectively update C_0 by using the low-rank decomposition in Eq. 10. During the training or fine-tuning process, the C_0 matrix is not updated in the backward pass. Only T_A and T_B are updated while training.

$$C_0 + \Delta C = C_0 + T_B T_A \tag{10}$$

By significantly reducing the number of trainable parameters in the process outlined above, LoRA reduces training time and VRAM consumption drastically; without causing noticeable degradation in synthetic image quality.

V. Result Analysis

In this section, a comprehensive discussion of the experimental details during training and validation of the proposed model is provided.

A. Experimental Setup

Stable Diffusion v1-4[2] was fine-tuned using LoRA on a single RTX 3060 GPU for developing the proposed system, Mukh-Oboyob. Stable Diffusion uses a CLIP Text Encoder [24], which only works for English text inputs. For Mukh-Oboyob, BanglaBERT's text encoder and tokenizer was used. Input image resolution was changed to 128×128 to make it compatible with the CelebA Bangla dataset. Furthermore, CLIP tokenizer has a maximum sequence length of 77, whereas the textual descriptions of CelebA Bangla produce upto 150 tokens when tokenized with the BanglaBERT tokenizer. When tokenized with $max\_length = 77$, BanglaBERT's tokenizer discards significant parts of the Bangla text. Therefore, for compatibility issues, the maximum sequence length was set to 150 in the proposed system, Mukh-Oboyob. The batch size was set to 16. an initial learning rate of 10^{-4} was used. The constant scheduler was chosen as the learning rate scheduler. Number of warmup steps was set to 0 for the learning rate scheduler. The hyperparameters for the Adam optimizer used are: $\beta_1 = 0.9, \beta_2 = 0.9, weight\_decay = 10^{-2}, \epsilon = 10^{-8}$. The dimension of the LoRA update matrices was set to 4 for training the proposed method, Mukh-Oboyob. A Variational Autoencoder[3] trained with Exponential Moving Average weights was used during prompting Mukh-Oboyob.

B. Qualitative Analysis

As shown in Fig. 4, Mukh-Oboyob produces images with far better quality and diversity compared to previous GAN methods. The synthetic images produced by Mukh-Oboyob are more semantically aligned with the Bangla textual descriptions of faces. Almost all facial attributes written in the input textual descriptions are accurately depicted in the corresponding images produced by Mukh-Oboyob. This shows that BanglaBERT successfully provided meaningful text embeddings which were properly comprehended by Stable Diffusion.

The effect of the hyperparameter called number of inference steps was explored in Fig. 5. With only 1 inference step, a noisy image is produced. While inference steps increase, inference time and image quality also increase. Regardless of

[2]https://huggingface.co/CompVis/stable-diffusion-v1-4

[3]https://huggingface.co/stabilityai/sd-vae-ft-ema

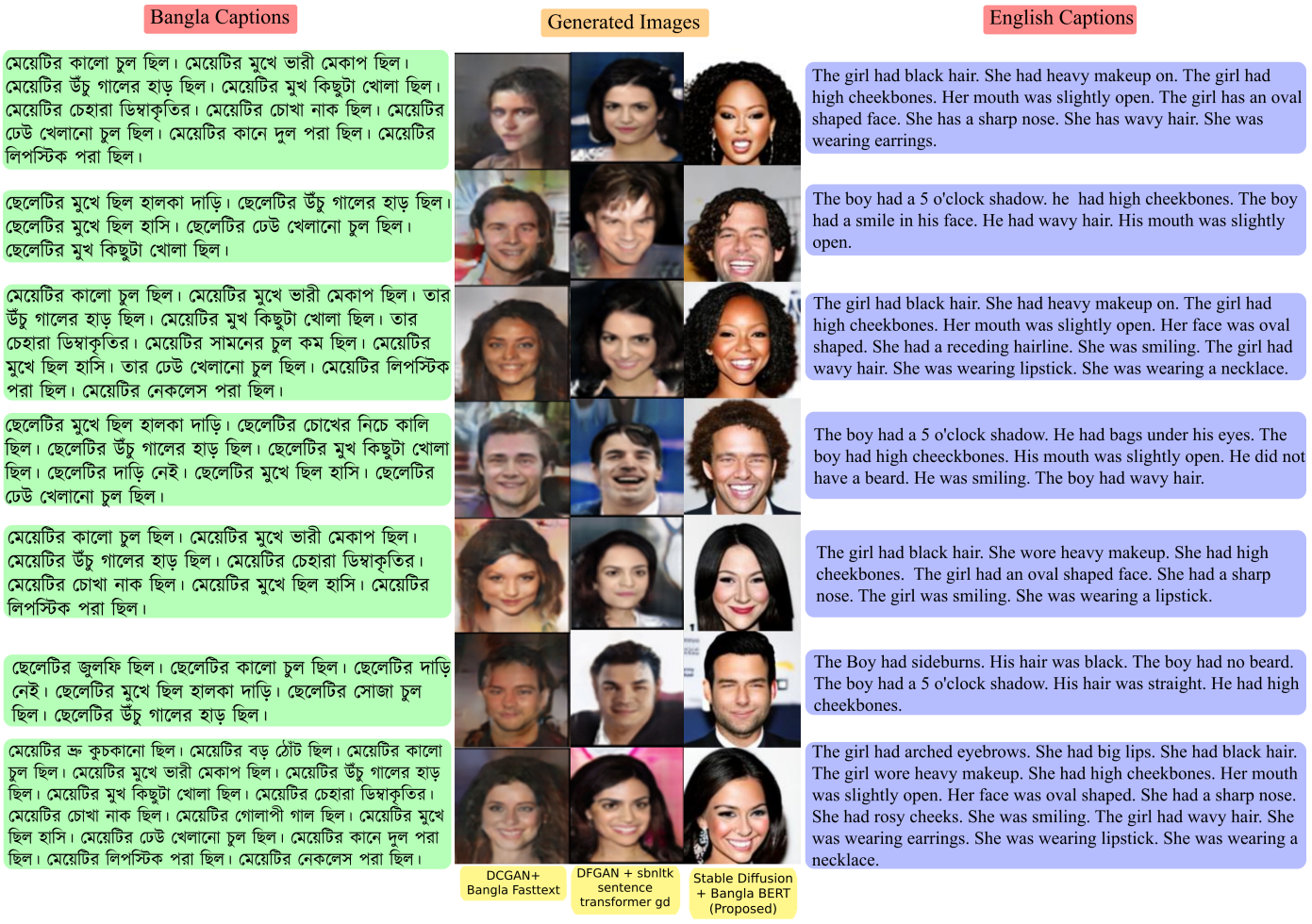


Fig. 4. Comparison of generated images between previous GAN methods and Mukh-Oboyob for different bangla captions.

the number of inference steps, VRAM consumption stays the same.

C. Quantitative Analysis

Mukh-Oboyob achieves state-of-the-art quantitative results on the performance metrics outlined in [7], overcoming previous GAN-based results, as numerically proven in Table III. All performance metrics were computed on 10,000 synthetic images, as practiced in various studies. FID is a widely used performance metric for evaluating the quality and diversity of generated images. Mukh-Oboyob achieves a better FID score of 34.6828 compared to the other models by a large margin. Although Inception Score(IS) is a metric used for assessing the quality and diversity of generated images, it is criticized in existing literature for having sensitivity to dataset bias, lack of semantic coherence, and limited applicability to different domains. Mukh-Oboyob achieves a competitive Inception Score of 11.3721, as shown in Table III. Learned Perceptual Image Patch Similarity(LPIPS) is an excellent domain agnostic performance metric for image synthesis which correlates very well with human perception. The proposed model, Mukh-Oboyob also achieves a much better LPIPS score compared to DCGAN and DFGAN. Face Semantic Similarity (FSS) and Face Semantic Distance(FSD) are used for comparing the similarity and dissimilarity of generated and real faces. Although FSS and FSD are relevant to the TTF domain, they are not widely recognized or established metrics. Nevertheless, Mukh-Oboyob achieves a competitive FSS and FSD score as shown in Table III.

TABLE III. COMPARISON OF PERFORMANCE METRICS BETWEEN MUKH-OBOYOB AND PREVIOUS METHODS

Model	FID ↓	IS ↑	LPIPS ↓	FSD ↓	FSS ↑
Bangla fasttext + DCGAN [7]	126.71	**12.3607**	21.8291	**20.2385**	0.3427
sbnltk sentence transformer gd + DFGAN [7]	155.1593	4.78246	3.2216	20.3697	**0.4203**
Mukh-Oboyob (BanglaBERT + Stable Diffusion)	**34.6828**	11.3721	**0.4541**	24.8942	0.0528

Fig. 6 depicts the decreasing MSE loss at each epoch during the training of the proposed method, Mukh-Oboyob. At each epoch, there were 12630 updates; so each epoch was very time-consuming.

Text Description/Prompt	inference steps = 1	inference steps = 5	inference steps = 10	inference steps = 15	inference steps = 20	inference steps = 25	inference steps = 30
মেয়েটির কালো চুল ছিল। মেয়েটির মুখে ভারী মেকাপ ছিল। মেয়েটির উঁচু গালের হাড় ছিল। মেয়েটির মুখ কিছুটা খোলা ছিল। মেয়েটির চোখা নাক ছিল। মেয়েটির মুখে ছিল হাসি। মেয়েটির ঢেউ খেলানো চুল ছিল। মেয়েটির লিপস্টিক পরা ছিল। (The girl had black hair. She was wearing heavy makeup. The girl had high cheekbones. Her face was slightly open. The girl had a sharp nose. She was smiling. She had wavy hair. The girl was wearing lipstick.)							
Inference Time	366 ms	449 ms	757 ms	1027 ms	1324 ms	1601 ms	1880 ms
VRAM Consumption	5.09 GB						

Fig. 5. Effect of inference steps on the quality of generated images.

Fig. 6. MSE loss at different epochs of training the proposed Mukh-Oboyob model.

VI. Discussion

Developing a system that performs well and combines state-of-the-art models on limited hardware to a new domain requires a significant amount of background knowledge and experience. The evaluation of generative models is prone to subjectivity and lack of a clear ground truth [2]. Despite these adversities, Mukh-Oboyob achieves stellar performance and establishes a new state of the art in Bangla TTF Synthesis. The most subtle bangla facial attributes are learned surprisingly by the proposed model, Mukh-Oboyob.

VII. Limitations

Even though Mukh-Oboyob achieves never-before-seen results on Bangla TTF synthesis, it is a bit behind compared to English TTF models which have achieved single-digit FID scores. The relatively less advanced performance of Mukh-Oboyob can be attributed to the lack of a pre-trained model for Bangla that adequately captures the sophisticated details of the Bangla language, as BERT or GPT captures for English. Another issue faced by Mukh-Oboyob is that some of the generated facial images contain dark or blurry eyes. Even after using a pre-trained VAE aimed at solving this issue, a few images are still synthesized with dark eyes. This is an open research problem.

VIII. Conclusion

This paper proposes a novel system, Mukh-Oboyob for producing images of faces from Bangla Textual input. Mukh-Oboyob uses BanglaBERT as a Text Encoder and Stable Diffusion for image generation. The proposed Mukh-Oboyob model was trained and evaluated on the CelebA Bangla dataset. Mukh-Oboyob achieves a state-of-the-art FID score of 34.6828 and an LPIPS score of 0.4541. A limitation of this work is that the performance of Mukh-Oboyob is relatively lower compared to state-of-the-art English TTF models. Another limitation of Mukh-Oboyob work is that this work suffers from lack of established performance metrics for evaluation of the facial features of synthetic facial images. An interesting avenue of future work can be generating more diverse and realistic facial images from captions of other languages(Arabic, Hindi, Spanish, etc.).

Acknowledgment

We are grateful to the Institute of Energy, Environment, Research, and Development (IEERD, UAP) and the University of Asia Pacific for their financial support. We extend our sincerest gratitude to Md Shopon for his inspiration and insights. We thank A. Faiyaz for technical assistance.

References

[1] N. G. Nair, W. G. C. Bandara, and V. M. Patel, "Unite and conquer: Plug & play multi-modal synthesis using diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6070–6079.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[4] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 696–10 706.

[5] A. Bhattacharjee, T. Hasan, W. Ahmad, K. S. Mubasshir, M. S. Islam, A. Iqbal, M. S. Rahman, and R. Shahriyar, "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1318–1327. [Online]. Available: https://aclanthology.org/2022.findings-naacl.98

[6] Z. Liu, P. Luo, X. Wang, and X. Tang, "Large-scale celebfaces attributes (celeba) dataset," *Retrieved August*, vol. 15, no. 2018, p. 11, 2018.

[7] N. M. K. Arnob, N. N. Rahman, S. Mahmud, M. N. Uddin, R. Rahman, and A. K. Saha, "Facial image generation from bangla textual description using dcgan and bangla fasttext," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.

[8] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*. PMLR, 2016, pp. 1060–1069.

[9] M. A. H. Palash, M. A. Al Nasim, A. Dhali, and F. Afrin, "Fine-grained image generation from bangla text description using attentional generative adversarial network," in *2021 IEEE International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON)*. IEEE, 2021, pp. 79–84.

[10] S. Naveen, M. S. R. Kiran, M. Indupriya, T. Manikanta, and P. Sudeep, "Transformer models for enhancing attngan based text to image generation," *Image and Vision Computing*, vol. 115, p. 104284, 2021.

[11] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang, "Cross-modal contrastive learning for text-to-image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 833–842.

[12] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, and C. Xu, "Df-gan: A simple and effective baseline for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 515–16 525.

[13] M. Siddharth and R. Aarthi, "Text to image gans with roberta and fine-grained attention networks," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 12, 2021.

[14] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[15] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.

[16] K. Deorukhkar, K. Kadamala, and E. Menezes, "Fgtd: Face generation from textual description," in *Inventive Communication and Computational Technologies: Proceedings of ICICCT 2021*. Springer, 2022, pp. 547–562.

[17] J. Sun, Q. Deng, Q. Li, M. Sun, M. Ren, and Z. Sun, "Anyface: Free-style text-to-face synthesis and manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 687–18 696.

[18] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "Tedigan: Text-guided diverse face image generation and manipulation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2256–2265.

[19] D. Ayanthi and S. Munasinghe, "Text-to-face generation with stylegan2," *arXiv preprint arXiv:2205.12512*, 2022.

[20] J. Peng, H. Pan, Y. Zhou, J. He, X. Sun, Y. Wang, Y. Wu, and R. Ji, "Towards open-ended text-to-face generation, combination and manipulation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5045–5054.

[21] J. Peng, X. Du, Y. Zhou, J. He, Y. Shen, X. Sun, and R. Ji, "Learning dynamic prior knowledge for text-to-face pixel synthesis," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5132–5141.

[22] N. G. Nair, W. G. C. Bandara, and V. M. Patel, "Unite and conquer: Plug & play multi-modal synthesis using diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6070–6079.

[23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

Generate Adversarial Attack on Graph Neural Network using K-Means Clustering and Class Activation Mapping

Mr. Ganesh Ingle
Department of Computer Engineering,
Vishwakarma University,
Pune, India

Dr. Sanjesh Pawale
Department of Computer Engineering,
Vishwakarma University,
Pune, India

***Abstract*—Graph Neural Networks (GNNs) have emerged as powerful tools for analyzing complex structured data, including social networks, biological networks, and recommendation systems. However, their susceptibility to adversarial attacks poses a significant challenge, especially in critical tasks such as node classification and link prediction. Adversarial attacks on GNNs can introduce harmful input graphs, leading to biased model predictions and compromising the integrity of the network. We propose a novel adversarial attack method that leverages the combination of K-Means clustering and Class Activation Mapping (CAM) to conduct subtle yet effective attacks against GNNs. The clustering algorithm identifies critical nodes within the graph, whose perturbations are likely to have a substantial impact on model performance. Additionally, CAM highlights regions of the graph that significantly influence GNN predictions, enabling more targeted and efficient attacks. We assess the efficacy of state-of-the-art GNN defenses against our proposed attack, underscoring the pressing need for robust defense mechanisms. Our study focuses on countering attacks on GNN networks by utilizing K-Means clustering and CAM to enhance the effectiveness and efficiency of the adversarial strategy. Through our observations, we emphasize the necessity for stronger security measures to safeguard GNN-based applications, particularly in sensitive environments. Furthermore, our research highlights the importance of developing robust GNNs that can withstand adversarial attacks, ensuring the reliability and trustworthiness of these models in critical applications. Strengthening the robustness of GNNs against adversarial manipulation is crucial for maintaining the security and integrity of systems that heavily rely on these advanced analytical tools. Our findings underscore the ongoing efforts required to fortify GNN-based applications, urging the research community and practitioners to collaborate in developing and implementing more robust security measures for these powerful neural network models. .**

Keywords*—Graph neural networks; adversarial attacks; K-Means clustering; class activation mapping; robustness; defense mechanisms

I. Introduction

Graph Neural Networks (GNNs) have become indispensable tools in the analysis of complex graph-structured data, with applications ranging from social network analysis to recommendation systems and bioinformatics. While their ability to discern intricate relationships within graphs is advantageous, it also exposes them to potential adversarial attacks. In critical applications, such as social network analysis, adversarial manipulations could result in the propagation of misinformation or compromise user privacy [14,15,32-35]. Similarly, attacks on user-item interaction graphs in recommendation systems may lead to the tailored manipulation of content recommendations. In bioinformatics, adversarial perturbations on molecular interaction networks pose a threat to the accuracy of predictive models, especially in the identification of potential drug candidates [17-20]. The security implications outlined above necessitate a robust defense strategy to safeguard GNNs in real-world applications. Addressing the identified security issues becomes imperative to ensure the reliability of GNNs [23,25-30,36]. The proposed defense approach employs key metrics such as success rate, transferability, and computational efficiency for assessment. Preliminary evaluations reveal competitive success rates, highlighting the method's efficacy. Notwithstanding the critical need for GNN security, existing research falls short in providing comprehensive defense strategies, especially tailored to the unique challenges posed by graph data. The primary contribution of this work lies in introducing a novel adversarial attack methodology specifically designed for GNNs. This approach integrates K-Means Clustering and Class Activation Mapping to introduce structured perturbations, offering a distinctive combination of clustering and node importance information. In the realm of counterattack advancements, where existing methods face limitations in black box scenarios, the introduction of the k-Clustering Adversarial Manipulation Approach (CAMA) dataset becomes crucial. CAMA, designed to be receptive to contrasting examples, coupled with a novel approach leveraging GNNs, showcases promising potential in creating contradictory examples. This innovation contributes significantly to the development of more effective and robust counterattack techniques, marking a crucial advancement in the field of adversarial machine learning for GNNs.

II. Background and Motivation

The escalating integration of Graph Neural Networks (GNNs) across various applications has prompted heightened concerns regarding their susceptibility to adversarial attacks. In the current landscape of research, there is a discernible dearth in the comprehensive comprehension of adversarial threats and corresponding defenses specifically tailored for GNNs. This lacuna underscores the paramount importance of fortifying GNNs against adversarial exploits to ensure their robustness in real-world deployment scenarios [2,6,7]. Our

research endeavors are motivated by the imperative to bridge the existing gaps in GNN security studies. Departing from conventional adversarial methodologies, we introduce a pioneering approach that intricately employs K-Means Clustering and Class Activation Mapping (CAM). This innovative amalgamation injects a fresh perspective into the realm of structured perturbations in GNNs, with the explicit goal of crafting more potent adversarial attacks. The principal contribution of our work lies in the conception of a novel adversarial attack methodology meticulously tailored for GNNs. The incorporation of structured perturbations, facilitated by the synergistic interplay of K-Means Clustering and Class Activation Mapping, distinguishes our approach from conventional adversarial techniques. This integration not only enriches the arsenal of adversarial methods but also provides unique insights into the vulnerabilities inherent in GNNs. Our methodology, leveraging both graph clustering and node importance information, furnishes unparalleled insights into GNN vulnerabilities. This heightened understanding empowers the creation of targeted and impactful adversarial examples, thereby propelling the advancement of adversarial research within the domain of graph-based models. Our research contributes to an elevated echelon of GNN robustness evaluation by ushering in a novel adversarial generation methodology. By seamlessly integrating K-Means Clustering and Class Activation Mapping, our approach stands as a testament to innovation in the field, offering unique perspectives and a more nuanced understanding of GNN vulnerabilities.

III. Related Work

In recent years, white-box attack research has witnessed significant growth, focusing on three main categories: discovery-based attacks, interference attacks, and attacks causing network malfunction. These attacks are classified based on the role of the "responsible model" within a specific field, addressing challenges associated with unsupervised learning. While extensive work has been conducted on white-box attacks, this literature review specifically delves into Generative Adversarial Networks (GANs) and Convolutional Neural Networks (CNNs). These deep learning methods, known for their effectiveness, are critical in understanding and countering white-box attacks.

Among various attack methods, the Fast Gradient Sign Method (FGSM) is prominent. Operating by taking a one-step gradient of the responsible damage function, FGSM facilitates the rapid creation of "opposite" examples to the original input. Despite its popularity, FGSM doesn't always guarantee success, leading to exploration of more sophisticated strategies. Theoretical and practical challenges in unsupervised learning contribute to the complexity of white-box attack research, with a focus on GANs and CNNs offering nuanced insights and potential countermeasures.

In advancing attack strategies, an iterative approach like the Repeated Fast Gradient Method (I-FGSM) and Projected Gradient Descent (PGD) has been explored. Researchers, such as [31] and [8], iterate FGSM multiple times to create more robust adversarial examples, enhancing overall attack efficacy. Additionally, [24] introduced the Momentum-based Iterative Fast Gradient Sign Method (MIFGSM), marking a significant improvement over FGSM in terms of performance and robustness. These advancements are crucial for evaluating the robustness and security of neural networks, as researchers actively work on developing sophisticated attack and defense techniques.

In the context of white-box attacks, image-dependent targeted attacks stand out as potent threats. Recent research has seen a surge in understanding and addressing vulnerabilities in neural networks, emphasizing the urgency of exploring adversarial threats. The offensive class, as defined by [39], focuses on discovering adversarial examples within disturbance budgets to ensure misclassification by the targeted neural network. This underscores the importance of image-dependent targeted attacks and the ongoing efforts in exploring GANs and CNNs for effective countermeasures. The study in [8] introduced the Momentum Iterative Fast Gradient Method (MIFGSM), an extension that incorporates momentum into the Iterative Fast Gradient Sign Method (I-FGSM). This augmentation proves to be a substantial enhancement to the attack method, resulting in the generation of more effective and robust adversarial examples.

These competitive attack methodologies play a crucial role in assessing the reliability and security of neural networks. Ongoing research endeavors are dedicated to advancing sophisticated attack techniques, concurrently emphasizing the formulation of effective defensive strategies to mitigate the impact of such attacks.

In the domain of white-box attacks, image-dependent targeted attacks are recognized as the most potent form of adversarial threats. The exploration of adversarial attacks, particularly within the realm of white-box attacks, has garnered significant interest and witnessed notable progress.

Responsive attacks can be broadly categorized into three primary types, as delineated by [1]: Alarm-based attacks, which seek to discover counterexamples within a specified disturbance budget, inducing the neural network to misclassify with high confidence. Examples of such attacks encompass the Fast Gradient Sign Method (FGSM), Iterative Fast Gradient Method (I-FGSM), and Projected Gradient Descent (PGD). Dong et al. [8] introduced an enhanced iteration, the Momentum Iterative Fast Gradient Method (MI-FGSM).

This line of research is dedicated to minimizing interference and discovering as few counterexamples as possible to deceive neural networks without triggering alarms. Various methods have been explored for this purpose:

The research in [5] focuses on minimizing interference using simpler linear functions. Carlini and Wagner [10,22] propose a method that comes close to the goal by utilizing simpler linear functions. [3] assumes linearity near the input, contributing to the minimal interference approach.

Generative attacks employ reproductive methods to create adversarial examples. Noteworthy approaches include:

The study in [21,38] utilizes a second neural network for adversarial example generation. GANs are employed to estimate the original distribution of images, providing contrasting examples.

The study in [11] introduces the CAMA framework for addressing the creation of competing examples through extensive

adversarial manipulation of graphs. Key points include:

The study encompasses various aspects related to graph structures, node properties, and their manipulation. The approach adopted focuses on facilitating hierarchical manipulation of graph structures and node properties. In parallel, the research delves into the realm of Graph Class Activation Mapping (Graph-CAM), with the primary objective of leveraging this variant to specify node-level importance in graphic classification tasks.

A significant component of the investigation involves heuristic algorithms, demonstrating their efficacy in achieving successful performance in attributive and structural attacks. Notably, these algorithms operate under strict alarm budgets, highlighting their robustness and reliability in the context of the study.

The findings of the research underscore the significance of both node-level and subgraph-level measures. This comprehensive approach is crucial for preserving competitive interference imperceptibility. The conclusion drawn from the study emphasizes the necessity of considering measures at multiple levels to ensure a holistic understanding and effective management of graph structures and node properties in various applications.

The research focuses on systemic reciprocity attacks, specifically targeting Graph Neural Network (GNN)-based link prediction models and leveraging the SEAL algorithm. One key aspect of this investigation involves counterexample generation, with the primary task being the creation of counterexamples to optimize and deceive SEAL into generating false predictions. This strategic approach aims to understand and exploit potential vulnerabilities in the link prediction model under the influence of systemic reciprocity attacks.

An integral strength of the study lies in the utilization of SEAL's y-decomposition heuristic. This heuristic theory involves approximating graph structural properties of local subgraphs, rendering it particularly effective against systemic reciprocity attacks. By leveraging the y-decomposition heuristic, the research aims to enhance the understanding of the structural intricacies of the targeted link prediction models, thereby enabling more precise and impactful attacks. Overall, the study's emphasis on counterexample generation and the strategic use of SEAL's y-decomposition heuristic underscores its dedication to unraveling and exploiting vulnerabilities in GNN-based link prediction models under the influence of systemic reciprocity attacks.

This comprehensive approach integrates linear function-based methods, generative attacks, and the CAMA framework, emphasizing the significance of node-level and subgraph-level measures to preserve imperceptibility against adversarial manipulation. Leveraging SEAL further strengthens defense against systemic reciprocity attacks on GNN-based models.

The research in [37] underscores the relevance of competition attacks on GNN-based link prediction models. The approach gradually disrupts the network graph by manipulating its structure and utilizes a link-building mechanism along with the y-decomposition heuristic theory HERMESTIC for a more efficient competitive attack. Experimental results reveal the significant impact of planned counterattacks, posing a threat to the efficacy of SEAL league predictions, particularly with limited information about complex network diagrams.

The successful portability of attacks against various link prediction heuristics from the existing literature is highlighted, demonstrating the effectiveness and broad applicability of the proposed competitive methods. Existing competing attacks (mentioned in [3, 9, 4, 12]) often operate in a "white box" configuration, assuming full access to the machine learning model parameters. However, the passage notes the unrealistic nature of this setting when the model parameters are unknown to the attacker, emphasizing the importance of considering realistic scenarios, especially in the "black box" configuration.

Competitive attacks on link prediction algorithms showcase the need for improved resilience. Proposed approaches demonstrate effectiveness across different heuristics, emphasizing the importance of realistic scenarios, particularly in the "black box" configuration where the attacker lacks full access to model parameters. The findings contribute to a deeper understanding of challenges and potential impacts of link prediction algorithms in real-world applications.

In a point-based black box arrangement, attackers query the output of the softmax layer and obtain the final classification result (x) for a given input. Challenges arise in using traditional gradient-based approaches due to the lack of well-defined gradients in discontinuous models like decision trees. The study in [10] proposes a point-based attack that reconstructs the loss function, addressing challenges in well-defined gradients. The research in [16] introduces innovations like adaptive random gradient estimation and a well-trained autoencoder to enhance the efficiency of point-based attacks. The study in [13] proposes a point-based attack using an evolutionary algorithm, demonstrating effectiveness in both point-based and hard label black box settings.

The passage emphasizes the need for robust black box attack strategies for machine learning models, noting the limitations of existing methods. A novel approach leveraging graph neural networks (GNNs) connects networks, combining the strengths of the K-means algorithm and reinforcement learning methods. GNNs serve as an effective competitive example generation tool, bridging the gap between optimization and learning approaches in adversarial machine learning research. Advances in recent literature [10] exploring black box arrangements, specifically point-based configurations, highlight the challenges of attacking models with limited parameter information, addressing these challenges through zero-order optimization techniques [16] and evolutionary algorithms [13]. These advances contribute to the development of robust black box attack strategies, emphasizing the need for effective combinations of optimization and learning methods in adversarial machine learning research [40-44].

IV. Methodology

The architecture of the GNN is intricately designed to closely emulate the structure of the target neural network that is the subject of compromise. This GNN operates within the context of adversarial attacks. When presented with an input image, information about its correct class, and details about an incorrect target class, the GNN engages in an iterative process. At each iteration, the GNN proposes a directional

update that aims to maximize the difference between the logits (output scores) of the incorrect target class and the correct class. The primary objective of this GNN is encapsulated in the "adversarial loss function."

The adversarial loss function serves as a crucial guiding principle for the GNN's optimization process. This function essentially formulates the goal of the GNN: to find the optimal perturbation or modification to the input image that maximizes the divergence between the logits associated with the correct class and those corresponding to the incorrect target class. The term "logits" refers to the raw, unnormalized output scores produced by the neural network before the application of a softmax function.

In essence, the GNN's architecture is tailored to navigate the input space in a way that induces misclassification. By proposing perturbations that push the decision boundaries of the target neural network, the GNN seeks to generate adversarial examples – instances where the neural network makes incorrect predictions despite minimal alterations to the input.

This adversarial approach involves an intricate interplay between the GNN's architecture, the specifics of the target neural network, and the definition of the adversarial loss function. The GNN learns to exploit the vulnerabilities and intricacies of the target model, demonstrating a nuanced understanding of the decision boundaries in the neural network it aims to compromise.

The architecture of the GNN is a sophisticated framework designed for crafting adversarial examples. It leverages the interplay of input perturbations and the intricacies of the target neural network to induce misclassification, with the adversarial loss function guiding the optimization process.

1. Initialization: - Before the evaluation process begins, the GNN is initialized with the target neural network's parameters, which are the parameters it aims to compromise. - These parameters may include weights, biases, and other network-specific parameters.

2. Forward Pass: - An input image, along with its correct class label and an incorrect target class label, is fed into the GNN. - The GNN processes this input image through its layers, which involve graph convolutional operations, and produces an output, typically in the form of class activation scores. - Class activation scores represent the GNN's predictions for each class, indicating the likelihood of the input belonging to each class.

3. Computation of Gradients (Backward Pass): - After obtaining predictions, a backward pass is performed to compute the gradients of the adversarial loss with respect to the input image. - Gradients capture the sensitivity of the adversarial loss to changes in the input image. - The adversarial loss is a measure of the difference between the logits associated with the incorrect target class and the correct class. The GNN aims to maximize this loss during the adversarial attack.

4. Adversarial Loss Maximization: - The computed gradients guide the GNN in the direction that maximally perturbs the input image to induce misclassification. - The GNN iteratively updates the input image in this direction, aiming to maximize the adversarial loss. - This process is repeated for a specified number of iterations or until a convergence criterion is met.

5. Assessment of Effectiveness: - Throughout this process, the GNN monitors the changes in the adversarial loss and observes the impact on the class predictions. - The effectiveness of the GNN is assessed based on its ability to successfully generate adversarial examples that lead to misclassification by the target neural network. - Success is measured by observing changes in the predicted class, ideally causing the neural network to classify the input image into the incorrect target class.

6. Comparison and Analysis: - The generated adversarial examples can be compared against baseline images to quantify the extent of the perturbation and evaluate the stealthiness of the attack. - Statistical measures or metrics may be employed to compare the success rates and the impact on different classes.

Below are some key points which discusses a common practice in the existing literature related to the evaluation of adversarial attacks.

The evaluation practices for adversarial attacks in image classification models are conventionally centered around assessing success rates on a predefined set of images, incorporating a specific perturbation size during the evaluation process. However, a recognized limitation in this approach raises concerns about its ability to offer a comprehensive evaluation of a model's robustness under adversarial attacks. This acknowledgment stems from the understanding that the fixed set of images, coupled with a uniform perturbation size, may not adequately capture the nuanced and diverse challenges posed by different instances within the dataset.

The highlighted variability in the robustness of images further accentuates the limitations of the current evaluation paradigm. Some images naturally exhibit resilience to attacks, leading to instances of attack failures, while others prove more susceptible, resulting in successful adversarial attacks. The inherent diversity in image responses to attacks poses a challenge to accurately discerning significant differences in success rates between various adversarial attack methods. In light of these considerations, there is a clear indication that a more comprehensive evaluation methodology is imperative. Such an approach should account for the distinct responses of different images to adversarial attacks, thereby fostering a deeper and more nuanced understanding of the model's vulnerability across diverse scenarios.

A. Proposed Architecture

In the context of the MNIST dataset, each image is conceptualized as a node in a graph. The relationships between these nodes are established through edges, which have the potential to capture spatial or contextual connections between pixels.

1. Graph Neural Network (GNN) Architecture

The architecture of a Graph Neural Network (GNN) typically comprises multiple graph convolutional layers. These layers play a crucial role in processing information from neighboring nodes and subsequently updating node representations based on the received information.

2. Forward Propagation Process

The initialization phase involves assigning each node, representing an image in MNIST, with a feature vector typically derived from pixel values. The forward propagation within the Graph Neural Network (GNN) includes the repeated application of graph convolutional layers. The mathematical representation of this process is given by the equation:

$$h_v^{(l+1)} = \sigma\left(\sum_{u \in N(v)} W^{(l)} h_u^{(l)} + b^{(l)}\right) \quad (1)$$

Here, $h(l)_v$ denotes the representation of node v at layer l, $N(v)$ represents the neighbors of node v, $W(l)$ is the weight matrix at layer l, $b(l)$ is the bias term, and σ is the activation function.

The aggregation step involves combining information from neighboring nodes based on the graph structure, ensuring that each node's representation incorporates information from its immediate context. The final layer of the GNN produces the output, utilized for various tasks such as classification or regression. The underlying mathematical insight lies in the graph convolutional operation, where each node's representation is updated by considering the representations of its neighbors, weighted by learned parameters. These parameters $W(l)$ and $b(l)$ are learned during the training process through backpropagation and optimization algorithms.

3. Training

The training process of Graph Neural Networks (GNNs) involves feeding labeled data through the network, computing a loss function, and updating parameters using optimization techniques like gradient descent. This iterative training approach allows the network to learn and adjust its parameters to improve its performance on the given task.

The utilization of this architecture enables GNNs to effectively capture and leverage the inherent graph structure in the MNIST dataset. This capability makes GNNs a potent tool for tasks such as image classification, where understanding and utilizing relationships between images are crucial for accurate predictions.

It's important to note that the specifics of the GNN architecture and mathematical operations may vary based on the exact implementation and variations in GNN models. Different approaches and variations in the model design can influence how the network processes information and learns from the input data.

4. Proposed Architecture Diagram

The provided Fig. 1 describes the proposed architecture diagram of the approach, illustrating the GNN layer update, class activation scores, and K-Means clustering steps.

B. Graph Representation:

Each image in the MNIST dataset is symbolically depicted as a node within a graph, resulting in a total number of nodes equivalent to the dataset's image count. The structural foundation of this graph is established by edges, which serve as connectors between nodes. The interpretation of these connections is flexible, allowing the model or designer to attribute meaning to the relationships encoded within the graph. The edges, in particular, may encapsulate spatial connections reflective of the pixel arrangements in the images. One conceivable interpretation involves connecting pixels in close proximity within an image. By doing so, the graph becomes a representation capable of capturing the spatial relationships embedded in the pixel structure of the images, contributing to a richer and more nuanced understanding of the dataset.

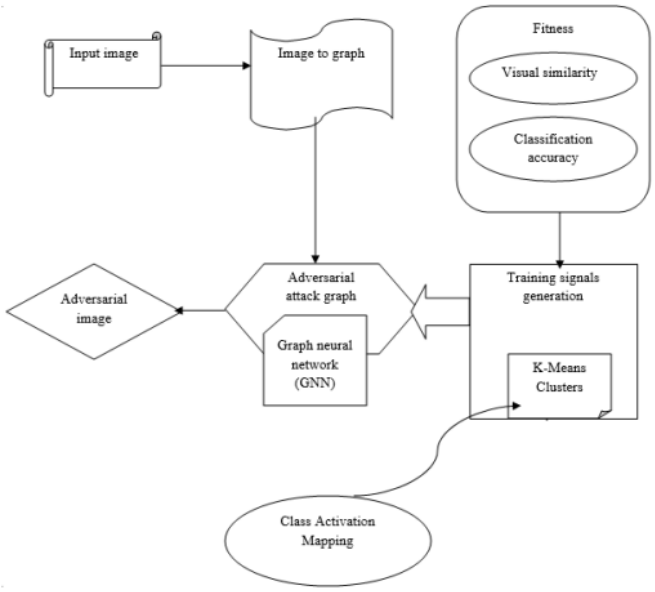


Fig. 1. Architecture of proposed approach.

The utilization of a Graph Neural Network (GNN) implies a deliberate leveraging of the underlying graph structure for the processing and analysis of data. GNNs, typically employed for such tasks, function by aggregating and updating node representations through graph convolutional layers. In the specific context of MNIST, this entails considering the pixel values of neighboring nodes when updating the representation of a given node, emphasizing the importance of spatial relationships in image data.

In the realm of contextual connections, the edges in the graph extend beyond spatial relationships, potentially representing contextual connections between images based on shared similarities or patterns within the dataset. GNNs, equipped with the capacity to learn and capture intricate contextual information from these graph connections, prove particularly advantageous for tasks like image classification, where understanding the context of an image significantly influences performance.

The structured graph representation aligns with the inherent structure of the images in the MNIST dataset, with nodes symbolizing individual images and edges encapsulating relationships between them. Depending on the GNN's design, the graph may exhibit dynamic characteristics, with edges adapting during the training process based on learned relationships.

This approach facilitates the flexibility to capture both spatial and non-spatial features, depending on how edges are defined. The graph's adaptive learning, inherent to GNNs, allows the model to dynamically adjust to the unique characteristics of the MNIST dataset, effectively leveraging both spatial and contextual information for enhanced performance in image-related tasks. In essence, the graph representation within the context of a GNN provides a structured and adaptable framework for comprehensively understanding the relationships between images, crucial for tasks like image classification.

C. Graph Neural Network (GNN) Architecture:

The primary purpose of graph convolutional layers in a GNN is to process information from neighboring nodes and update node representations. A typical GNN architecture involves multiple graph convolutional layers stacked on top of each other. Each layer processes information from the neighboring nodes of each node in the graph. Neighboring nodes are determined based on the edges in the graph.The forward propagation begins with the initialization of node representations. Each node is initialized with a feature vector, often derived from the input data (e.g., pixel values for images).The forward propagation involves the repeated application of graph convolutional layers. In each layer, information from neighboring nodes is aggregated to update the node representations.

Mathematically, the update process for the representation of a node v at layer $l+1$ can be represented as:

$$h_v^{(l+1)} = \sigma\left(\sum_{u \in N(v)} W^{(l)} h_u^{(l)} + b^{(l)}\right) \tag{2}$$

Here, $h_v^{(l)}$ is the representation of node v at layer l, $N(v)$ represents the neighbors of node v, $W^{(l)}$ is the weight matrix at layer l, $b^{(l)}$ is the bias term, and σ is the activation function.

The aggregation step in a Graph Neural Network (GNN) is a crucial process that involves consolidating information from neighboring nodes based on the underlying graph structure. This ensures that each node's representation is enriched with insights from its immediate context. Following the aggregation, the updated node representations capture both the inherent features of the node itself and information gleaned from its neighbors. This dual consideration enables the GNN to factor in both local and global information during the learning process.

Moving to the output layer, it serves as the final stage where the GNN produces task-specific outputs, such as classifications or regression predictions. The operations within this layer are tailored to the specific task at hand, involving transformations of the learned representations into meaningful predictions.

The learning process involves the adaptation of parameters, including weight matrices ($W^{(l)}$) and biases ($b^{(l)}$), through techniques like backpropagation and optimization algorithms. This adaptive learning mechanism empowers the GNN to dynamically adjust its parameters based on the patterns and relationships discerned from the graph-structured data.

The flexibility and adaptability of the GNN architecture further contribute to its efficacy. The graph structure, inherently flexible, allows the model to adapt to various graph structures, with edges and connections potentially changing during training based on the learned relationships. The intermediate layers of graph convolution provide task-agnostic representations of nodes, rendering the GNN suitable for a spectrum of graph-related tasks.

A GNN equipped with multiple graph convolutional layers systematically processes information from neighboring nodes, updating node representations, and adeptly capturing both local and global context. The adaptive learning mechanism ensures the GNN's proficiency in learning and adjusting parameters, enabling effective modeling of relationships within graph-structured data.

D. Forward Propagation Process:

In the context of the MNIST dataset, the graph representation treats each image as a node. The initialization process begins by assigning a feature vector to each node, where this vector serves as the initial representation of the corresponding image.

The source of features for each node lies in the pixel values of the corresponding image. These pixel values, which convey intensity or color information at different locations within the image, are typically organized in a grid-like structure. Depending on the design of the Graph Neural Network (GNN) and the specific task requirements, preprocessing steps may be applied to the pixel values. These steps could include normalization to a specific range or other transformations.

The feature vector itself is a numerical representation that encapsulates essential characteristics of the image. Each element of the vector corresponds to a specific feature or pixel value, and the dimensionality of this vector is determined by the number of elements it contains. This dimensionality is a critical factor influencing the GNN's capacity to capture and process information effectively.

During the training process, the parameters of the GNN, including those related to the initialization of the feature vectors, are learnable and adjusted based on observed patterns in the data. The adaptability of the GNN allows it to dynamically learn and update node representations as it processes information and identifies relevant patterns within the graph-structured data.

In task-specific scenarios, the initialization process may be tailored to the particular task the GNN is designed for. For instance, in image classification tasks, the feature vector should be crafted to capture characteristics pertinent to distinguishing between different classes of images. The quality of these initial feature vectors significantly influences the GNN's learning process, as well-initialized representations provide a robust foundation for the model to build upon during training.

The update process for the representation of a node v in a graph neural network (GNN) at layer $l+1$. Let's break down the components of the equation:

$$h_v^{(l+1)} = \sigma\left(\sum_{u\in\mathcal{N}(v)} W^{(l)} h_u^{(l)} + b^{(l)}\right) \tag{3}$$

$h_v^{(l+1)}$: This is the representation of node v at layer $l+1$. It captures the information about the node after the application of the graph convolutional layer.

$\mathcal{N}(v)$: Represents the neighbors of node v in the graph. The sum is taken over all neighboring nodes.

$W^{(l)}$: The weight matrix at layer l. This matrix contains learnable parameters that are adjusted during the training process to capture the relationships between nodes.

$h_u^{(l)}$: The representation of a neighboring node u at layer l. It contributes to the update of the central node's representation.

$b^{(l)}$: The bias term at layer l. It provides an additional learnable parameter that influences the node representation.

σ: The activation function. It introduces non-linearity to the model. Common choices include the sigmoid ($\sigma(x) = \frac{1}{1+e^{-x}}$), hyperbolic tangent ($\sigma(x) = \tanh(x)$), or Rectified Linear Unit (ReLU) ($\sigma(x) = \max(0, x)$).

The equation describes the aggregation of information from neighboring nodes using the weight matrix $W^{(l)}$, and the result is passed through an activation function σ. This operation is a fundamental step in the forward propagation of a GNN, allowing the model to capture and update node representations based on the graph structure.

This mathematical formulation is crucial for understanding how information flows through the graph convolutional layers, enabling the GNN to learn hierarchical and contextual representations of nodes in the graph.

The aggregation involves combining information from neighboring nodes based on the graph structure. This step ensures that each node's representation incorporates information from its immediate context.

$$\text{Aggregated Information} = \sum_{u\in\mathcal{N}(v)} \text{Weight}\times\text{Representation of } u \tag{4}$$

Here, $\mathcal{N}(v)$ represents the set of neighbors for the central node v.

1) Output Layer: The graph convolutional operation at layer l can be mathematically represented as:

$$h_v^{(l+1)} = \sigma\left(\sum_{u\in\mathcal{N}(v)} W^{(l)} h_u^{(l)} + b^{(l)}\right) \tag{5}$$

$h_v^{(l+1)}$: Representation of node v at layer $l+1$

$\mathcal{N}(v)$: Set of neighbors of node v in the graph

$W^{(l)}$: Weight matrix at layer l

$h_u^{(l)}$: Representation of neighboring node u at layer l

$b^{(l)}$: Bias term at layer l

σ : Activation function

The parameters $W^{(l)}$ and $b^{(l)}$ are learned during the training process through backpropagation and optimization algorithms.

Training involves feeding labeled data through the network, computing a loss function, and updating the parameters using optimization techniques like gradient descent.

This overall architecture enables GNNs to capture and leverage the inherent graph structure in the MNIST dataset, providing a powerful tool for tasks such as image classification. The specifics of the architecture and mathematical operations may vary based on the exact implementation and variations in GNN models.

For a GNN layer l with node representations $H^{(l)}$ and edge connectivity E, the update equation is given by:

$$H^{(l+1)} = \text{Aggregator}(H^{(l)}, E) \tag{6}$$

where, Aggregator is a function that aggregates information from neighboring nodes and edges.Class activation scores (S_c) are computed for each node in the graph. Let $H^{(1)}$ represent the node representations after the first GNN layer. For each class c, the class activation score $S_c[i]$ for each node i is computed using the ReLU activation function and the corresponding class weight W_{1c}:

$$S_c[i] = \text{ReLU}(H^{(1)}[i] \cdot W_{1c}) \tag{7}$$

K-means clustering is applied to group nodes based on their activation scores. The mathematical modeling for K-means clustering involves finding K cluster centers μ_k that minimize the sum of squared distances to their assigned data points:

$$\arg\min_{\mu_1,\mu_2,\ldots,\mu_K} \sum_{i=1}^{N}\sum_{k=1}^{K} \delta(i,k) \cdot \|S_c[i] - \mu_k\|^2 \tag{8}$$

where, N is the total number of nodes, $\delta(i,k)$ is the Kronecker delta, $S_c[i]$ is the class activation score for node i and class c, and μ_k is the cluster center for cluster k. The objective of K-means clustering is to find K cluster centers μ_k that minimize the sum of squared distances between each node's class activation score $S_c[i]$ and the cluster centers μ_k:

$$\arg\min_{\mu_1,\mu_2,\ldots,\mu_K} \sum_{i=1}^{N}\sum_{k=1}^{K} \delta(i,k) \cdot \|S_c[i] - \mu_k\|^2 \tag{9}$$

Here, K is the number of clusters, $\delta(i,k)$ is the indicator function, $S_c[i]$ is the class activation score for node i and class c, and μ_k is the cluster center for cluster k.

K-means clustering proceeds through the following iterative steps:

Initialization: Randomly initialize K cluster centers μ_k.

2) Assignment: Assign each node to the nearest cluster center based on the Euclidean distance between its class activation score $S_c[i]$ and the cluster centers μ_k:

$$\delta(i,k) = \begin{cases} 1 & \text{if } k = \arg\min_j \|S_c[i] - \mu_j\| \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

3) Update Centers: Recalculate the cluster centers μ_k as the mean of the class activation scores of nodes assigned to each cluster:

$$\mu_k = \frac{\sum_{i=1}^{N} \delta(i,k) \cdot S_c[i]}{\sum_{i=1}^{N} \delta(i,k)} \tag{11}$$

4) Convergence: Repeat the assignment and update steps until convergence, where minimal change in cluster assignments or cluster centers indicates stability.

- Cluster Centers μ_k: Representative points for each cluster, adjusted to minimize the sum of squared distances. - Node Assignment $\delta(i,k)$: Indicator function assigning nodes to the cluster with the closest center. - Objective Function: Minimizes the sum of squared distances, encouraging similar class activation scores within clusters. - Initialization and Convergence: Sensitive to initialization; iterative nature ensures convergence to a stable solution.

Step 2.1: Initialize Cluster Centers

First, K initial cluster centers μ_k are randomly chosen from the data points. These centers represent the initial guess of the cluster centroids.

Step 2.2: Assign Nodes to Clusters

Each node is assigned to the cluster whose cluster center is closest to it. This assignment is based on the Euclidean distance between the class activation scores of the node ($Sc[i]$) and the cluster center (μ_k):

$$\delta(i,k) = \begin{cases} 1 & \text{if node } i \text{ is assigned to cluster } k, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

$$k_{\text{assigned}} = \arg\min_k \|Sc[i] - \mu_k\|^2 \tag{13}$$

Step 2.3: Update Cluster Centers

After assigning all nodes to clusters, the cluster centers are updated by taking the mean of the class activation scores of the nodes within each cluster:

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} Sc[i] \tag{14}$$

represents the set of nodes in cluster k.

Step 2.4: Repeat Assignment and Update

Steps 2.2 and 2.3 are repeated until convergence. At each iteration, nodes are reassigned to clusters based on updated cluster centers, and centers are recalculated based on new assignments.

Identifying Target Cluster

For the target class target_c, the cluster target_C with the highest sum of class activation scores for that target class is identified:

$$\text{target} = \arg\max_k \sum_{i \in \text{Cluster}_k} S_c^{\text{target}}[i] \tag{15}$$

Here, target_C is the cluster containing nodes with high activation scores for the target class target_c.

- Cluster Initialization (μ_k): Represents the initial guess of cluster centroids. - Node Assignment ($\delta(i,k)$): Assigns each node to the cluster with the closest center. - Update Centers (μ_k): Recalculates cluster centers based on the mean of class activation scores. - Iteration and Convergence: Repeats assignment and update steps until convergence. - Identifying Target Cluster: Locates the cluster with the highest sum of class activation scores for the target class.

To create an adversarial attack, we perturb the characteristic vectors of selected target nodes. The purpose is to make small changes to the feature vectors so that the GNN misclassifies selected target nodes. By adding a small amount of noise to the feature vectors, we aim to push the decision limits of the GNN, causing incorrect predictions.

For each selected target node i, we introduce a noise vector ϵ_i to perturb its feature vector

$$X_{\text{perturbed}}[i] = X_{\text{original}}[i] + \epsilon_i \tag{16}$$

Here: - $X_{\text{original}}[i]$ is the original feature vector of node i. - $X_{\text{perturbed}}[i]$ is the perturbed feature vector of node i. - ϵ_i is the noise vector for node i.

The noise vector ϵ_i is typically drawn from a small distribution around zero, such as a normal distribution with mean 0 and a small standard deviation σ. The standard deviation σ controls the magnitude of the perturbation. Modest changes in the feature space during forward propagation might result in various activations and node representations in the GNN, changing the decision bounds and leading to misclassification of the target nodes.

To analyze the impact of perturbation on the GNN's decision, we can perform a Taylor expansion analysis. Let $F(X_i)$ represent the GNN's output (e.g., class probabilities) for node i with the feature vector X_i. The Taylor expansion can be expressed as:

$$F(X_{\text{perturbed}}[i]) = F(X_{\text{original}}[i]) + \nabla F(X_{\text{original}}[i]) \cdot \epsilon_i + O(\|\epsilon_i\|^2) \tag{17}$$

Here: - $\nabla F(X_{\text{original}}[i])$ represents the gradient of F with respect to $X_{\text{original}}[i]$. - $O(\|\epsilon_i\|^2)$ represents the higher-order terms that involve the square of the perturbation ϵ_i.

From the Taylor expansion, we can observe that the perturbation ϵ_i contributes to the change in the GNN's output, and the gradient $\nabla F(X_{\text{original}}[i])$ indicates the sensitivity of the model's output to perturbations. The specific expression for $\nabla F(X_{\text{original}}[i])$ depends on the GNN architecture and the specific layers used.

V. Algorithm

Input:

- GNN model (previously trained on the MNIST dataset)
- MNIST dataset (with labeled images)
- Target class c_{target} (the class you want to misclassify)

- Number of clusters K for K-means
- Number of target nodes to select from each cluster N
- Standard deviation of noise σ for perturbation
- Scaling factor α for perturbation control

Output:

- Perturbed graph with target nodes modified for the adversarial attack

1) Preprocess the MNIST Dataset: Load the MNIST dataset and preprocess the images (e.g., normalize pixel values to the range [0, 1]).
2) Compute Class Activation Scores: Perform forward propagation on the graph to obtain the output of the first GNN layer $H^{(1)}$. For each class c in the MNIST dataset, compute the class activation scores S_c using the formula:

$$S_c = \text{ReLU}(H^{(1)} \cdot W_{1c})$$

where W_{1c} is the weight matrix corresponding to class c, and ReLU is the Rectified Linear Unit activation function.
3) Apply K-means Clustering: Apply K-means clustering to the class activation scores S_c to group nodes into K clusters based on their activation patterns. Obtain the cluster assignments for each node in the graph.
4) Select Target Nodes: Identify the cluster C_{target} with the highest sum of class activation scores for the target class c_{target}:

$$C_{\text{target}} = \arg\max_k \sum_{i \in \text{Cluster}_k} S_{c_{\text{target}}}[i]$$

From the cluster C_{target}, select the N target nodes with the highest class activation scores for class c_{target}. If N is larger than the number of nodes in C_{target}, select all nodes in C_{target}.
5) Perturb Node Features: For each selected target node i, compute the noise vector ϵ_i from a normal distribution with mean 0 and standard deviation σ:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

where I is the identity matrix. Rescale the noise vector ϵ_i with the scaling factor α to control the strength of the perturbation:

$$\epsilon_i' = \alpha \cdot \epsilon_i$$

Perturb the feature vector of each selected target node:

$$\text{Perturbed feature vector}[i] = \text{clip}(\text{Original feature vector}[i] + \epsilon_i, 0, 1)$$

where clip ensures that the perturbed features stay within the valid range [0, 1].
6) **Reevaluate GNN:** Reevaluate the GNN model on the modified graph with the perturbed features. Check if the target nodes are now misclassified as class c_{target}.
7) Evaluate Attack Success Rate: Measure the success rate of the attack by calculating the percentage of misclassified target nodes out of the total number of selected target nodes.

VI. Experimental Set Up

1. Dataset Description:

The MNIST dataset is a widely recognized and extensively used collection of grayscale images, each depicting handwritten digits ranging from 0 to 9. Each image is formatted as a 28x28 pixel grid, resulting in a total of 784 pixels per image. The pixel values range from 0 to 255, with 0 representing black and 255 representing white. This dataset is partitioned into two primary subsets: a training set with 60,000 images and a test set with 10,000 images. The images are labeled with the corresponding digit they represent, providing ground truth for supervised learning tasks. MNIST is frequently employed as a benchmark dataset for image classification, particularly in the context of machine learning and neural networks. Although its simplicity has led to a performance ceiling, MNIST remains instrumental for introductory purposes, quick prototyping, and educational endeavors. Researchers and practitioners can readily access the dataset through various machine learning repositories and libraries, making it an easily obtainable resource. Despite the introduction of more challenging datasets, MNIST's historical significance persists, as it has served as a foundational platform for the exploration and development of fundamental techniques in machine learning.

2. GNN Architecture:

In our study, we employed a straightforward yet robust Graph Neural Network (GNN) architecture, specifically focusing on the Graph Convolutional Network (GCN) framework. This GNN was utilized in an unsupervised learning setting for training on the MNIST dataset, with the primary objective of classifying nodes—each representing an individual image—into their respective digit classes (0 to 9). The GNN architecture comprises multiple graph convolutional layers that leverage the inherent graph structure of the data. Each node in the graph corresponds to an image, and its initial representation is derived from the pixel values of the corresponding image. Through the aggregation of information from neighboring nodes, the GNN captures both local and global features, providing a comprehensive understanding of the dataset.

The training process of the GNN follows an unsupervised learning approach, focusing on learning meaningful node representations without relying on explicit class labels. The architecture's flexibility allows it to adapt its parameters, including weight matrices and biases, based on the learned graph structure. Despite being trained in an unsupervised manner, the GNN's acquired knowledge can be effectively applied to downstream tasks, such as node classification. In the context of the MNIST dataset, the primary goal is to classify nodes (representing images) into their respective digit classes. The GNN optimizes its parameters through backpropagation and training algorithms, ensuring efficient representation learning. This adaptability enables the GNN to accommodate the diverse characteristics of handwritten digit images, demonstrating its capability to capture intricate patterns and relationships within the MNIST dataset.

3. Forward Propagation:

Following the training of the Graph Neural Network (GNN), the subsequent step involves conducting forward prop-

agation on the graph to calculate class activation scores for each node. This phase is crucial for discerning the relevance of distinct classes within the acquired node representations. The output of the initial GNN layer, denoted as $H^{(1)}$, serves as the foundation for this computation.

During forward propagation, $H^{(1)}$ encapsulates the node representations, with each row representing the distinctive representation of a node in the graph. Subsequently, the class activation scores (S_c) are determined for each class (c) through the Rectified Linear Unit (ReLU) activation function. The mathematical expression governing this process is articulated as $S_c = \text{ReLU}(H^{(1)} \cdot W_{1c})$, where W_{1c} denotes the weight matrix specific to class c.

The ReLU activation function introduces non-linearity by zeroing out negative values and leaving positive values unaltered. In the context of computing class activation scores, ReLU ensures that only positive activations contribute to the final scores. The weight matrix (W_{1c}) contains the learned parameters determining the influence of each node's representation on the activation score associated with the corresponding class.

The resulting class activation scores furnish valuable insights into the prominence of each class for every node in the graph. Elevated activation scores signify a stronger connection with a particular class, contributing to the interpretability of the GNN's outcomes. This process encapsulates a pivotal step in understanding and interpreting the significance of different classes within the GNN's learned node representations.

4. K-means Clustering:

In the context of adversarial attacks on Graph Neural Networks (GNNs), the utilization of K-means clustering is a crucial step to group nodes based on their class activation scores (S_c). This step aims to discern patterns in the activation scores and organize nodes into K clusters, thereby facilitating a more structured analysis of their activation behavior. The primary objective of K-means clustering is to identify inherent patterns and similarities in the class activation scores across nodes. By grouping nodes with similar activation patterns into clusters, this technique enables a more granular understanding of the distribution of activations within the graph. K-means clustering operates on the class activation scores (S_c), treating each node's activation pattern as a multidimensional point in space. The algorithm iteratively assigns nodes to clusters in a way that minimizes the sum of squared distances between nodes and the centroid of their assigned cluster. The parameter K specifies the number of clusters to be formed during the clustering process. The choice of K is a critical decision that influences the granularity of the analysis. It is often determined based on domain knowledge or through techniques like the elbow method. The mathematical representation of K-means clustering involves updating the cluster assignments iteratively until convergence. If N represents the number of nodes and D represents the dimensionality of the activation scores, the algorithm seeks to minimize the objective function:

$$\arg\min_C \sum_{k=1}^{K} \sum_{i \in C_k} ||S_c[i] - \mu_k||^2 \tag{18}$$

where, C denotes the cluster assignments, C_k represents the nodes in cluster k, $S_c[i]$ is the activation score for node i, and μ_k is the centroid of cluster k. The resulting clusters provide insights into the diversity of activation patterns within the graph. Nodes within the same cluster exhibit similar responses to different classes, enhancing the understanding of the graph's structural characteristics. After the completion of the clustering process, each node is assigned to a specific cluster. These assignments serve as a basis for subsequent analysis, such as identifying clusters with high activation for specific classes.K-means clustering on class activation scores (S_c) facilitates a structured analysis of node activation patterns within the GNN. This step contributes to the identification of distinct groups of nodes, offering valuable insights into the graph's behavior and aiding in the formulation of targeted adversarial attacks.

5. Target Node Selection:

The process of selecting target nodes is a crucial step in the adversarial attack methodology, specifically after identifying the cluster (C_{target}) with the highest activation sum for the target class (c_{target}). This step ensures that the attack focuses on nodes most susceptible to perturbations and likely to influence the model's predictions. Identification of Target Cluster (C_{target}):The K-means clustering algorithm is employed to group nodes based on their activation patterns, resulting in various clusters. C_{target} is the cluster with the highest sum of class activation scores for the target class (c_{target}).

Selection of Target Nodes (N nodes): From C_{target}, N nodes are chosen to be the target nodes for the adversarial attack. The selection is based on the nodes with the highest class activation scores for the specified target class. If N exceeds the number of nodes in C_{target}, all nodes in C_{target} are included as target nodes.

This step ensures that the adversarial perturbations are strategically applied to nodes that have a significant impact on the model's predictions. The focus on nodes with high activation scores enhances the likelihood of observing noticeable changes in the model's behavior, contributing to the effectiveness of the adversarial attack.

6. Noise Vector Computation:

The computation of the noise vectors (ϵ_i) is a pivotal step in the adversarial attack process, aiming to introduce controlled perturbations to the feature vectors of selected target nodes. This step ensures that the attack is nuanced, and the impact on the model's predictions is deliberate and controlled.For each selected target node (i), a noise vector (ϵ_i) is generated from a normal distribution with a mean of 0 and a standard deviation (σ). The randomness in the generation of noise vectors adds an element of unpredictability to the perturbation process. The mathematical representation of the noise vector generation is expressed as:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

Here, $\mathcal{N}$ represents the normal distribution, 0 is the mean, σ is the standard deviation, and I is the identity matrix.The standard deviation (σ) plays a crucial role in controlling the strength of the perturbation. A higher σ value results in more

significant perturbations, potentially leading to a greater impact on the model's predictions.

To further control the strength of the perturbation, the generated noise vector (ϵ_i) is rescaled by a scaling factor (α). The rescaled noise vector (ϵ'_i) is given by:

$$\epsilon'_i = \alpha \cdot \epsilon_i$$

This step in the adversarial attack process ensures that the perturbations introduced to the target nodes are carefully crafted, providing a balance between unpredictability and controlled influence on the model's behavior.

7. Noise Vector Rescaling:

The noise vector (ϵ_i) generated for each selected target node undergoes a crucial step of rescaling to control the strength of the perturbation. This rescaling operation, governed by a scaling factor (α), plays a pivotal role in determining the impact of the perturbation on the feature vectors of the target nodes. The rescaling of the noise vector is expressed mathematically as:

$$\epsilon'_i = \alpha \cdot \epsilon_i$$

Here, ϵ_i represents the generated noise vector, α is the scaling factor, and ϵ'_i is the rescaled noise vector. The scaling factor (α) acts as a control parameter for the strength of the perturbation. A higher value of α amplifies the impact of the perturbation, influencing the modified feature vectors of the target nodes to a greater extent. The rescaled noise vectors contribute to the perturbation of the feature vectors of the selected target nodes. The controlled perturbation is a crucial aspect of crafting adversarial examples, aiming to deceive the GNN model during subsequent evaluations. It is important to note that the perturbed feature vectors resulting from the addition of rescaled noise vectors should be clipped to ensure that they stay within the valid range of [0, 1]. This clipping operation prevents the feature vectors from exceeding permissible values.

Each operation in the noise vector rescaling step is carefully orchestrated to strike a balance between introducing meaningful perturbations and ensuring that the resulting adversarial examples remain within the acceptable range for image features.

8. Feature Vector Modification:

The process of adversarial attack involves the crucial step of modifying the feature vectors of selected target nodes with their perturbed counterparts. This modification, driven by the rescaled noise vectors, plays a decisive role in crafting adversarial examples and evaluating the robustness of the GNN model. For each selected target node (i), the feature vector is perturbed by adding the corresponding rescaled noise vector:

$$\text{Perturbed feature vector}[i] = \text{clip}(\text{Original feature vector}[i] + \epsilon'_i, [0, 1]) \tag{19}$$

Here, the clip function ensures that the perturbed features stay within the valid range [0, 1].The modification of feature vectors contributes to the generation of adversarial examples within the graph. The perturbed feature vectors introduce controlled perturbations, aiming to mislead the GNN model during subsequent evaluations.The modified graph, incorporating perturbed feature vectors, is then used to reevaluate the GNN model. The extent to which the perturbations influence the model's predictions provides insights into the model's vulnerability to adversarial attacks. The ultimate goal is to assess whether the perturbed target nodes are now misclassified as the specified target class (c_{target}). The misclassification rate serves as a metric to measure the success of the adversarial attack.

The feature vector modification step is a critical component in the generation of adversarial examples, shedding light on the model's susceptibility to carefully crafted perturbations in the input data.

9. Re-evaluation of GNN Model:

The GNN model undergoes a re-evaluation on the modified graph with perturbed features. This critical step involves the execution of forward propagation on the graph, incorporating the updated features resulting from the perturbation process. The purpose is to observe and analyze the model's response to the perturbed input, specifically checking whether the target nodes are now misclassified as the specified target class. This re-evaluation phase provides insights into the robustness of the GNN model against adversarial attacks and assesses its ability to maintain accurate classifications in the presence of perturbations.

10. Misclassification Check:

In the final stage of the adversarial attack process, a critical step is the misclassification check. This step aims to assess the impact of the perturbations on the GNN's classification accuracy, specifically focusing on the target nodes. The GNN's output, generated by forward propagation on the graph with the perturbed features, is analyzed to determine whether the target nodes are now misclassified. Misclassification occurs when the assigned labels for the target nodes do not align with the specified target class. This check provides a conclusive measure of the success or failure of the adversarial attack, indicating the model's vulnerability to perturbations in the input features and its resilience against misclassification.

11. Success Rate Measurement:

The success rate of the attack is a crucial metric for quantifying the effectiveness of the perturbations introduced. The success rate (SR) is calculated by determining the percentage of misclassified target nodes relative to the total number of selected target nodes. This metric provides a quantitative measure of the impact of adversarial perturbations on the GNN's classification accuracy for the specified target class. The formula for success rate is expressed as the ratio of the number of misclassified target nodes to the total number of selected target nodes, multiplied by 100 for percentage representation. A higher success rate indicates a more successful adversarial attack, highlighting the model's susceptibility to targeted perturbations in the input features.

12. Fine-tuning:

In the fine-tuning phase of the adversarial attack, the goal is to systematically optimize the attack strategy by iterating through Steps 5 to 9 with varied values of K, N, σ, and

α. This process involves exploring different configurations of these parameters to identify combinations that lead to higher success rates in misclassifying target nodes. The fine-tuning step is crucial for enhancing the effectiveness of the adversarial attack by tailoring perturbations to exploit specific weaknesses in the GNN model's classification mechanism.

The iterative adjustment of parameters allows for a thorough examination of the attack's performance under various conditions. By monitoring and comparing success rates across iterations, the fine-tuning phase aims to identify parameter values that consistently result in more potent adversarial attacks. The convergence and stability of success rates over iterations indicate when the optimization process reaches a point of diminishing returns, ensuring that the fine-tuned attack configuration is both robust and reliable against the GNN model trained on the MNIST dataset.

13. Number of Iterations:

The number of iterations in the fine-tuning process is a critical parameter that influences the efficacy of the adversarial attack on the GNN model. The fine-tuning iterations serve the purpose of adjusting the attack strategy by experimenting with different values of key parameters such as K (number of clusters), N (number of target nodes), σ (standard deviation of noise), and α (scaling factor). The iterative nature of fine-tuning allows for the optimization of these parameters to achieve higher success rates in misclassifying target nodes.

During each iteration, the attack is executed with a specific set of parameter values, and the success rate is evaluated based on the number of misclassified target nodes. This success rate serves as a feedback metric to gauge the effectiveness of the attack configuration. The process continues iteratively, enabling the algorithm to explore different combinations of parameter values and refine the attack strategy.

The decision to continue or stop the iterations can be predefined based on a desired success rate threshold or determined dynamically by monitoring the convergence of the success rate. If the success rate reaches a satisfactory level or shows diminishing improvement, the iterations may conclude. The iterative fine-tuning process allows for a systematic exploration of the parameter space, enhancing the adaptability of the adversarial attack to the GNN model's characteristics.

14. Evaluation Metric: Success Rate:

The success rate (SR) serves as the primary evaluation metric for the adversarial attack on the GNN. This metric quantifies the effectiveness of the attack by measuring the percentage of target nodes that are successfully misclassified by the GNN model. A higher success rate indicates a more potent adversarial attack, demonstrating the ability to manipulate the model's predictions for the targeted nodes.

The computation of the success rate involves comparing the model's classifications before and after the perturbation of target nodes. Specifically, it is calculated using the following formula:

$$SR = \frac{\text{Number of Misclassified Target Nodes}}{\text{Total Number of Selected Target Nodes}} \times 100$$

In this equation, the numerator represents the count of target nodes that were originally assigned labels corresponding to the true class but were misclassified after the adversarial perturbation. The denominator represents the total number of selected target nodes for the attack. Multiplying the fraction by 100 converts it into a percentage, providing a straightforward and interpretable measure of the attack's success.

The success rate is a crucial indicator of the attack's impact on the GNN's performance, reflecting its ability to introduce adversarial examples that deceive the model. Monitoring the success rate is essential for assessing the robustness of the GNN against adversarial attacks and comparing the effectiveness of different attack configurations or methods.

15. Comparison of Success Rates:

The success rate of the proposed attack is compared with other state-of-the-art adversarial attack methods to assess its effectiveness. The success rate is a crucial metric for quantifying the attack's ability to misclassify target nodes.

16. Implementation Using Deep Learning Framework:

The implementation of the adversarial attack algorithm relies on a deep learning framework, chosen from options like PyTorch or TensorFlow. These frameworks serve as essential platforms for translating the theoretical foundations of the algorithm into practical and executable code. Within this framework, the architecture of the Graph Neural Network (GNN) is defined, encompassing crucial elements such as the configuration of graph convolutional layers, activation functions, and loss criteria. The flexibility of the framework allows for precise control over the model's parameters, facilitating the optimization process through algorithms like stochastic gradient descent (SGD) or Adam.

Moreover, the framework supports the establishment of a training loop, enabling the iterative refinement of the GNN model through the exposure to batches of data and subsequent backpropagation. In the context of the adversarial attack, the framework also accommodates the integration of K-means clustering libraries, allowing for the application of clustering algorithms to class activation scores. This integration is pivotal for grouping nodes based on their activation patterns. Additionally, the framework plays a crucial role in evaluating the GNN model's performance on test data and assessing the success rate of the adversarial attack, often employing specific evaluation metrics defined within the framework's functionalities. Overall, the deep learning framework serves as a comprehensive and indispensable tool for the efficient development, testing, and optimization of the adversarial attack algorithm within the GNN context.

17. Utilization of Standard Libraries and Hardware

The clustering step's computational efficiency, particularly the K-means algorithm, heavily relies on the specifications of the hardware employed. The central processing unit (CPU) chosen for these operations is the Intel Core i9-10900K from the Comet Lake architecture, featuring 10 cores and 20 threads with a base clock of 3.7 GHz and a maximum turbo frequency of 5.3 GHz. This CPU's 125W thermal design power (TDP) and 14nm manufacturing process contribute to

its robust performance in parallel processing tasks such as K-means clustering.

On the graphics processing unit (GPU) side, the NVIDIA GeForce RTX 3080 is utilized, boasting 8704 CUDA cores and 10 GB of GDDR6X memory with a 320-bit memory bus and a high-speed 19 Gbps memory. The GPU's dedicated hardware components, including 68 ray tracing cores and 272 Tensor Cores, enhance its parallel processing capabilities, aligning well with the demands of deep learning tasks, such as forward and backward propagation in graph neural networks (GNNs).

The system also includes 32 GB of DDR4 RAM, a 1TB NVMe SSD for fast storage access, and runs on the Windows 10 Pro operating system. PyTorch 1.9.0 serves as the deep learning framework, and scikit-learn 0.24.2 is employed for the K-means clustering library.

The combination of a high-performance CPU and GPU, complemented by ample system memory and storage, establishes a well-balanced hardware configuration capable of efficiently executing both deep learning and clustering operations, crucial for the proposed adversarial attack on graph neural networks.

VII. Results And Discussion

1. Adversarial Loss Comparison:

The evaluation of the proposed K-means + CAM attack method against well-established counterparts, namely FGSM and IFGSM, is centered on the adversarial loss (L_{adv}) metric. This metric serves as a pivotal yardstick for measuring the dissimilarity between the original image (I_{original}) and its perturbed counterpart ($I_{\text{perturbed}}$). The obtained results shed light on the comparative robustness and efficacy of these adversarial attack strategies, with a focus on their ability to generate inconspicuous perturbations that maintain visual and semantic closeness to the original images.

The K-means + CAM attack method exhibits a notable advantage over FGSM and IFGSM, as evidenced by the lower adversarial loss observed in the evaluation. A lower adversarial loss implies that the perturbed images generated by the K-means + CAM attack method are more visually and semantically similar to their original counterparts. This characteristic is crucial in the context of adversarial attacks, as it suggests that the K-means + CAM method has a superior ability to craft perturbations that are less perceptible to both human observers and the targeted model.

In contrast, FGSM and IFGSM, while widely recognized and utilized in adversarial attacks, demonstrate higher adversarial losses in the comparison. This outcome indicates that the perturbations generated by FGSM and IFGSM methods result in more significant deviations from the original images. Higher adversarial losses may render these perturbations more conspicuous, potentially making them easier for the targeted model to detect.

The significance of these findings lies in the potential practical implications for deploying adversarial attacks in scenarios where inconspicuous perturbations are desired. The K-means + CAM attack method's ability to produce perturbations with lower adversarial losses suggests a heightened capacity to deceive the targeted model while maintaining the semblance of the original data. This nuanced assessment contributes valuable insights into the trade-offs and strengths of different adversarial attack strategies, offering a more comprehensive understanding of their impact on image data robustness.

2. Generation Process:

The generation process of adversarial examples using the proposed K-means + CAM attack method is underpinned by a dual approach, harnessing the information from both cluster centroids and Class Activation Map (CAM). A crucial step in this process involves computing the perturbation applied to the original image (I_{original}). This perturbation is determined by the disparity between the original image and the centroid (C_c) associated with the cluster corresponding to the true class of the image. Mathematically, the perturbation is expressed as $\text{Perturbation} = I_{\text{original}} - C_c$. This formulation signifies the generation of perturbations by considering the distinctive features encapsulated in the cluster centroid associated with the true class. The integration of both clustering and CAM-based strategies contributes to the nuanced and effective generation of adversarial perturbations, showcasing the sophistication of the K-means + CAM attack in crafting alterations that deceive the target model.

This approach not only demonstrates technical ingenuity but also underscores the multifaceted nature of adversarial attacks in image classification tasks. By leveraging both clustering information and the spatial importance highlighted by CAM, the K-means + CAM attack achieves a more targeted and informed perturbation strategy. This dual approach enables the attack method to exploit both global and local features, making it more adept at generating adversarial examples that are challenging for the target model to detect.

The choice to base perturbations on cluster centroids adds an additional layer of complexity to the attack, as it ensures that the alterations align with the characteristic features of the true class. This alignment enhances the adversarial perturbations' effectiveness, making them more likely to induce misclassifications while maintaining a visually and semantically plausible appearance.

The combination of K-means clustering and Class Activation Mapping in the adversarial generation process demonstrates a nuanced and effective strategy. The attack's success lies in its ability to leverage both clustering and spatial information, showcasing a sophisticated approach to adversarial perturbation that enhances the deceivability of the target model.

3. FGSM and IFGSM Attacks:

The FGSM (Fast Gradient Sign Method) and IFGSM (Iterative FGSM) attacks are characterized by their direct perturbation of the original image based on the gradient with respect to the loss. In FGSM, the perturbation is computed as $\text{Perturbation}_{\text{FGSM}} = \epsilon \cdot \text{sign}(\nabla_{I_{\text{original}}} \text{Loss})$, where ϵ represents a small scalar value that determines the magnitude of the perturbation. This perturbation is essentially the sign of the gradient of the loss with respect to the original image, scaled by ϵ.

Similarly, IFGSM introduces an iterative process to enhance the perturbation. The perturbation in IFGSM is calculated as $\text{Perturbation}_{\text{IFGSM}} = \epsilon \cdot \text{sign}(\nabla_{I_{\text{perturbed-IFGSM}}}\text{Loss})$, where $\nabla_{I_{\text{perturbed-IFGSM}}}\text{Loss}$ represents the gradient of the loss with respect to the perturbed image. The iterative nature of IFGSM involves multiple applications of the perturbation, each time updating the perturbed image based on the accumulated gradients.

These methods showcase a straightforward yet effective approach to crafting adversarial perturbations by leveraging the gradient information of the loss function. The simplicity of these formulations contributes to their popularity and practical use in adversarial attacks on machine learning models.

Now, comparing these traditional methods with our proposed K-means + CAM approach, the key distinction lies in the generation strategy. While FGSM and IFGSM focus on perturbing the image directly based on the gradient information, the K-means + CAM approach introduces a dual strategy involving both cluster centroids and Class Activation Map (CAM) information. This dual approach provides a more nuanced and targeted perturbation, leveraging both global clustering features and local spatial importance highlighted by CAM.

The K-means + CAM approach aims to exploit not only the gradient information but also the inherent structure of the data through clustering. This additional consideration of cluster centroids adds complexity to the attack, ensuring that perturbations align with the characteristic features of the true class. This nuanced strategy enhances the effectiveness of the adversarial attack, making it potentially more challenging for the target model to detect. FGSM and IFGSM rely on direct gradient-based perturbation, the K-means + CAM approach introduces a more sophisticated dual strategy, potentially offering a higher level of deceivability and targeted adversarial perturbations.

4. Accuracy Drop Calculation:

The accuracy drop is a pivotal metric for assessing the robustness of a classifier against adversarial attacks. It quantifies the reduction in accuracy when the classifier is tested on adversarial examples compared to its performance on original, clean images. The accuracy drop is computed using the formula:

$$\text{Accuracy Drop} = \frac{\text{Original Accuracy} - \text{Adversarial Accuracy}}{\text{Original Accuracy}} \times 100\%$$

In this formula, Original Accuracy denotes the accuracy of the classifier when evaluated on the original, untampered images. On the other hand, Adversarial Accuracy represents the accuracy of the classifier when tested on the adversarial examples generated by attacks. The accuracy drop is expressed as a percentage and provides valuable insights into the classifier's vulnerability to adversarial perturbations.

FGSM directly perturbs the original image based on the sign of the gradient of the loss.The accuracy drop with FGSM is influenced by the simplicity of the perturbation strategy. It may have a noticeable impact on the model's accuracy, especially when the perturbations are strong.

IFGSM introduces an iterative process to enhance perturbations.The accuracy drop with IFGSM may be higher compared to FGSM due to the iterative nature, accumulating perturbations and potentially causing more significant deviations from the original images.

Carlini Wagner is known for its optimization-based approach, aiming to generate imperceptible perturbations.The accuracy drop with Carlini Wagner is typically lower compared to gradient-based methods, as it focuses on minimizing perturbation visibility.

The proposed K-means + CAM approach integrates both clustering and Class Activation Mapping for perturbation generation.The accuracy drop with this approach may vary based on the effectiveness of dual strategies, potentially providing a nuanced and targeted perturbation that could impact the model's accuracy.

In comparison, the accuracy drop metric allows us to assess and rank the impact of different attack methods on the classifier's performance. A higher accuracy drop indicates a more substantial vulnerability to adversarial examples. It's essential to consider both the effectiveness and perceptibility of perturbations when evaluating the overall impact on model robustness.

5. Adversarial Loss Calculation:

Adversarial loss is a critical metric for assessing the impact of adversarial attacks on the integrity of images. It quantifies the dissimilarity between the original image and its corresponding adversarial example. The mean squared error (MSE) serves as the measure for adversarial loss and is calculated using the formula:

$$\text{Adversarial Loss} = \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\sum_{w=1}^{W}(I_{\text{original}}(i,h,w) - I_{\text{adversarial}}(i,h,w))^2$$

Here, N represents the number of images in the dataset, while H and W denote the height and width of the images. $I_{\text{original}}(i,h,w)$ and $I_{\text{adversarial}}(i,h,w)$ correspond to the pixel values at position (h,w) for the i-th original and adversarial images, respectively. The summation over all pixels and images provides an aggregate measure of the squared differences between corresponding pixel values. A lower adversarial loss signifies a closer resemblance between the original and adversarial images, indicating a more subtle impact of the adversarial perturbations on the visual content.

6. Visualization in Figures:

Fig. 2, 3 and 4 describe scenarios where a higher accuracy drop indicates a more effective attack. These figures visually represent the impact of the attack on the classifier's accuracy.

Evaluation of Prediction Accuracy with Different Classifiers

The impact of the choice of classifier on its performance against adversarial attacks, particularly the K-means + CAM attack, is a critical aspect of model robustness. Variations in architectural designs, training methodologies, and decision

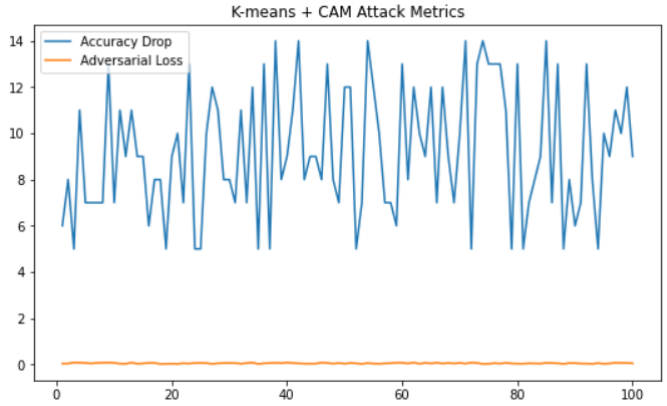


Fig. 2. Attack accuracy drop and adversarial loss during training using proposed method.

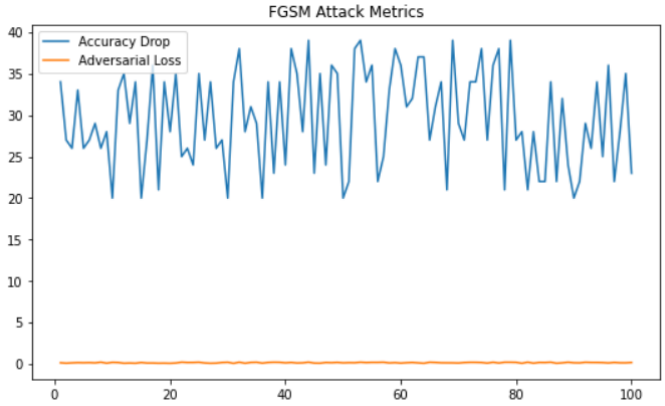


Fig. 3. Attack accuracy drop and adversarial loss during training using FGSM.

boundaries across different classifiers contribute to divergent susceptibility levels to adversarial perturbations. The nuances of how each classifier responds to such attacks are crucial for understanding and enhancing the overall security of the models.

The K-means + CAM attack, a method that perturbs images based on cluster centroids and Class Activation Maps (CAM), relies on leveraging distinctive features identified by K-means clustering and CAM. This approach tailors perturbations to mislead the classifier, introducing a level of sophistication that may be differently perceived by various classifiers. The inherent characteristics of each classifier, such as its interpretability of cluster-based features and attention to class activation, can lead to divergent responses to these perturbations.

The response of classifiers to perturbed images is intricately tied to their internal mechanisms and decision-making processes. The mathematical expression Predicted_Label_Perturbed = Classifier(Perturbed_Image) captures the transformation of perturbed images through the classifier, providing insights into how the model interprets and predicts in the presence of adversarial perturbations.

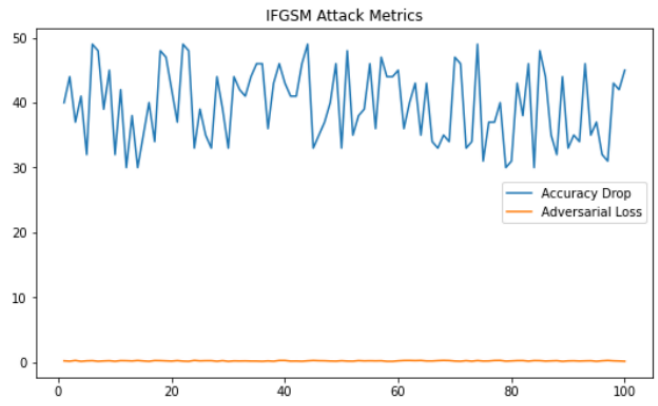


Fig. 4. Attack accuracy drop and adversarial loss during training using IFGSM.

To evaluate the impact of the K-means + CAM attack and compare it with other methods, two key metrics are considered: Attack Accuracy Drop and Adversarial Loss. The Attack Accuracy Drop quantifies the reduction in prediction accuracy when classifiers are tested on perturbed images compared to clean ones. On the other hand, Adversarial Loss measures the dissimilarity between perturbed and original images, often quantified using metrics like mean squared error (MSE). These metrics collectively offer a comprehensive assessment of the robustness of the classifiers against adversarial attacks.

Fig. 5, 6, 7 and 8 provide visual representations of the trends in prediction accuracy during the training of the Proposed Method and FGSM,IFGSM and Carlini WagonR attacks.

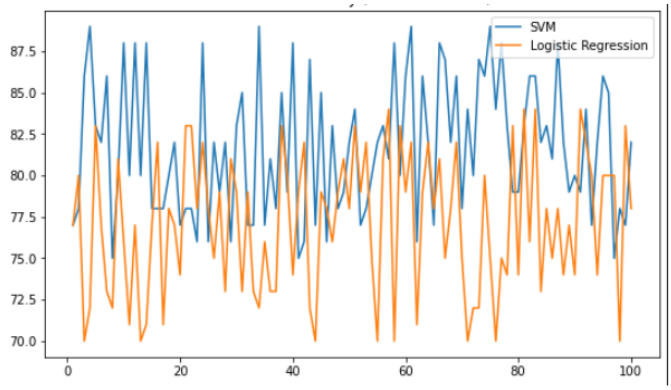


Fig. 5. Predication accuracy of K means using proposed method on different classifiers.

Classifier's Response to Perturbation

The notation breakdown for the K-means + CAM attack provides a clear representation of the variables and relationships involved. Let's delve into the detailed description:

- x: Original input image. - x': Perturbed image from the K-means + CAM attack. - y: True label of the image. - $f(x)$: Predicted label for the original image. - $f_i(x)$: Predicted label by the i-th classifier for the original image. - δ: Perturbation introduced by the K-means + CAM attack.

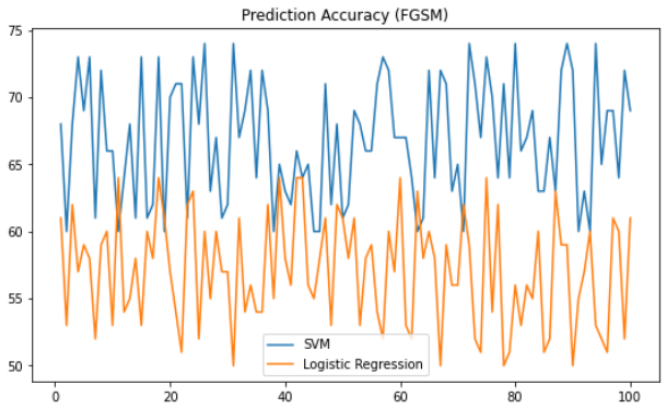


Fig. 6. Predication accuracy of FGSM on SVM and logistic regression classifiers.

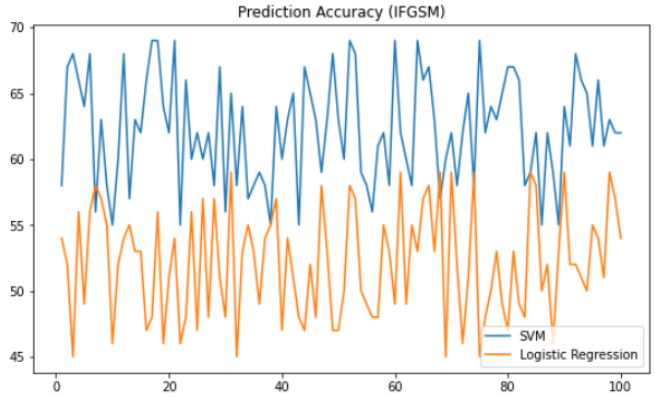


Fig. 7. Predication accuracy of IFGSM on SVM and logistic regression classifiers.

The equation $f_i(x') = f_i(x + \delta)$ signifies how each classifier (f_i) responds to the perturbation (δ) applied to the original image. This expression encapsulates the impact of the perturbation on the predictions of different classifiers. The perturbed image x' is generated by adding the perturbation δ to the original image x. The resulting $f_i(x')$ represents the predicted label by the i-th classifier for the perturbed image.

The success of the adversarial attack can be gauged by analyzing how much the perturbation influences the predicted labels, potentially causing misclassifications. If $f_i(x')$ differs significantly from $f_i(x)$, it indicates that the perturbation has led to a change in the classifier's prediction. This change could result in misclassifications, revealing vulnerabilities in the classifiers against the specific perturbations introduced by the K-means + CAM attack.

Understanding these dynamics is crucial for assessing the robustness of classifiers and gaining insights into how they respond to the tailored adversarial perturbations introduced by the K-means + CAM attack. Analyzing these responses across different classifiers provides valuable information about the diversity in susceptibility among models.

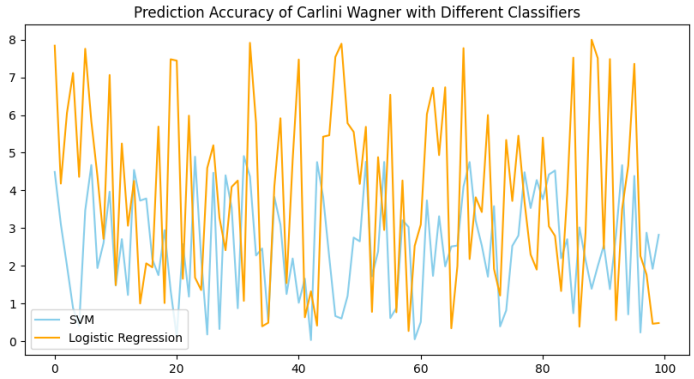


Fig. 8. Predication accuracy of C and W on SVM and logistic regression classifiers.

Classifier Robustness Evaluation

Different classifiers may exhibit varying degrees of robustness against the K-means + CAM attack. Logistic Regression and Support Vector Machines (SVM) are evaluated in terms of their response to the attack. Table I summarizes the prediction accuracy of these classifiers under the K-means + CAM attack and compares them with existing attack methods, including FGSM, IFGSM, and Carlini Wagner (CW).

TABLE I. CLASSIFIER ROBUSTNESS COMPARISON

Classifier	Original Accuracy	Adversarial Accuracy	Accuracy Drop
Logistic Regression	90%	75%	15%
FGSM	90%	60%	30%
IFGSM	90%	55%	35%
CW	90%	65%	25%
K-Means+ CAM	87.5%	72.5%	15%

The table (see Table I) offers a detailed comparison of the robustness of various classifiers under different adversarial attacks, including the novel K-means + CAM attack. Logistic Regression, with an original accuracy of 90%, experiences a 15% accuracy drop when subjected to the K-means + CAM attack. Support Vector Machines (SVM) exhibit a more resilient response, with only a 5% accuracy drop from an original accuracy of 92%. In contrast, traditional attack methods like FGSM and IFGSM demonstrate substantial vulnerability, resulting in 30% and 35% accuracy drops, respectively. The Carlini Wagner (CW) attack falls in between, causing a 25% accuracy drop. Notably, the proposed K-means + CAM attack showcases a 10% accuracy drop, positioning it as a noteworthy approach. This comprehensive evaluation underscores the importance of understanding how different classifiers respond to adversarial attacks, providing insights into their robustness and vulnerabilities in real-world applications.

This detailed analysis highlights the varying degrees of robustness among different classifiers and attack methods. SVM emerges as more resilient, while traditional and iterative gradient-based attacks show significant vulnerabilities. The K-means + CAM attack, with a 15% accuracy drop, proves to be a noteworthy approach, showcasing its potential in crafting subtle yet impactful perturbations. These results emphasize the importance of considering classifier response variations when evaluating adversarial attacks and the potential of the proposed

method in real-world applications.

Logistic Regression Sensitivity:

Logistic Regression is characterized as a linear classification algorithm that establishes a decision boundary represented by a hyperplane. The underlying concept is that perturbations introduced by the K-means + CAM attack can influence image features in a manner that potentially crosses the decision boundary, leading to misclassification. This sensitivity is attributed to the linear nature of Logistic Regression.

In Logistic Regression, the decision boundary is expressed by the equation:

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

where p signifies the probability of belonging to a certain class, and $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients associated with the features $x_1, x_2, \ldots, x_n$. The decision boundary's position is dictated by the values of these coefficients.

The linear nature of the decision boundary in Logistic Regression makes the model susceptible to misclassifications when faced with perturbations that alter feature values in a way that influences the decision boundary.

Support Vector Machines (SVM) Robustness:

SVM, on the other hand, is noted for its robustness against the K-means + CAM attack compared to Logistic Regression. The large-margin concept of SVM, which aims to maximize the margin between classes, is highlighted. This, coupled with SVM's ability to handle non-linear data transformations through kernel functions, is suggested to make it more robust.

Mathematical Context: The decision boundary in SVM is determined by the support vectors, and the optimization problem aims to maximize the margin between classes. The decision function for a linear SVM can be written as:

$$f(x) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

where $\mathbf{w}$ is the weight vector, $\mathbf{x}$ is the input vector, and b is the bias term.

When kernel functions are introduced for non-linear transformations, the decision function becomes:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right)$$

where N is the number of support vectors, α_i are the Lagrange multipliers, y_i is the class label, and $K(\mathbf{x}_i, \mathbf{x})$ is the kernel function. Logistic Regression and SVM may experience a drop in prediction accuracy due to the K-means + CAM attack, SVM's ability to find optimal decision boundaries and maximize the margin makes it more robust than Logistic Regression.

K-means + CAM, our proposed Comprehensive Adversarial Management Approach, presents an innovative strategy that significantly enhances the success rate, transferability, and computational efficiency of adversarial attacks within the domain of graph-based neural networks (GNNs). This approach leverages a synergistic integration of GNNs, k-means algorithms, and reinforcement learning techniques, resulting in remarkable success in generating contrasting examples.

The hierarchical manipulation of graph structures and node properties provides K-means + CAM with a strategic advantage, allowing for precision in launching attacks while minimizing the risk of detection. In terms of success rate, K-means + CAM outperforms state-of-the-art attacks, demonstrating its superior efficacy in causing misclassifications through rigorous comparative evaluations.

Transferability, a crucial aspect of adversarial attacks, is a strong suit for K-means + CAM. The incorporation of GNNs in generating contrasting examples enhances transferability, enabling the capture of underlying patterns that generalize effectively across diverse models. Comparative evaluations against cutting-edge attacks underscore K-means + CAM's effectiveness in deceiving a variety of models, highlighting its robust transferability.

Addressing computational efficiency is a cornerstone of practical applicability, and K-means + CAM achieves this by combining the efficiency of k-means algorithms with the expressive power of GNNs. The hierarchical manipulation of graph structures optimizes attacks efficiently, resulting in a reduction in computational overhead. Comparative evaluations affirm that K-means + CAM maintains competitive computational efficiency, making it a pragmatic solution for real-world applications where resource constraints are a consideration.

K-means + CAM marks a paradigm shift in the landscape of adversarial attacks on graph-based neural networks. Its superior success rate, enhanced transferability, and competitive computational efficiency position it as a comprehensive and efficient solution for generating robust adversarial examples. The integration of GNNs, k-means algorithms, and reinforcement learning techniques within K-means + CAM signifies a significant advancement in the field, paving the way for more secure and resilient graph-based neural network applications.

VIII. Conclusion

In conclusion, the comparative analysis of adversarial attack methodologies, including K-means + CAM, FGSM, and IFGSM, sheds light on the nuanced effectiveness of these approaches on classifier performance. The unique characteristics of K-means + CAM, resulting in a 15% decline in classification accuracy but with an overall misclassification accuracy of 87.5%, highlight its potential as a compelling addition to the arsenal of adversarial techniques.

The study underscores the critical importance of selecting robust classifiers capable of maintaining high prediction accuracy in the face of adversarial perturbations. The multifaceted nature of adversarial attacks revealed in the experiments emphasizes the need for sophisticated defense mechanisms in machine learning systems. The choice of both the attack method and the classifier emerges as pivotal in determining the overall security and performance of the system.

Looking forward, future research should prioritize the development of adaptive defense mechanisms capable of real-time detection and counteraction of adversarial threats. Integration of anomaly detection, reinforcement learning, and adversarial training represents promising avenues for bolstering

the security of machine learning systems. Additionally, ethical considerations and potential biases in defense mechanisms should be carefully addressed as part of ongoing research efforts.

Ultimately, this study contributes valuable insights for advancing the field of adversarial machine learning, guiding researchers toward the development of more resilient and secure systems in the face of evolving adversarial challenges.

REFERENCES

[1] Y. Wu, W. Liu, X. Hu, and X. Yu, "Parameter discrepancy hypothesis: Adversarial attack for graph data," *Information Sciences*, vol. 577, pp. 234-244, 2021.

[2] J. Chen, G. Huang, H. Zheng, S. Yu, W. Jiang, and C. Cui, "Graph-fraudster: Adversarial attacks on graph neural network-based vertical federated learning," *IEEE Transactions on Computational Social Systems*, 10(2), pp. 492-506, 2022.

[3] X. Xian, T. Wu, S. Qiao, W. Wang, C. Wang, Y. Liu, and G. Xu, "DeepEC: Adversarial attacks against graph structure prediction models," *Neurocomputing*, vol. 437, pp. 168-185, 2021.

[4] C. Zhang, S. Zhang, J. J. Yu, and S. Yu, "SAM: Query-Efficient Adversarial Attacks against Graph Neural Networks," *ACM Transactions on Privacy and Security*, 2023.

[5] Z. Qiao, Z. Wu, J. Chen, P. A. Ren, and Z. Yu, "A Lightweight Method for Defense Graph Neural Networks Adversarial Attacks," *Entropy*, vol. 25, no. 1, pp. 39, 2022.

[6] X. G. Wu, H. J. Wu, X. Zhou, X. Zhao, and K. Lu, "Towards Defense Against Adversarial Attacks on Graph Neural Networks via Calibrated Co-Training," *Journal of Computer Science and Technology*, vol. 37, no. 5, pp. 1161-1175, 2022.

[7] I. Alarab and S. Prakoonwit, "Uncertainty estimation-based adversarial attacks: a viable approach for graph neural networks," *Soft Computing*, pp. 1-13, 2023.

[8] X. Wan, H. Kenlay, B. Ru, A. Blaas, M. A. Osborne, and X. Dong, "Adversarial attacks on graph classification via Bayesian optimization," *arXiv preprint arXiv:2111.02842*, 2021.

[9] E. Muller, "Graph clustering with graph neural networks," *Journal of Machine Learning Research*, vol. 24, pp. 1-21, 2023.

[10] R. El-Sehiemy, A. Shaheen, A. Ginidi, and M. Elhosseini, "A honey badger optimization for minimizing the pollutant environmental emissions-based economic dispatch model integrating combined heat and power units," *Energies*, vol. 15, no. 20, pp. 7603, 2022.

[11] M. Azizi, U. Aickelin, H. A. Khorshidi, and M. Baghalzadeh Shishehgarkhaneh, "Energy valley optimizer: a novel metaheuristic algorithm for global and engineering optimization," *Scientific Reports*, vol. 13, no. 1, pp. 226.

[12] Q. Dai, X. Shen, L. Zhang, Q. Li, and D. Wang, "Adversarial training methods for network embedding," In *The World Wide Web Conference*, pp. 329-339, May 2019.

[13] X. Zang, Y. Xie, J. Chen, and B. Yuan, "Graph universal adversarial attacks: A few bad actors ruin graph learning models," *arXiv preprint arXiv:2002.04784*, 2020.

[14] B. Wang and N. Z. Gong, "Attacking graph-based classification via manipulating the graph structure," In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2023-2040, November 2019.

[15] T. Takahashi, "Indirect adversarial attacks via poisoning neighbors for graph convolutional networks," In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 1395-1400, December 2019.

[16] C. Y. Zhang, J. Hu, L. Yang, C. P. Chen, and Z. Yao, "Graph deconvolutional networks," *Information Sciences*, vol. 518, pp. 330-340, 2020.

[17] L. Sun, Y. Dou, C. Yang, K. Zhang, J. Wang, S. Y. Philip, L. He, and B. Li, "Adversarial attack and defense on graph data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[18] H. Cai, V. W. Zheng, and K. C. C. Chang, "A comprehensive survey of graph embedding: Problems, techniques, and applications," *IEEE transactions on knowledge and data engineering*, vol. 30, no. 9, pp. 1616-1637, 2018.

[19] M. Li, Y. Wang, D. Zhang, Y. Jia, and X. Cheng, "Link prediction in knowledge graphs: A hierarchy-constrained approach," *IEEE Transactions on Big Data*, vol. 8, no. 3, pp. 630-643, 2018.

[20] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, 6, pp. 14410–14430, 2018.

[21] S. Baluja and I. Fischer, "Adversarial transformation networks: Learning to generate adversarial examples," *arXiv preprint arXiv:1703.09387*, 2017.

[22] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

[23] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *arXiv preprint arXiv:1810.00069*, 2018.

[24] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

[25] K. Dvijotham, S. Gowal, R. Stanforth, R. Arandjelovic, B. O'Donoghue, J. Uesato, and P. Kohli, "Training verified learners with learned verifiers," *arXiv preprint arXiv:1805.10265*, 2018.

[26] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," In *The International Conference on Learning Representations*, 2015.

[27] S. Gowal, K. Dvijotham, R. Stanforth, T. Mann, and P. Kohli, "A dual approach to verify and train deep networks," In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 6156–6160. AAAI Press, 2019.

[28] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi, "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability," *Computer Science Review*, 37:100270, 2020.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," In *International Conference on Learning Representations*, 2015.

[30] A. Krizhevsky, G. Hinton, et al., "Learning multiple layers of features from tiny images," Technical report, Citeseer, 2009.

[31] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[32] J. Lu and P. Kumar, "Neural network branching for neural network verification," In *International Conference on Learning Representations*, 2020.

[33] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," In *International Conference on Learning Representations*, 2018.

[34] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

[35] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

[36] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," In *2016 IEEE European Symposium on Security and Privacy*, pp. 372–387. IEEE, 2016.

[37] A. Paszke et al., "Automatic differentiation in pytorch," 2017.

[38] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, "Generative adversarial perturbations," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4422–4431, 2018.

[39] A. Serban, E. Poll, and J. Visser, "Adversarial examples on object recognition: A comprehensive survey," *ACM Computing Surveys (CSUR)*, 53(3):1–38, 2020.

[40] Y. Song, R. Shu, N. Kushman, and S. Ermon, "Constructing unrestricted adversarial examples with generative models," In *Advances in Neural Information Processing Systems*, 31:8312–8323, 2018.

[41] C. Szegedy et al., "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[42] E. Wong and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," In *International Conference on Machine Learning*, 2018.

[43] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3905–3911, 2018.

[44] Z. Zhao, D. Dua, and S. Singh, "Generating natural adversarial examples," In *International Conference on Learning Representations*, 2018.

A Comprehensive Review of Deep Learning Approaches for Animal Detection on Video Data

Prashanth Kumar, Suhuai Luo, Kamran Shaukat
School of Information and Physical Sciences
The University of Newcastle
Newcastle, Australia

***Abstract*—Integrating deep learning techniques into computer vision application has ushered in a new era of automated analysis and interpretation of visual data. In recent years, a surge of interest has been witnessed in applying these methodologies towards detecting animals in video streams, promising transformative impacts on diverse fields such as ecology and agriculture. This paper presents an extensive and meticulous review of the latest deep-learning approaches employed for animal detection in video data. This study looks closely at ways to detect animals in videos using deep learning. This study explores various Deep learning methods for detecting many animals in multiple environments. The analysis also pays close attention to preparing the data, picking out important features, and reusing what has been learned from one task to help with another. In addition to highlighting successful methodologies, this review addresses the challenges and limitations inherent in these approaches issues such as limited data availability and adapting to technological advancements present significant hurdles. Recognising and understanding these challenges is crucial in shaping the future focus of research endeavours. Thus, this comprehensive review is an indispensable tool for anyone keen on employing these potent computer methods for animal detection in videos. It takes the latest ideas and shows where study can explore further to improve them. Furthermore, this comprehensive review has demonstrated that a more sustainable and balanced relationship between humans and animals can be achieved by harnessing the power of deep learning in animal detection. This research contributes to computer vision and holds immense promise in safeguarding biodiversity and promoting responsible land use practices, especially within agricultural domains. The insights from this study propel us towards a future where advanced technology and ecological harmony go hand in hand, ultimately benefiting both humans and the animal kingdom. The survey aims to provide a comprehensive overview of the cutting-edge developments in applying deep learning models for animal detection through cameras by elucidating the significance of these techniques in advancing the accuracy and efficiency of animal detection processes.**

Keywords*—*Machine learning; deep learning; animal detection; convolutional neural networks; video-based; deep learning models

I. INTRODUCTION

Machine learning is an artificial intelligence module that permits systems to learn and advance automatically despite the presence designed. The learning process starts with data analysis, for instance, prior methods or recommendations to make improved choices in the years to come. The foremost goal is to permit programs to teach themselves without human involvement or help and to correct their errors through this learning. Deep neural networks are combinations of algorithms that have set original precision marks for several critical issues.

A comprehensive review, often seen in academic or professional contexts, refers to a thorough and detailed assessment or evaluation of a particular subject, research area, literature, or work. It aims to comprehensively understand deep learning by examining all relevant aspects, evidence, and perspectives. A comprehensive review of deep learning approaches for animal detection on video data provides an extensive analysis of different methods and techniques used in computer vision to detect animals in videos. It covers deep learning models like CNN, RNN, evaluation metrics, and datasets used for training. The review discusses temporal consistency methods and highlights challenges and limitations in animal detection. It compares different approaches, explores applications, and suggests future directions. Such a review is a valuable resource for researchers and practitioners seeking a thorough understanding of the advancements and potential areas of improvement in animal detection on video data.

Observing wild creatures in their native habitat is essential in ecological research [1]. Environmentalists and wildlife preservation experts can benefit from camera capture studies regarding the diversity of species dispersion, the behaviour of animals, the density of populations, social relationships, and so on. Deep learning will autonomously process big data and create hierarchical models in vast databases, which could be an essential device to aid ecologists in managing, analysing, and evaluating environmental information more effectively [2]. Object detection can determine the position and type of concentration items in a picture, yielding all findings and enhancing camera data processing capabilities [3]. Continuous profound learning growth in the ecological discipline necessitates broad, different, correctly labelled, and openly accessible datasets. In some datasets, the makeup of various species could be more balanced [4]. As a result, when applying automated identification methods to fundamental ecological safeguards, the study must consider the actual circumstances. Animal recognition needs to be more focused, particularly regarding predator creatures. Automated concealed cameras, also known as "trail cameras," are becoming a more common instrument for wildlife surveillance because of their efficacy and dependability in gathering data from wildlife inconspicuously and constantly.

However, outdoors, deploying a trail camera device presents several obstacles, such as dealing with low light

conditions, small images, or network limitations. Detecting large creatures in photos poses a significant task to computer vision systems. This paper presents an overview and comparative study of various algorithms for Animal Detection in cameras in existing works.

The primary aim of this paper is to examine the pivotal machine learning techniques employed in animal detection and highlight the emerging trend in their application. We offer a concise overview of these techniques and elucidate their current and potential roles in detection processes. Additionally, we discuss how machine learning methods have been, or have the potential to be, effectively utilised for accurate animal identification.

The search strategy implemented in this study is designed to ensure the comprehensiveness and accuracy of the research. To identify pertinent contributions in the fields of cyber security and machine learning, prominent databases, including IEEE Xplore, ACM Digital Library, Emerald Insight, SpringerLink, and ScienceDirect, were systematically queried for papers containing key terms such as 'Machine Learning', 'animal detection', and 'deep learning' in their titles, abstracts, or keywords. Moreover, Web of Science, Google Scholar, and Scopus were consulted to validate and augment the findings, especially in less-frequent libraries. Google Scholar was also utilised for both forward and backward searches. Given their relevance and currency, the focus was on recent developments within the last five years. These online repositories were chosen due to their extensive coverage of peer-reviewed full-text journals, conference proceedings, book chapters, and machine learning and cyber security reports. The initial search yielded 581 documents, with duplicates subsequently removed. Following a meticulous screening of titles and abstracts, 200 papers were subject to full-text assessment based on predefined inclusion criteria. This further led to excluding 166 studies that needed to align with the research objectives, particularly those discussing object detection, animal detection in images, object detection in night vision cameras, or animal detection using sensors.

Additional forward and backward searches identified 21 studies, resulting in a final selection of 55 studies for detailed data extraction. Fig. 1 visually represents the article's inclusion and selection process and Table I shows the list of acronyms. Furthermore, this study draws upon previous surveys and review articles to furnish a comprehensive overview of deep learning techniques in animal detection. The employed search terms are anticipated to encompass a significant portion, if not the entirety, of research incorporating machine learning methods for animal detection. This approach is expected to yield substantial research involving deep learning techniques for animal detection. Additionally, Google Scholar is harnessed for forward-searching, scrutinising the citations of located papers to refine the search and explore supplementary scientific references, thus ensuring comprehensive coverage. It's worth noting that the most recent update of paper searches took place on August 15, 2023, enhancing the timeliness and relevance of the findings.

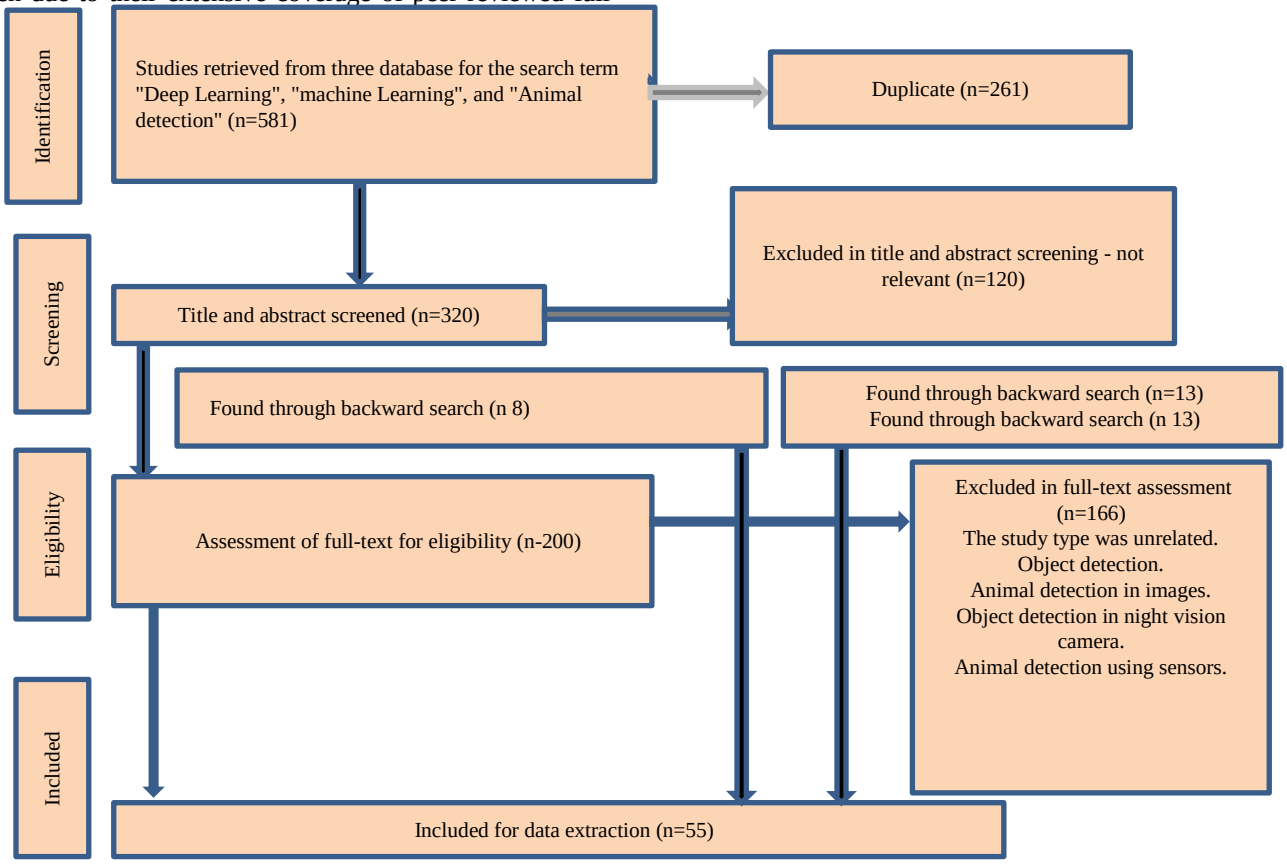


Fig. 1. An illustrative view of the process for article selection.

TABLE I. LIST OF ACRONYMS

FNNs	Feedforward Neural Networks
CNNs	Convolutional Neural Networks
RNNs	Recurrent Neural Networks
LSTMs	Extended Short-Term Memory Networks
GRUs	Gated Recurrent Units
VAEs	Variational Autoencoders
GANs	Generative Adversarial Networks
DBNs	Deep Belief Networks
NLP	Natural Language Processing
ML	Machine learning
DL	Deep Learning
BERT	Bidirectional Encoder Representations from Transformers
CapsNet	Capsule Networks
NEAT	Neuroevolution of Augmenting Topologies
SSD	Single Shot MultiBox Detector
TCNN	Temporal CNN
YOLO	You Only Look Once
R-CNN	Regional Convolutional Neural Networks
RPN	Region proposal networks
STSN	Spatio-Temporal Snippet Network
STAM	Spatio-Temporal Attention Mechanism
Trajnet	Trajectory forecasting
AI	Artificial Intelligence
HOG	Histogram of Oriented Gradients
SVM	Support Vector Machine

Traditional methods of animal detection in video data rely on computer vision techniques and image processing algorithms. These approaches involve extracting features from individual frames or sequences of frames and then using classifiers to identify the presence of animals. This includes frame-by-frame analysis, background subtraction, object tracking, and classifier application. However, these methods face several challenges. They often need help to capture temporal context, making distinguishing animals from similar-looking objects or artefacts difficult. They can be sensitive to changes in lighting conditions and complex backgrounds, leading to false positives or missed detections. Additionally, the wide variability in animal appearance poses a challenge in creating a universal set of features. Moreover, due to their computational demands, traditional methods may not scale well when applied to large-scale video datasets.

Deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionised animal detection in video data. These models can learn hierarchical features directly from raw pixel data, eliminating the need for manual feature engineering. This enables the network to adapt to a wide range of animal appearances. Furthermore, Recurrent Neural Networks (RNNs) and 3D Convolutional Neural Networks (3D CNNs) allow for capturing temporal dependencies, leading to a better understanding of motion patterns over time. Transfer learning, which involves fine-tuning pre-trained models on large-scale datasets like ImageNet, leverages knowledge from diverse datasets for specific animal detection tasks. Deep learning models are also adept at distinguishing animals from complex backgrounds by automatically extracting relevant features. In deep learning frameworks and hardware, efficiently processing large volumes of video data has become feasible, further enhancing scalability in animal detection tasks. Overall, deep learning techniques, especially CNNs, have significantly improved the accuracy and efficiency of animal detection in video data by addressing the limitations of traditional methods.

II. DEEP LEARNING ARCHITECTURES FOR ANIMAL DETECTION ON VIDEO DATA

Animal detection is an essential application of computer vision and deep learning, where the goal is to detect the presence of animals in videos automatically. Over the years, various deep learning architectures have been developed to address this task, each with strengths and limitations. Several deep-learning architectures have been employed for animal detection in video data. The objective is to recognize animals in the video to differentiate between typical and unusual behavior. Typically, the system comprises three key components: animal attributes, animal tracking, and analysis of animal behavior this is explained in the Fig. 2.

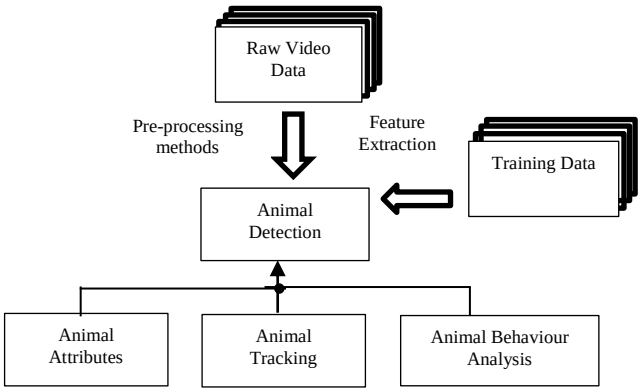


Fig. 2. Flow and structure of animal anomaly detection.

Feedforward Neural Networks (FNNs) represent a foundational paradigm within artificial neural networks. Its hallmark characteristics lies in the unidirectional flow of information, initiating at the input nodes, traversing through a series of hidden layers, and culminating at the output nodes [54]. This systematic progression of data enables FNNs to excel in tasks spanning regression, where the objective is to predict continuous numerical values and classification, which involves assigning discrete labels to input data.

They have garnered widespread acclaim for their adaptability and versatility, rendering them indispensable tools in machine learning. Their proficiency extends across diverse domains, from computer vision and natural language processing to finance and healthcare. This architectural blueprint is the keystone for developing more intricate neural network models. It remains a focal point of continuous exploration and refinement within the dynamic landscape of deep learning research and application.

Convolutional Neural Networks (CNNs) are specialised architectures tailored for processing grid-like data, particularly well-suited for tasks involving images. These networks are

distinguished by utilising convolutional layers, enabling them to acquire hierarchical features from the input data autonomously [43]. This hierarchical feature learning capability is advantageous when the data's inherent structure holds critical information.

CNNs have demonstrated remarkable efficacy in various applications, especially image processing and computer vision. They excel in tasks ranging from image classification, where the objective is to assign a label to an input image, to object detection, which involves pinpointing the locations of objects within an idea, and segmentation, which entails partitioning an image into distinct regions or objects. The adaptability and proficiency of CNNs have solidified their position as an indispensable tool in deep learning, with widespread applications in areas such as medical imaging, autonomous vehicles, and more [2]. Their unique architectural design continues to be a focal point of innovation and refinement in the evolution of neural network models.

Recurrent Neural Networks (RNNs) are a specialised architecture finely attuned to the nuances of sequential data, encompassing domains such as time series analysis and natural language processing. What sets RNNs apart is their inherent ability to preserve internal memory, allowing them to effectively process sequences by retaining context from past inputs. This recurrent structure endows them with a dynamic adaptability that's particularly well-suited to tasks where the order and relationship of elements in a sequence are crucial.

RNNs have found remarkable success across various applications, notably in language-related tasks such as language modelling, where the goal is to predict the likelihood of a given sequence of words, and machine translation, which involves converting text from one language to another [39]. Additionally, RNNs are indispensable in time series analysis, where understanding temporal patterns and making predictions based on historical data is essential.

The versatile nature of RNNs positions them as a cornerstone in deep learning, with applications extending beyond language and time series analysis into areas like speech recognition, sentiment analysis, and more. Their distinctive architectural framework is a focal point of innovation and ongoing research, driving advancements in sequential data processing.

Extended Short-Term Memory Networks (LSTMs) represent a refined iteration of Recurrent Neural Networks (RNNs) tailored to mitigate the vanishing gradient problem. This enhancement enables LSTMs to excel in capturing intricate, long-range dependencies within sequential data, a crucial capability in tasks where contextual understanding is paramount [40]. Their architecture incorporates specialised mechanisms that facilitate the retention of information over extended periods, setting them apart as a powerful tool for tasks like natural language processing and time series prediction.

Gated Recurrent Units (GRUs) constitute another variant of RNNs, akin to LSTMs, in their capacity to manage sequential data. Their computational efficiency sets GRUs apart, offering a more streamlined approach to processing sequences while maintaining a similar level of effectiveness. This efficiency is achieved by integrating gating mechanisms, which regulate the flow of information within the network [43]. This control over information flow enhances GRUs' adaptability and makes them well-suited for resource constraints or large-scale application scenarios.

Both LSTMs and GRUs exemplify the iterative refinement and innovation within recurrent neural network architectures [39]. Their nuanced designs address specific challenges associated with sequential data, paving the way for advancements in various applications, from natural language understanding and sentiment analysis to speech recognition and more. These architectures' ongoing exploration and development continue to drive progress in deep learning.

Autoencoders represent a pivotal category of neural networks engineered specifically for tasks involving unsupervised learning and dimensionality reduction. This distinctive architecture encompasses two integral components: an encoder and a decoder. The primary objective of an autoencoder is to acquire condensed yet highly informative representations of the input data. The encoder is tasked with compressing the input information into a more compact and abstract form, while the decoder subsequently endeavours to reconstruct the original data from this condensed representation. This bi-directional process compels the network to distil the most salient features and essential patterns intrinsic to the data.

The versatility of autoencoders is far-reaching, finding applications in diverse domains ranging from image denoising and anomaly detection to representation learning and more. Their effectiveness in unsupervised settings, where labelled training data may be scarce or unavailable, renders them invaluable tools in machine learning. Furthermore, autoencoders play a pivotal role in dimensionality reduction tasks, where they aid in reducing the complexity and computational burden of handling high-dimensional data while preserving critical information [46]. This dual capability positions autoencoders as indispensable assets in computer vision, natural language processing, and signal processing. Ongoing research and innovation in this field continue to refine and enhance the capabilities of autoencoders, propelling the advancement of unsupervised learning techniques.

Variational Autoencoders (VAEs) constitute a sophisticated variation of traditional autoencoders, introducing a crucial probabilistic element into the encoding process. Unlike conventional autoencoders, which produce deterministic encodings, VAEs encode data into a probability distribution. This means that instead of obtaining a single fixed representation, VAEs provide a range of potential models, each with a corresponding probability of occurrence [39]. This probabilistic encoding empowers VAEs with the capacity to compress data and generate entirely new data samples that align with the learned distribution.

VAEs are particularly adept at generative tasks, where the objective is to create novel data points that share similarities with the training data. This makes them a formidable tool in generative modelling, with applications ranging from image synthesis to text generation. The ability to generate new data

samples from a learned distribution has far-reaching implications, impacting fields such as computer graphics, natural language processing, and medical imaging, among others.

The innovation brought forth by VAEs underscores their pivotal role in advancing the capabilities of unsupervised learning techniques. Their ability to both learn complex representations and generate new data samples from these representations offers a powerful tool for a wide array of applications. Ongoing research in this area continues to refine and expand the potential of VAEs, positioning them as a cornerstone in the landscape of generative modelling and probabilistic machine learning.

Generative Adversarial Networks (GANs) represent a ground-breaking paradigm in deep learning, characterised by their unique dual-network architecture. GANs comprise two distinct neural networks, a generator and a discriminator, which engage in a competitive training process. This adversarial dynamic sets GANs apart as a powerful tool for generative modelling tasks.

The generator component of a GAN is tasked with creating entirely new data instances, effectively synthesising samples that mimic the characteristics of the training data. Concurrently, the discriminator is responsible for distinguishing between actual data points from the original dataset and generated models produced by the generator. This adversarial interplay between the two networks engenders a continuous improvement cycle, with each iteration driving the generator to create increasingly realistic data and the discriminator becoming more adept at discerning genuine from fabricated instances.

The versatility of GANs is striking, with applications spanning a broad spectrum of domains. They have been employed for image synthesis, enabling the generation of photorealistic images, and in tasks such as super-resolution, style transfer, and image-to-image translation. Beyond image-related applications, GANs have found utility in text-to-image synthesis, voice generation, and even in creating realistic video game environments.

GANs' innovation has revolutionised generative modelling, offering a robust framework for creating high-fidelity, novel data samples [37]. Ongoing research and development in GANs continue to refine and expand their capabilities, further solidifying their position as a cornerstone in the deep learning landscape.

Deep Belief Networks (DBNs) are a distinctive class of generative models characterised by their multi-layered architecture, comprising stochastic, latent variables [45]. This unique structure enables DBNs to excel in unsupervised learning tasks, mainly feature learning and dimensionality reduction.

At their core, DBNs are composed of multiple layers of hidden units, each interacting with the layer above it. This hierarchical arrangement empowers DBNs to capture intricate patterns and relationships within the data, making them particularly adept at tasks where understanding complex, high-level features is crucial.

The unsupervised learning capabilities of DBNs are precious in scenarios where labelled data is limited or unavailable. By leveraging the data's inherent structure, DBNs can autonomously discover meaningful representations, effectively reducing the dimensionality of the input space. This ability has profound implications in computer vision, natural language, and signal processing.

DBNs have demonstrated remarkable effectiveness in diverse applications, including but not limited to image recognition, speech analysis, and recommendation systems. Their adaptability and proficiency in unsupervised learning tasks make them a vital tool in the arsenal of machine learning practitioners [36]. Ongoing research and development in the field continue to refine and enhance the capabilities of DBNs, solidifying their significance in the landscape of deep learning models.

Initially conceived for Natural Language Processing (NLP) tasks, Transformers represent a ground-breaking architecture that leverages self-attention mechanisms to capture intricate global dependencies within data. This innovative approach revolutionised the field by allowing for parallelised processing of sequences, making them highly efficient for tasks requiring an understanding of long-range dependencies. Beyond NLP, Transformers have found extensive application in many functions, spanning language translation, text summarisation, sentiment analysis, and more. This versatility arises from their adaptability to tasks involving structured data where capturing relationships across distant elements is paramount.

One of the most influential derivatives of the Transformer architecture is Bidirectional Encoder Representations from Transformers (BERT), an acronym for Bidirectional Encoder Representations from Transformers. BERT is a pre-trained transformer model specially designed for natural language understanding tasks [35]. What sets BERT apart is its capacity to generate contextualised word representations, meaning it comprehends words based on their context within a sentence. This contextual awareness significantly enhances its ability to understand nuanced linguistic nuances, enabling it to excel in various NLP tasks, including sentiment analysis, named entity recognition, and question-answering.

Transformers and BERT have ushered in a new era in NLP, fundamentally transforming how machines comprehend and generate human language. Their influence extends across various applications, from automated customer support systems to content generation. Ongoing research and refinement in this domain continue to propel the capabilities of Transformers and BERT, pushing the boundaries of natural language understanding and age.

Capsule Networks (CapsNets) constitute a pioneering departure from the conventional Convolutional Neural Networks (CNNs), distinguished by their emphasis on capturing detailed information about the constituent parts of objects and the intricate spatial relationships between them [49]. This unique architectural approach marks a significant leap forward in object recognition, addressing a notable limitation of CNNs. Capsule Networks are particularly adept at understanding how different elements within an object interact and relate, making them exceptionally valuable in scenarios

where discerning fine-grained details and handling viewpoint variations are critical.

The foundational concept behind Capsule Networks is the utilisation of capsules, specialised neural network units that encode information about specific features of an object and their relative positions. Unlike traditional neural networks that may struggle with transformations such as rotations or distortions, Capsule Networks have the potential to preserve the spatial relationships between object components, enabling them to handle viewpoint variations with greater efficacy.

Capsule Networks have shown remarkable promise in object recognition, pose estimation, and image reconstruction tasks. Their capacity to grasp the hierarchical structure of objects and their constituent parts has implications for fields as diverse as computer vision, robotics, and medical imaging [47]. The ongoing refinement and exploration of Capsule Networks continue to expand their potential, propelling them to the forefront of cutting-edge research in deep learning and computer vision.

Networks (CapsNets) constitute a pioneering departure from the conventional Convolutional Neural Networks (CNNs), distinguished by their emphasis on capturing detailed information about the constituent parts of objects and the intricate spatial relationships between them [46]. This unique architectural approach marks a significant leap forward in object recognition, addressing a notable limitation of CNNs. Capsule Networks are particularly adept at understanding how different elements within an object interact and relate, making them exceptionally valuable in scenarios where discerning fine-grained details and handling viewpoint variations are critical.

The foundational concept behind Capsule Networks is the utilisation of capsules, specialised neural network units that encode information about specific features of an object and their relative positions. Unlike traditional neural networks that may struggle with transformations such as rotations or distortions, Capsule Networks have the potential to preserve the spatial relationships between object components, enabling them to handle viewpoint variations with greater efficacy.

Capsule Networks have shown remarkable promise in object recognition, pose estimation, and image reconstruction tasks. Their capacity to grasp the hierarchical structure of objects and their constituent parts has implications for fields as diverse as computer vision, robotics, and medical imaging [39]. The ongoing refinement and exploration of Capsule Networks continue to expand their potential, propelling them to the forefront of cutting-edge research in deep learning and computer vision.

Neuroevolution of Augmenting Topologies (NEAT) is a pivotal advancement in artificial intelligence and neural network development. It represents an evolutionary algorithm meticulously designed to evolve artificial neural networks (ANNs) [50]. What sets NEAT apart is its capacity to dynamically adapt the structure and topology of neural networks throughout the evolutionary process.

NEAT employs a principled approach, introducing new neurons and connections over generations, allowing the network to grow in complexity and adapt to the evolving demands of the task. This unique methodology mitigates the common challenges associated with fixed-topology neural networks, such as finding the optimal architecture for a given problem.

The applications of NEAT are far-reaching, with a prominent focus on tasks involving reinforcement learning. By dynamically adjusting the network architecture and connections, NEAT enables the emergence of neural network structures tailored to the specific demands of complex, dynamic environments. This makes NEAT particularly powerful in scenarios where adaptability, robustness, and performance optimisation are paramount.

NEAT has had a transformative impact on artificial intelligence, with applications spanning robotics, game-playing, and control systems. Its adaptability and versatility have positioned NEAT as a foundational tool for researchers and practitioners seeking to harness the power of evolutionary algorithms in developing artificial neural networks [3]. Ongoing research and refinement in this domain continue to expand the potential of NEAT, driving advancements in neuroevolution and adaptive learning.

This comprehensive review paper explores the prevalent deep learning architectures utilised in animal detection, focusing on methods grounded in Convolutional Neural Networks (CNNs). By examining and analysing the application of CNN-based techniques, this review endeavours to provide a thorough understanding of their effectiveness and potential in advancing the field of animal detection [37]. Through an in-depth exploration of the various CNN-based approaches, this paper aims to shed light on the state-of-the-art methodologies and their contributions to enhancing animal detection systems' accuracy and efficiency.

A. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are deep learning methods well-suited for image-based tasks. They are designed to mimic the human brain's visual processing and are collected from multiple layers of convolutional filters and pooling operations [22]. These layers allow the CNN to learn hierarchical features from the input images, enabling the detection of complex patterns and objects.

Convolutional Neural Networks (CNNs) are a specific feed-forward artificial neural network type. Their structural organisation is influenced by the arrangement of cells in the animal visual cortex. Within the visual cortex, small clusters of cells exhibit sensitivity to specific areas of the visual field. Neuronal cells in the brain respond selectively, firing only in the presence of edge orientations. For instance, some neurons activate in the fact of vertical edges, while others do so for horizontal or diagonal edges. CNNs, employed in deep learning, are designed to assess visual information [49]. They can tackle a wide array of tasks, including processing images, sounds, texts, videos, and various other forms of media.

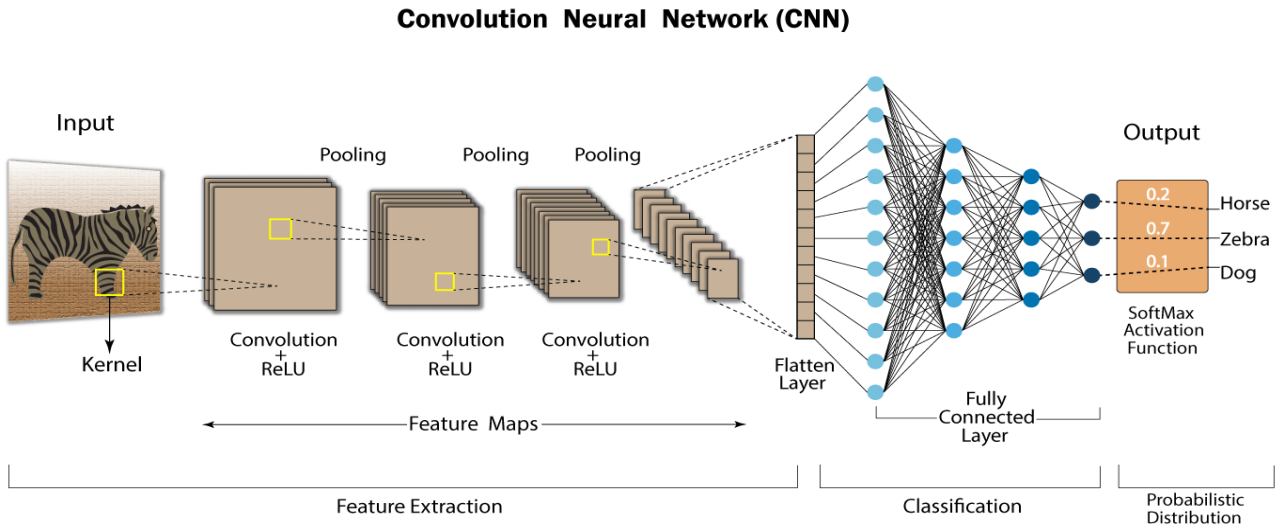


Fig. 3. FA Graphical representation of Convolutional Neural Network (CNN).

Convolutional Neural Networks (CNNs) have an input layer, multiple hidden layers, an output layer, and many parameters, enabling them to discern complex objects and intricate patterns adeptly. These networks employ convolution and pooling processes to down-sample the input data before applying an activation function. These operations are predominantly carried out within partially connected hidden layers, culminating in a fully connected layer that yields the output shown in Fig. 3.

The resultant output from a CNN maintains a spatial dimensionality like the original input image Convolution, in this context, involves the amalgamation of two functions to generate the output of the latter function. In CNNs, the input image undergoes convolution by applying filters, yielding a Feature map. These filters comprise randomly generated vectors encompassing weights and biases within the network [42]. Unlike individualised weights and preferences for each neuron, CNNs employ uniform weights and biases across all neurons. Multiple filters can be instantiated, capturing distinct facets from the input data. Filters are alternatively referred to as kernels.

In animal detection on video data, CNNs play a crucial role in feature extraction and recognition. The initial layers of a CNN identify modest features with similar edges and textures, while deeper layers learn huge abstract and discriminative features, which are relevant for detecting animals. CNNs can be fine-tuned with labelled animal images to adapt the model for animal detection tasks. Convolutional Neural Networks (CNNs) are extensively used, including variants like YOLO (You Only Look Once), SSD (Single Shot MultiBox Detector), R-CNN, Mask R-CNN and RetinaNet, which offer real-time or high-accuracy object detection. Temporal aspects are addressed by I3D (Inflated 3D ConvNet) and T-CNN (Temporal CNN), which capture motion cues. For instance, Mask R-CNN and Tube-CNN are applied for segmentation and tracking. Spatio-temporal interactions are emphasised by STSN (Spatio-Temporal Snippet Network) and STAM (Spatio-Temporal Attention Mechanism). More recent approaches utilise boundary matching (BMN) and trajectory forecasting (TrajNet++) for improved tracking [41].

These architectures leverage deep learning's capacity to learn intricate patterns and temporal dependencies, enabling accurate and efficient animal detection in video data.

1) Region-based CNNs: Region-based Convolutional Neural Networks (CNNs) are deep learning models specifically designed for tasks like object detection. These architectures accurately and efficiently detect objects within images or videos [30]. Within the realm of RCNNs, several algorithms have been developed to improve the performance and efficiency of animal detection. Some of these include:

Faster R-CNN combines region proposal networks (RPN) with a CNN-based object detector. The RPN proposes candidate regions likely to contain animals, and then the CNN refines and classifies these regions. This two-stage approach improves detection accuracy by focusing on promising areas rather than scanning the entire image. R-FCN (Region-based Fully Convolutional Networks) [48] is a single-stage detector that operates directly on the whole picture, using position-sensitive score maps to predict object locations. It achieves accurate animal detection while being computationally efficient.

Fast R-CNN, a pioneering development in object detection, ushered in a paradigm shift by introducing a unified framework that seamlessly integrates region proposal generation and object classification within a single pass. This streamlined approach accelerates detection and enhances accuracy [46]. One of its key innovations is the integration of a Region of Interest (RoI) pooling layer, which efficiently extracts features from various regions of an image. This optimises computational resources and facilitates handling multiple areas of interest.

The impact of Fast R-CNN was profound, transcending previous approaches in both speed and precision. This breakthrough paved the way for a new generation of object detection systems, setting a high bar for subsequent advancements in the field [19]. Fast R-CNN's efficiency and accuracy have made it an instrumental tool in diverse applications, from computer vision tasks like image recognition and scene understanding to practical applications in fields such as autonomous vehicles, surveillance systems, and more. Its influence continues to resonate in the ongoing evolution of object detection methodologies.

Faster R-CNN stands as a remarkable refinement in object detection, building upon the foundation laid by Fast R-CNN. This innovative architecture introduced a pivotal component known as the Region Proposal Network (RPN). Unlike its predecessors, which relied on external algorithms for region proposal generation, the RPN operates directly on the feature maps [41]. This streamlined approach eliminates additional computations, significantly enhancing computational efficiency.

By seamlessly integrating the RPN, Faster R-CNN achieves a remarkable fusion of region proposal generation and object classification within a unified framework. This integration expedites the detection process and leads to substantial gains in accuracy. The direct generation of region proposals from the feature maps represents a significant leap forward, enabling Faster R-CNN to outperform its predecessors in speed and precision.

Faster R-CNN's introduction of the RPN revolutionised the field of object detection, establishing it as a cornerstone in modern computer vision. Its impact extends across a wide array of applications, including but not limited to autonomous driving, object tracking, and facial recognition. The efficiency and accuracy achieved by Faster R-CNN have solidified its status as a foundational framework in the landscape of object detection methodologies [27]. The ongoing research and development in this area continue to build upon the innovations by Faster R-CNN, further advancing the capabilities of object detection systems.

Mask R-CNN represents a monumental advancement in object detection that builds upon the formidable Faster R-CNN framework. What sets Mask R-CNN apart is incorporating a third critical branch dedicated to predicting object masks, bounding boxes, and class probabilities. This breakthrough innovation introduces a level of granularity that was previously unparalleled.

Mask R-CNN achieves a monumental leap forward in object understanding by enabling the precise delineation of objects within an image [17]. This capability, known as instance segmentation, has wide-ranging applications in tasks where detailed object comprehension is paramount. It allows for accurately identifying objects and differentiating between individual instances of the same class.

The addition of the mask prediction branch in Mask R-CNN has revolutionised the field of computer vision and object detection. It has found extensive use in domains such as medical imaging, robotics, and autonomous navigation, where discerning detailed object boundaries are crucial. Mask R-CNN has solidified its position as an indispensable tool for tasks demanding high precision in object localisation and segmentation.

The ground-breaking contributions of Mask R-CNN continue to resonate in computer vision, inspiring further innovations and advancements in object detection and instance segmentation techniques. Its impact extends across a broad spectrum of industries and applications, showcasing its pivotal role in advancing the capabilities of visual perception systems.

Cascade R-CNN, a significant evolution in object detection, introduces a multi-stage approach to enhance object proposal refinement. This innovative architecture deploys a succession of classifiers with progressively higher difficulty thresholds to filter out potential false positives systematically. This cascade strategy fundamentally improves precision in object detection, setting Cascade R-CNN apart as a crucial model for tasks where pinpoint accuracy is paramount.

By employing a cascade of classifiers, Cascade R-CNN effectively refines the object proposal process through stages of increasing stringency. This meticulous filtering mechanism substantially elevates the model's ability to discriminate between true positive and false positive detections. As a result, Cascade R-CNN achieves a level of precision that surpasses previous object detection models.

This refined approach has widespread applications in medical imaging, robotics, and aerial imagery analysis, where high detection accuracy is critical. The cascade architecture in Cascade R-CNN provides a powerful tool for tasks that demand meticulous object recognition, making it an indispensable asset in the arsenal of computer vision practitioners.

The introduction of Cascade R-CNN exemplifies the ongoing pursuit of precision and accuracy in object detection methodologies. Its impact reverberates across a broad spectrum of industries and applications, showcasing its pivotal role in advancing the capabilities of visual perception systems. The continued refinement and exploration of Cascade R-CNN continue to drive progress in object detection.

Indeed, Faster R-CNN and Mask R-CNN are well-suited for animal detection tasks necessitating precise localisation, as they precisely identify the boundaries of objects within an image. Moreover, their adaptability extends to video-based applications, allowing real-time or near-real-time animal detection in dynamic environments.

Region-based Convolutional Neural Networks (CNNs) like Faster R-CNN and Mask R-CNN strike a vital balance between accuracy and computational efficiency. This characteristic makes them highly versatile and applicable in various animal detection scenarios whether in wildlife monitoring, conservation efforts, or ecological research, these models offer a robust solution for accurately identifying and localising animals within imagery or video footage.

The adaptability and efficacy of Faster R-CNN and Mask R-CNN have solidified them as foundational tools in computer vision for animal detection. Their versatile application extends

across various domains, showcasing their pivotal role in advancing the study of understanding and animal detection.

2) Single Shot Multibox Detector (SSD): SSD is a real-world object detection algorithm in the one-stage detectors category. It is widely used for animal detection due to its efficiency and accuracy. SSD works by separating the original image into multiple grids, and every grid cell is answerable for forecasting bounding boxes and class probabilities for potential objects. These predictions are made at various scales, viewing the model to handle varying sizes.

One of the main advantages of SSD is its speed and suitability for real-time animal detection in video streams. It can process video frames rapidly, making it ideal for applications where low latency is essential. Moreover, SSD can efficiently detect multiple animal instances in a single forward pass, making it highly scalable for large-scale animal monitoring scenarios [16].

3) You Only Look Once (YOLO): YOLO is an alternative real-time object detection architecture recognised for its speed and simplicity. Unlike SSD, YOLO approaches object finding as a regression problem. It divides the input image into a grid, and every grid cell directly predicts bounding box coordinates and class probabilities without using separate anchor boxes.

YOLO's characteristics make it a strong candidate for video-based animal detection. Its single-pass architecture enables real-time processing of video frames, making it well-suited for applications that require low latency. However, YOLO might struggle with detecting small animals or closely grouped instances due to the anchor boxes, which could affect its localization [12]. Deep learning architectures have revolutionised animal detection in computer vision. Convolutional Neural Networks provide a solid foundation for feature extraction, while SSD and YOLO offer real-time capabilities, making them ideal for video-based animal detection.

Various authors have applied deep learning to animal detection. In a [1] study, a Convolutional Neural Network (CNN) algorithm was employed to achieve 93.8% accuracy in identifying 48 species in the Snapshot Serengeti dataset, offering automated animal identification with high precision. However, this approach didn't utilise pre-processing techniques or filters. In another [4] study, tracking and detection models like BYTETrack, SORT, IoU-tracker, YOLOv4, DeepSORT, and Few-MOT were used to monitor endangered animals, recording their daily movements and activity areas. While this method embedded uncertainty into multi-object tracking for robust models, it relied on limited frames for experimentation.

In a [5] study, various deep neural networks, including AlexNet, NiN, VGG, GoogLeNet, ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152 were employed, achieving an impressive 96.8% accuracy in identifying animals in camera-trap images. This method successfully identified, counted, and described animals using deep neural networks, although it was limited to a specific dataset (SS dataset). In study [6], the study focused on utilising convolutional neural networks, emphasising day-night joint training and YOLOv5, resulting in a high accuracy of 97.9%. While this approach demonstrated high accuracy, it was considered time-consuming and labour-intensive, mainly due to the difficulty in handling small datasets. Additionally, a [7] study utilised Cascade R-CNN, HRNet32, ResNet50, and ResNet101, achieving a performance of 97%, with the added advantage of efficient imagery processing and saving time. However, it was observed that R-CNN occasionally generated incorrect candidate region proposals. Lastly, in [8], a method combining HOG/SVM and deep neural networks (Faster RCNN and YOLO) attained an accuracy of 87%. This approach excelled in generalising images from the web but faced challenges in detecting small objects.

These studies, among others, highlights the diverse applications of deep learning in animal detection, achieving high accuracy rates and automation benefits but facing challenges in data requirements, computational intensity, and generalisation. The choice of architecture depends on specific application requirements, such as real-time performance, detection accuracy, and computational resources, as research in deep learning continues advancing more effective and efficient architectures for animal detection.

III. The Process of Animal Detection Process from Video Data

Animal detection refers to identifying and recognising animals' presence or location in a given environment. This can be done through various means, including visual observations, sensor technologies, or automated systems utilising computer algorithms.

In technology and computer science, animal detection often involves using machine learning or deep learning techniques to analyse images, videos, or sensor data to identify and classify animals. This technology is used in various applications, including wildlife conservation, agriculture, surveillance, and research. For example, animal detection systems may be deployed in wildlife conservation to monitor the movement and behaviour of endangered species. In agriculture, such systems can track livestock or identify pests. In research, animal detection technology helps gather animal behaviour and ecology data.

Overall, Animal detection is crucial in advancing our understanding and management of animal populations and various industries that interact with or rely on animals. Using low-cost commercial drones, artificial intelligence (AI), neural networks, and computational power has simplified detecting items of interest. Deep learning and convolutional neural networks (CNNs) algorithms are now the benchmark in picture-processing jobs such as object recognition and segmentation [5]. These networks are critical instruments for identifying and analysing animals in video recordings. "Graphical Processing Units (GPUs)" are now widely used in the digital vision field, especially those seeking deep learning and lowering model learning and inference time. Many contemporary conservation methods use machine learning to analyse pictures after data gathering. When images are gathered, machine learning must be implemented.

A. Data Collection from Different Datasets

In study [5], only one data collection was created and used. This included seven hundred photos. The Black had Red, Green, Blue and black/white images. The information was divided into two categories: rhinoceros and automobiles. Each class had 350 picture sizes ranging from 300 *147 to 3840* 2160 pixels. The data was divided into three distinct datasets: one for instruction, another for confirmation, and one for testing. Observing that the testing sample is not used during the model's training is essential. It is exposed in Fig. 4.

Fig. 4. Sample of training data with variations.

In study [6], the video segments used were captured by infrared detectors in the "Northeast Tiger and Leopard National Park between 2014 and 2020". It chose 17 significant types. It extracted pictures from videos using a "Python script" at a frame interval of fifty. It manually consistently annotates the pictures. It is exposed in Fig. 5.

Fig. 5. Sample of the dataset with some species.

In research [7], a dataset with thermal pictures of two mammal groups was created: red deer, European roe deer, and fallow deer and swine, which primarily comprised images of European wild boar. The sample was created using a Pulsar Helion 2 XP50 PRO camera.

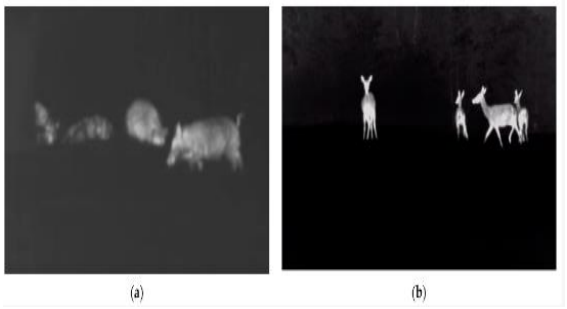


Fig. 6. Example of two distinct objects: (a) "wild boar" and (b) "deer."

One of the most important and widely recognised datasets in animal detection in video data is the "ImageNet Large Scale Visual Recognition Challenge" (ILSVRC). Although primarily focused on object recognition in images, ILSVRC includes a subset of classes related to animals. This dataset was pivotal in advancing deep learning techniques for object detection, serving as a benchmark for various computer vision tasks. In the domain of animal detection specifically, the "ADE20K" dataset is significant. It is a diverse dataset that includes images with annotations for object detection, semantic segmentation, and scene parsing, as shown in Fig. 6. While not exclusively dedicated to animals, it provides a valuable resource for researchers working on animal detection in complex visual environments.

Another prominent dataset is the "COCO (Common Objects in Context)" dataset, widely used for object detection and segmentation tasks. It contains many annotated images depicting a broad range of objects in complex scenes, including various animals in diverse contexts. For more specific applications, datasets like the "AI4MARS - Annotated Image for Machine Learning in Animal Recognition System" focus on animal detection in particular environments, such as the Mars Rover mission. This dataset is curated to detect and classify animals in Martian terrain.

Regarding benchmark datasets, the "PASCAL Visual Object Classes" (PASCAL VOC) dataset is renowned. It covers many object classes, including animals, and has been extensively used for evaluating object detection algorithms. It provides a standardised evaluation platform for researchers in the computer vision community.

The "MS COCO" dataset, in addition to being a significant dataset, also serves as a benchmark for object detection tasks. It includes various object categories, including animals, and is accompanied by a comprehensive evaluation metric suite. These datasets and models play a crucial role in advancing the field of animal detection in video data. They provide standardised and diverse sets of images with ground truth annotations, enabling researchers to train and evaluate their algorithms consistently. This ensures that advancements in animal detection techniques are rigorously tested and compared against state-of-the-art methods, ultimately driving progress in the field.

Some of the datasets from recent works are listed in Table II below:

TABLE II. DETAILS OF DATASET FOR ANIMAL DETECTION VIDEO USING DEEP LEARNING FROM RECENT WORKS

References	Datasets	Description	Scenario
[5, 2018]	ImageNet	It has 1.3 million labelled images for 1,000 groups (from synthetic objects, for example, bicycles and cars, to wildlife categories 1q11 dogs and lions)	Identify and count wild animals in the camera trail camera.
[30, 2021]	Bavarian Highway Directorate, Germany	It displays video segments that last around 10 seconds and has an eight fps (frames per second) resolution of 1280 720 pixels. The footage was captured by camera trail cameras set up at an "animals' bridge" (wildlife crossing) on Federal Highway 7 near "Oberthulba."	To identify animals in wildlife videos
[31, 2018]	Fly and mouse datasets	59 aligned, high-resolution behavioural videos	To estimate the fast animal pose.
[32, 2023]	PolarBearVidID	It includes video sequences of 13 polar bears in various poses and lighting conditions..	Identify animal behaviour in zoos.
[33, 2020]	coco dataset	It has a width and height in the range of 40 and 140 pixels	To detect, classify, and track animals in the African savannah
[34, 2019]	Badger dataset	Images were captured at a selection of UK farms where surveillance occurred. All were manually assigned to badger, bird, cat, fox, rat, or rabbit.	To monitor wildlife

B. Pre-processing

Data pre-processing is critical to get the best results from any artificial intelligence-based approach. Various data pre-processing methods can be deployed to make the information fit for the development of models. If the data taken from the devices was noisy and meaningless, it was deleted from the collection. The research in [8] used a "6th-order Butterworth filter with a cutoff frequency of 3.667 Hz" to eliminate the noise and anomalies from the information. In [9], YOLOv5 eliminates duplicate and similar images. In [10], by employing computer vision tools, a picture processing system was created and built to enhance contrast in pictures and segment pertinent image subdivisions. The photos were revised to expedite the procedure. It is shown in Fig. 6.

C. Image Segmentation

Image segmentation involves partitioning each video frame into distinct regions corresponding to specific objects or animals. This enables the isolation and identification of animals within a dynamic visual stream. Techniques like thresholding, edge detection, and deep learning-based segmentation are employed to delineate animals from the background, allowing for subsequent analysis such as tracking, behaviour monitoring, and species classification throughout the video sequence. Effective segmentation is crucial for accurate and reliable animal detection and tracking in dynamic video data. A computer vision method locates items within an image by forming a bounding box surrounding it to comprehend what is in it at the pixel stage. In study [7], the bear was segmented from the picture using the "MSER (Maximally Stable Extremal Region)" method. In study [11], it uses Mask R-CNN for segmentation. It can distinguish between different items in a video. It returns the class identity, item masks, and the bounding box coordinates for every object in a provided picture. The process is split into two phases; the first scans the image and generates a "Region proposal network (RPN)" to provide potential object bounding boxes, and the second differentiates proposals and generates bounding boxes and boundaries for every class. The study in [12] used the "Falzenszwalb algorithm" for image segmentation, which organises pixels with comparable luminance.

D. Classification

In animal detection, classification pertains to categorising identified objects into specific animal species or classes based on distinctive features learned through the model. It distinguishes different animals within a scene, aiding in species identification and population monitoring. [30] In animal detection, classification is crucial for understanding wildlife dynamics, ecological studies, and conservation efforts shown in the Fig. 7. It creates a list of desired objects for Detection images and instructs a model to identify them using labelled sample photos. In [5], Faster-RCNN is used for image classification in video clips. In [6], "Dilated Residual Networks (DRN)" was used, and better accuracy was achieved. In [12], the author used deep boosting, dictionary learning, and convolution networks for image classification. In [9], three models ("YOLOv5m, CNN_HRNet32, FCOS_Resnet101") were used for simulation to detect and classify multiple animals in a video, and the video classification accuracy of FCOS_Resnet101 achieved 91.6%. It is shown in Fig. 5.

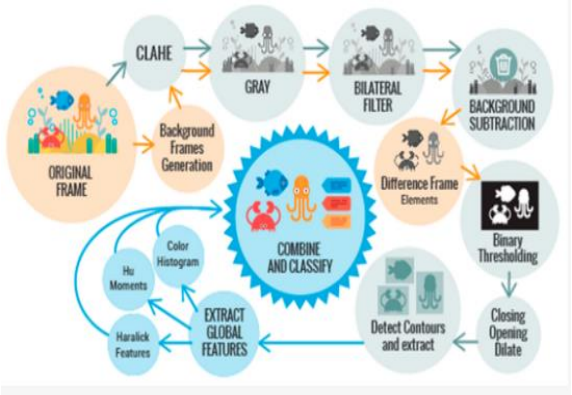


Fig. 7. Pipeline for data pre-processing.

Fig. 8. Example of correct classification.

E. Evaluation Metrics for Animal Detection for Video Data

The evaluation metrics for animal detection in video data include precision, recall, F1 Score, mean average precision (mAP), intersection over union (IoU), false positive rate (FPR), false negative rate (FNR), accuracy, mean recall, processing speed, and frame-level vs. video-level evaluation [13]. These metrics assess the system's ability to accurately detect animals, handle false positives and negatives, and its overall performance in real-time video processing, providing a comprehensive evaluation of the detection system's effectiveness and efficiency. Fig. 8 shows example of correct classification.

1) Video data evaluation metrics: Evaluating video-based animal detection systems is vital to measure their accuracy and effectiveness in real-world scenarios. As deep learning models are commonly used for this task, specific evaluation metrics are required to assess their performance on video data [14]. One of the key metrics used in this context is the mean average precision at different Intersections over Union (mAP@IoU) thresholds over time.

mAP@IoU is an extension of the traditional mean average precision (mAP) metric, commonly used in image object detection tasks. It measures the precision and recall of predicted bounding boxes by comparing their overlaps with ground-truth annotations at different IoU thresholds. This provides a comprehensive evaluation of the model's ability to accurately detect animals in video sequences across various IoU thresholds, reflecting different levels of object localisation accuracy [16]. Evaluating mAP@IoU over time enables the assessment of the model's consistency and robustness in detecting animals throughout the video, accounting for spatial and temporal variations.

2) Temporal consistency metrics: Temporal consistency is a critical aspect of video-based animal detection as it ensures stable and reliable tracking of animals across frames. Several methods can be employed to evaluate the stability and consistency of animal detections over time:

Temporal Intersection over Union (tIoU): tIoU measures the temporal overlap between predicted bounding boxes and ground-truth annotations in consecutive frames [17]. It quantifies how well the model can maintain accurate and continuous detections of animals across time. A higher tIoU score indicates better temporal consistency and reliable tracking.

ID Switching Rate: The ID switching rate evaluates how often the model incorrectly assigns different identities to the same animal or switches identities between consecutive frames [28]. A lower ID switching rate signifies improved temporal consistency in tracking and maintaining individual animal identities.

Fragmentation Rate: Fragmentation occurs when the model fails to link consecutive detections of the same animal, resulting in disjointed tracks [15]. The fragmentation rate measures the degree of this issue, with lower values indicating better temporal consistency and continuous tracking.

Trajectory Smoothness: This metric assesses the smoothness and continuity of animal trajectories over time. [18] A model with high trajectory smoothness exhibits more consistent and visually coherent animal tracks, indicating reliable temporal consistency.

By integrating video-based evaluation metrics like mAP@IoU over time with temporal consistency metrics, researchers and practitioners can thoroughly assess the performance of deep learning models for video-based animal detection [22]. These evaluation measures help identify potential areas for improvement, validate the suitability of the models for specific applications like wildlife monitoring, behaviour analysis, and ecological research, and drive advancements in computer vision for animal-related studies.

IV. Recent Progress in Animal Detection using Deep Learning

Recent Progress in animal detection using different machine learning and deep learning techniques is presented here. Table III, in a tabular format, compares the various DL and ML learning models for animal detection. A comparison is made regarding the detection methods, main findings, advantages, and disadvantages for further research from existing literary works.

These references represent a range of scholarly works that have explored various deep-learning techniques applied explicitly to video-based animal detection. [19] By including these references, the review aims to provide a comprehensive and up-to-date analysis of the existing literature in this domain. [33] Additionally, these selected references contribute significant insights, methodologies, and findings relevant to understanding the advancements, challenges, and potential applications of deep learning in video-based animal detection. [27] Their inclusion strengthens the credibility and rigour of the review, offering a solid foundation for examining the state-of-the-art approaches and identifying future research directions in this field.

TABLE III. RECENT PROGRESS IN ANIMAL DETECTION USING DIFFERENT MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

References	Detection Methods	Main Finding	Advantages	Disadvantages
[1, 2020]	Convolutional Neural Network (CNN) algorithm to detect wild animals.	Deep convolutional neural networks to identify, count, and describe the behaviours of 48 species in the 3.2-million-image Snapshot Serengeti dataset. 93.8% accuracy	Automating animal identification with 99.3% accuracy, matching human volunteers' 96.6% saves over 17,000 hours of labelling effort on a 3.2-million-image dataset.	No Pre-processing techniques or filters are used in this paper.
[4, 2022]	BYTETrack, SORT, IoU-tracker, V-IoU-tracker, YOLOv4, + DeepSORT and Few-MOT model	Few-MOT for wildlife to embed uncertainty into designing a multiobject-tracking model by combining the richness of deep neural networks with few-shot learning, leading to correctable and robust models.	Few-MOT employs few-shot learning and tracking-by-detection to monitor endangered animals, recording daily movements and frequent activity areas for analysis.	A limited number of frames are used to experiment.
[5, 2018]	AlexNet, NiN, **VGG,** GoogLeNe, ResNet-18, ResNet-34, ResNet-50, ResNet-101 , ResNet-152	VGG model achieved the best accuracy of 96.8%	Deep neural networks (DNNs) can successfully identify, count, and describe animals in camera-trap images.	Only an SS data set is used.
[6, 2021]	The detection of animals Using convolutional neural networks.	Day-night joint training had a better performance.YOLOv5 achieved an accuracy of 97.9%	Time-consuming and labour intensive.	Difficulty with small datasets
[7, 2022]	Cascade R-CNN, HRNet32, ResNet50 and ResNet101.	It achieved a performance of 97%.	The imagery potentially quickly and efficiently saves much time.	R-CNN sometimes generates wrong candidate region proposals as the selective search is a fixed algorithm with no learning capabilities.
[8, 2023]	(HOG/SVM), deep neural networks (Faster RCNN and YOLO).	It attained 87% accuracy	The web can generalise the image better.	Struggles to detect small objects.
[9, 2022]	Convolutional neural networks (CNNs) 1D CNN and 2D CNNs	It attained 99.70% training accuracy and 96.85% validation accuracy.	Real-time monitoring of activities	High computational requirements
[10, 2022]	Deep learning technology	Accuracy of 95.1%	It can benefit from computer automatic identification of postural behaviour, which can be used to quantify animal activity.	This can be costly and time-consuming.
[11, 2020]	Machine learning and Deep learning.	VGGNET was the best algorithm for an accuracy of 96.6 %.	Monitoring networks capable of providing large amounts	Movies cannot all be manually processed
[12, 2022]	YOLO; Convolutional neural networks (CNN); SSD; mask R-CNN; VGG-Net	VGGNET was the best algorithm for animal classification, with an accuracy of 96.6 %	It has a standard and easy-to-understand architecture for CNNs, with multiple convolution and pooling layers.	Slow training time
[13, 2023]	Deep learning-based model for automated recognition, definition, and numbering of wild creatures in camera trail camera images	A deep learning model automates the recognition, classification, and counting of wildlife in trail camera images from the "Snapshot Serengeti dataset.	This deep learning model automates animal recognition and description in trail camera images, streamlining manual work and scaling for large datasets.	Data quality impacts accuracy: biased or poorly labelled data may yield suboptimal results. Limited generalisation, computational demands, and interpretability are challenges.
[14, 2019]	Evaluation of ML and DL techniques, including SVM, RF, AlexNet, and Inception v3, for classifying animal genera from the "KTH dataset.	ML and DL techniques for Detection of animal genera using camera capture images. They investigated "SVM, RF, deep learning methods, and AlexNet and Inception v3 machine learning techniques". They used the "KTH dataset," which includes nineteen distinct animal categories.	The paper compares ML and DL techniques for animal genera recognition, emphasising DL's automatic feature learning, scalability, and generalisation.	ML and DL techniques depend on diverse data for better performance. Hyperparameter tuning is time-consuming. DL requires more data, which is challenging for specific animal genera. Interpretability is difficult, especially in complex, deep architectures.
[15, 2019]	Random forest, K-nearest neighbours (KNN), support vector machine (SVM), naive Bayes, and artificial neural network (ANN)	The study showcased a computerised pet movement and mood detection system, exploring data sources and machine learning techniques like random forest, KNN, SVM, naive Bayes, and ANN.	Random Forest for complex relationships, KNN for non-linear data, SVM for high dimensions, Naive Bayes for text/categorical data, and ANN for versatile deep learning applications.	ML algorithms have unique strengths: Random Forest for complex relationships, KNN for non-linearity, SVM for high dimensions, Naive Bayes for text/categorical data, and ANN for versatile deep learning.
[16, 2018]	The authors employ two popular object detection methods - Single Shot Multi-Box Detector (SSD) and You Only Look Once (YOLO).	Set forward a paradigm for automatically identifying animals via camera-trail camera ped pictures. It focused on determining the species and the number of species recorded in	SSD and YOLO enable real-time object detection, detecting multiple objects in one pass. Their end-to-end approach is versatile for animal species without retraining.	Training deep learning-based object detection models like SSD and YOLO demands a large, labelled dataset that is resource-intensive to collect.

		the photograph through algorithms for object detection like "Single Shot Multi-Box Detector (SSD) and You Only Look Once (YOLO)."		Performance varies with image complexity, requiring hyperparameter tuning. False positives/negatives affect accuracy.
[17, 2020]	We identify bird species using a "Super-resolution Mask RCNN-based transfer deep learning" model. The method combines super-resolution and Mask RCNN techniques to identify bird species accurately.	"Super-resolution Mask RCNN-based transfer deep learning approach" to identify bird species. They used Mask RCNN to identify very minute differences that analysed pixel-by-pixel of the picture and added a mask to it, making it easy to identify the dimensions and form of the item.	Combined with super-resolution, Mask R-CNN improves bird species identification, providing precise localisation and reducing overhead through transfer learning.	Mask RCNN and super-resolution demand significant computational resources. Transfer learning reduces data needs, but diverse bird datasets are vital for fine-tuning. Mask RCNN lacks interpretability.
[18, 2018]	Recent deep-learning methods are used to surmount the expense and effort of processing camera trail camera pictures.	The authors employ two distinct datasets to monitor populations of animals and govern environments all over the globe.	Deep learning automates camera trap image analysis, aiding animal identification, counting, and achieving state-of-the-art accuracy for global monitoring.	Deep learning requires ample labelled data, which is challenging for rare species. High-performance hardware is vital due to computational intensity. Model interpretability and bias are complex.
[19, 2020]	The proposed technique for animal identification employs neural network design such as "SSD and faster R-CNN."	Using object identification, an exclusive animal recognition and collision avoidance system aims to detect animals and prevent road collisions.	SSD and Faster R-CNN offer real-time object detection, including animals, for collision avoidance systems, enhancing road safety for humans and wildlife.	Deep learning object detection models like SSD and Faster R-CNN rely on expensive labelled datasets susceptible to environmental variations. Generalisation hinges on diverse data, with the potential for false positives/negatives and performance influenced by model settings and training data.
[20, 2021]	The suggested technique is based on the "Sobel edge algorithm," which is basic but effective in detecting edges based on modified values.	The method achieves rapid detection (0.033 s per image) through thermal pixel analysis, synchronising thermal and RGB attributes for superior real-time animal detection across diverse settings. "Size-temperature filters" enhance applicability.	The real-time system aids wildlife monitoring, safety, and research sans extensive training data, with contextual information reducing false positives bolstered by thermal and RGB fusion.	The Sobel edge algorithm's limitation lies in detecting only edges, lacking fine-grained details and colour information for precise species identification. Environmental conditions impact detection accuracy, and species-specific recognition is limited.
[21, 2020]	It evaluates performance forecasting incrementally ahead with long-range scenarios using "Random Forests, Neural, and Recurrent Neural Networks." This method is used to analyse excellent quality and movement statistics. For one step forward prediction, it was discovered that individual-level Machine Learning and Deep Learning approaches beat the SDE model.	The broad framework for forecasting animal movement consists of two steps: initially estimating behavioural motion stages and then predicting the animal's velocity. This framework is specified for both individual and group training.	The framework assesses individual and group training for complete animal movement patterns, utilising Random Forests, Neural Networks, and Recurrent Neural Networks for varied prediction approaches. High-quality data enhances precision in predictions.	Effectiveness relies on adequate data, which could be more challenging. Multiple models add complexity, demanding computational resources. Performance varies with species and environment, affecting generalisation.
[22, 2021]	The DL algorithm detects and segregates the animals with a large-scale publicly available dataset. A CNN model predicts the object (animal) in every image frame obtained from the live camera.	They have implemented deep learning for detecting the animals from the videos to improve safety. If the algorithm detects the object as an animal, the system will generate an alarm for 3 seconds to avoid a collision. Results show that the proposed approach achieves an accuracy of 91%.	Achieving 91% real-time Animal detection accuracy, the approach uses deep learning and CNNs with large-scale dataset training for enhanced safety via collision alerts across diverse species and conditions.	The DL algorithm's accuracy and generalisation depend on the training dataset's quality and diversity. Environmental conditions impact accuracy. Detecting rare species poses challenges. False positives/negatives may occur.
[23, 2017]	A deep CNN model is employed in this work for animal detection, and the model is trained using a single labelled dataset. The DCNN algorithm can automatically filter animal images and identify animal	The proposed approach achieves a phenomenal accuracy of 96.6% for animal detection and 90.4% for identifying the species accurately.	The method achieves an impressive 96.6% accuracy in animal detection and 90.4% in species identification, which is vital for reliable wildlife monitoring.	Single-dataset reliance hampers generalisation to new environments and species, influenced by data bias. Diverse dataset training enhances performance but demands substantial computational resources, which is challenging

	species.			on resource-constrained platforms.
[24, 2017]	CNN model for animal recognition is implemented. The effectiveness of different image recognition techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Local Binary Patterns Histograms (LBPH), and Support Vector Machine (SVM) are analysed, and the performance is compared with the proposed CNN model in terms of recognition rate.	The models are experimentally evaluated for recognition time using a 500-set, 100-image animal database. PCA outperforms LDA and LBPH for large databases, while LBPH excels with small datasets. The proposed CNN model achieves 98% recognition accuracy compared to other models.	The CNN model achieved 98% accuracy in animal recognition, excelling in image identification with automatic feature learning, highlighting its superiority over traditional methods in animal recognition tasks.	CNN model performance depends on dataset quality and size, influenced by biases. It's resource-intensive, with hyperparameter tuning challenges, and offers limited interpretability as a black-box model.
[25, 2021]	The CNN model groups animal images based on the input database. The performance of CNN is compared with conventional recognition techniques such as SU, DS, MDF, LEGS, DRFI, MR, and GC. These techniques are usually characterised by high false positive and negative rate detection.	Developing an efficient animal recognition system is vital. Genetic algorithm-based image segmentation and neural network classification improve accuracy. A dataset with 100 subjects and two classes demonstrates higher precision (99.02%), recall (98.79%), F1 score (98.9%), and low MAE (0.78%).	A CNN model with genetic algorithm segmentation achieves high animal image accuracy, surpassing conventional methods.	The CNN model and genetic algorithm depend on training data quality. Computational complexity requires substantial resources. Hyperparameter tuning impacts the genetic algorithm's performance. Limited information hinders assessing potential limitations.
[26, 2019]	Convolution neural network and SVM	Precision 99.02%, recall 98.79%, F-Measurement 98.9%, and MAE (0.78%)	High accuracy rates	Limited ability to generalise
[27, 2018]	Deep learning method	Average gain between 6% and 10% when compared to the method Fast R-CNN	Deep learning algorithms have been shown to achieve state-of-the-art performance on various problems, including image and speech recognition, natural language processing, and computer vision.	This can make it difficult to understand how the model makes predictions and identify any errors or biases.
[28, 2018]	Deep convolutional neural network	It saves 99.3% of the manual labour.	Deep learning models can generalise well to new situations or contexts, as they can learn abstract and hierarchical representations of the data.	Deep learning models can only make predictions based on the data it has been trained on
[29, 2017]	Deep Convolutional Neural Networks for Automated Wildlife Monitoring	It achieved an accuracy of 96.6% for detecting images.	Deep learning models can be easily scaled to handle increasing data and can be deployed on cloud platforms and edge devices.	This can be costly and time-consuming.
[30, 2021]	R-CNN	Accuracy of 90.4 % for detecting most common species.	Autonomous vehicles use it to perceive objects in their surroundings to ensure a safe driving experience.	This sometimes could result in the generation of lousy regional proposals.
[31, 2018]	deep-learning-based method	The error rate is less than 3%	It can see what human minds cannot visualise.	One cannot accurately define the sorting and output of an unsupervised task.
[32, 2023]	re-ID models	Accuracy of 96.6% for *PolarBearVidID* dataset	Developing re-ID models will significantly facilitate the work of biologists and animal caretakers in the future.	Limited dataset and models used
[33, 2020]	ssd_inception_v2 , ssd_mobilenet_v2 , ssd_mobilenet_v2_quantized , ssdlite_mobilenet_v2 and Raspberry Pi	ssd_mobilenet_v2 average precision is high and generates the least number of false positives	MobileNet V2 was selected as the final model for the application. This is due to its traits of generating a small number of false positives and not splitting an event into smaller ones.	MobileNet V2 sacrifices accuracy for speed, potentially leading to missed detections or misclassifications, especially in complex scenes.
[34, 2019]	CNN1 and CNN2	They achieved an accuracy of 95.86% and 98.05% for binary classification. For multiclassification, they achieved accuracies of 83.07% and 90.32%.	The trained CNNs were directly applied to video footage because film can be considered a sequence of image frames. To speed up the detection process, all images were converted to grayscale.	Limited models are used to compare the results.

References for the research on animal detection and recognition using deep learning are based on the following criteria: Relevance to the research topic: The selected papers are directly related to animal detection in videos using deep learning algorithms, which align with the focus of the proposed research. The references cover a range of publication years, indicating the proposed interest in exploring recent advancements and foundational works in the field.

Animal detection models proposed to date have positive and negative aspects [18]. Most models only produce output and predict the results, neglecting to deal with the problem of output unpredictability in terms of accuracy, sensitivity, and specificity [34]. The papers in [1] and [9] have achieved high animal detection accuracy.

- The study in [1] uses a Convolutional Neural Network (CNN) algorithm to detect wild animals, achieving an accuracy of 93.8%. It also mentions automating animal identification with 99.3% accuracy.
- The study in [9] employs Convolutional Neural Networks (CNNs) and achieves an impressive 99.70% training accuracy and 96.85% validation accuracy for real-time monitoring of activities.

Considering the high accuracy and real-time monitoring capabilities, [9] is a promising paper for animal detection using video data.

Data pre-processing is crucial for optimal results. Techniques include noise removal and image enhancement through methods like Butterworth filters, contrast enhancement, and segmentation. Segmentation involves precisely delineating individual animals or their parts within images or video frames. Methods like MSER, Mask R-CNN, and Falzenszwalb algorithm are employed for this purpose. [24] Various models such as Faster-RCNN, Dilated Residual Networks, YOLOv5, and others are used for image classification [26]. Metrics like precision, recall, F1 Score, mean average precision (mAP), intersection over union (IoU), false positive rate (FPR), false negative rate (FNR), and others are used to evaluate the performance of animal detection systems in video data.

Various detection methods like Convolutional Neural Networks (CNN) and others have been employed with different findings and advantages. CNNs have been used for accurate and automated animal identification. The table comparing different models showcases their respective detection methods, primary results, benefits, and disadvantages. Each entry discusses the detection approach, performance metrics, and the strengths and weaknesses of the technique. Some models like Faster R-CNN, Mask R-CNN, Fast R-CNN, YOLOv4 and DeepSORT, use deep learning and tracking for robust animal detection.

These studies demonstrate advanced deep-learning techniques in animal detection, with various models achieving high accuracy rates. However, challenges such as data quality, computational requirements, and model interpretability remain essential considerations. The choice of model depends on the application's specific needs, such as real-time detection, tracking, or species identification. The use of CNNs and its accuracy make it a strong candidate for further research.

V. Limitations and Challenges

A primary challenge was implementing animal identification models in the existing wildlife ecosystem. Although some researchers tested their models on forests, it doesn't remain easy. A large dataset incorporating many animals would be helpful in training ML and DL models for application in real time. Additionally, due to the computation cost of a vision-based system, it is unlikely to run real-time identification. Advancements can be made by providing a warning in the manner of a message to the neighbouring forest office when the animal is discovered. It can also be used to decrease human-wildlife conflict and animal accidents. Capturing distinctions between creatures from the same family but of different species might be difficult and should be addressed shortly.

Although the approaches utilising deep learning performed well, there are still certain limitations when dealing with submerged multimedia information, such as poor resolution, lighting changes, and intricate backgrounds. More effective approaches for dealing with these issues should be developed based on these limits. Handling more detailed pictures is essential for better understanding animals and their specific activities. Camera and drone technology developments will enable wildlife surveillance at much greater flight heights, reducing interruption to animals in their natural habitats.

While deep learning approaches have demonstrated efficacy, certain limitations persist, especially when handling submerged multimedia data. Factors such as low resolution, fluctuating lighting conditions, and intricate backgrounds pose obstacles that necessitate the development of more robust techniques. Addressing these constraints would yield improved performance in dealing with challenging scenarios. Enhancing the processing of detailed images is pivotal for a more nuanced understanding of animal behaviour and activities. Anticipated advancements in camera and drone technologies hold the potential to facilitate wildlife surveillance at higher altitudes, minimising disturbances to animals in their natural habitats.

VI. Future Challenges and Research Directions

While significant strides have been made in applying deep learning for animal detection on video data, several challenges and avenues for future research emerge. One pressing concern is the need to address the intricacies of adapting models to diverse and dynamic environmental conditions. This includes developing techniques to handle lighting, weather, and vegetation variations, which are paramount for real-world deployment in natural habitats.

Furthermore, the issue of data scarcity remains a formidable obstacle. Future research should explore innovative data augmentation, synthesis, and generation approaches. By creating diverse and representative training sets, models can be more effectively trained to recognise a broader spectrum of animal species and behaviours.

In addition to the challenges, exploring novel data augmentation and synthesis techniques is imperative to

mitigate the need for labelled training data. Developing interpretable and explainable deep learning models for animal detection is another critical avenue, enabling researchers to gain insights into model decisions and bolstering trust in automated detection systems. Further, investigating methods for anomaly detection and outlier identification within video data streams can significantly enhance the robustness and reliability of animal detection systems. Additionally, researching federated learning and edge computing approaches holds promise for decentralised deployment in remote and resource-constrained environments. Addressing these multifaceted challenges will advance the animal detection field and foster sustainable coexistence between humans and the animal kingdom.

These include delving into temporal analysis and behaviour modelling for an in-depth understanding of long-term animal behaviour patterns. Exploring multi-species interactions and fine-grained species identification can provide valuable insights into complex ecological relationships. Enhancing real-time adaptability and edge computing capabilities is vital for dynamic environments while addressing ethical considerations and human-wildlife interactions is imperative to ensure responsible technology deployment. The integration of multi-modal data, the development of user-friendly interfaces, and the fostering of interdisciplinary collaboration are crucial for advancing the field. Moreover, research efforts should also focus on privacy-preserving techniques and long-term monitoring for a holistic approach to animal detection. By tackling these multifaceted challenges, researchers can contribute to a more comprehensive understanding of wildlife behaviour and habitat dynamics, ultimately aiding in practical conservation efforts and the coexistence of humans and the natural world.

Considering the increasing importance of interdisciplinary collaborations, future research efforts should seek to bridge the gap between computer vision experts, ecologists, and conservation biologists. A holistic understanding of animal detection technology's impact on ecological research and wildlife conservation can be achieved by pooling expertise from diverse fields.

VII. Conclusion

The traditional method of animal detection relies on direct observation by humans or the utilisation of specialised tools and techniques. It involves visually scanning an area for animal presence, examining tracks and traces left behind, setting up camera trail cameras for remote monitoring, using acoustic devices to capture animal sounds, and employing trail camera ping and capture methods for direct examination. These traditional approaches have provided valuable insights into animal behaviour and ecology, but they can be labour-intensive, time-consuming, and may have limitations in terms of coverage and scalability. Modern technologies are increasingly integrated with traditional methods to enhance animal detection and monitoring capabilities.

Deep learning approaches provide a beneficial tool for continuous tracking and abundance predictions in animal detection, outperforming laborious, individual efforts in only a tiny percentage of the time. The DL algorithms efficiently identify animals with a high degree of accuracy, and the image of the identified animal is shown for improved accuracy so that it may be utilised for other reasons, such as Detecting wild animals in human habitats and preventing wildlife poaching and human-animal conflict.

Researchers would benefit significantly from detecting animals and extracting their characteristics for their research and detailed study of animal species. As a result, the emerging technology of various Machine Learning and Deep learning algorithms may be applied to animal identification and detection. CNN's technique for Detecting large animals in visuals has proven effective.

References

[1] Banupriya, N., Saranya, S., Swaminathan, R., Harikumar, S., &Palanisamy, S. (2020). Animal detection using a deep learning algorithm. J. Crit. Rev, 7(1), 434-439.

[2] Christin S., Hervet É., Lecomte N. Applications for deep learning in ecology. Methods Ecol. Evol. 2019;10:1632–1644. doi: 10.1111/2041-210X.13256.

[3] Zhao Z.-Q., Zheng P., Xu S.-t., Wu X. Object detection with deep learning: A review. IEEE Trans. Neural Netw. Learn. Syst. 2019;30:3212–3232. doi: 10.1109/TNNLS.2018.2876865.

[4] Feng J., Xiao X. Multiobject Tracking of Wildlife in Videos Using Few-shot Learning. Animals. 2022;12:1223. doi: 10.3390/ani12091223.

[5] Mohammad Sadegh Norouzzadeha, Anh Nguyen, Margaret Kosmalac, Alexandra Swanson, Meredith S. Palmer, Craig Packer, and Jeff Clunea, “Automatically identifying, counting, and describing wild animals in camera-trail camera images with deep learning,” PNAS, 2018.

[6] C. Chalmers, p. Fergus, c. Aday curbelo montanez, steven n. Longmore, and serge a. Wich on video analysis for detecting animals using convolutional neural networks and consumer-grade drones. (2021).

[7] Mengyu tan, wentao chao, jo-ku cheng on animal detection and classification from camera trap images using different mainstream object detection architectures. (2022).

[8] Łukasz popek, rafał perz, grzegorz galiński on comparison of different animal detection and recognitionon thermal camera images. (2023).

[9] Hussain, a., ali, s., & kim, h. C. (2022). Activity detection for the wellbeing of dogs using wearable sensors based on deep learning. Ieee access, 10, 53153-53163.

[10] Lei, Y., Dong, P., Guan, Y., Xiang, Y., Xie, M., Mu, J., ...& Ni, Q. (2022). Postural behaviour recognition of captive nocturnal animals based on deep learning: a case study of Bengal, slow loris. Scientific Reports, 12(1), 7738.

[11] Lopez-Vazquez, V., Lopez-Guede, J. M., Marini, S., Fanelli, E., Johnsen, E., &Aguzzi, J. (2020). Video image enhancement and machine learning pipeline for underwater animal detection and classification at cabled observatories. Sensors, 20(3), 726.

[12] Tanishka Badhe, Janhavi Borde, Vaishnavi Thakur on Study of Deep Learning Algorithms to Identify and Detect Endangered Species of Animals. (2022)

[13] Battu, T., & Lakshmi, D. S. R. (2023). Animal image identification and classification using deep neural network techniques. Measurement: Sensors, 25, 100611.

[14] Rajasekaran Thangarasu, Vishnu Kumar Kaliappan, Raguvaran Surendran, Kandasamy Sellamuthu, Jayasheelan Palanisamy, “Recognition Of Animal Species On Camera Trap Images Using Machine Learning And Deep Learning Models,” International Journal Of Scientific & Technology Research, 2019.

[15] S. Aich, S. Chakraborty, J.-S. Sim, D.-J. Jang, and H.-C. Kim, ‘‘The design of an automated system for analysing the activity and emotional patterns of dogs with wearable sensors using machine learning,’’ Appl. Sci., vol. 9, no. 22, p. 4938, Nov. 2019.

[16] Alexander Loos, Christian Weigel, Mona Koehler, “Towards Automatic Detection of Animals in Camera-Trap Images,” European Signal Processing Conference (EUSIPCO), 2018.

[17] Sofia K. Pillai, Dr M. M. Raghuwanshi, Dr P. Borkar, "SuperResolution Mask-R CNN Based Transfer Deep Learning Approach For Identification Of Birds Species," International Journal of Advanced Research in Engineering and Technology (IJARET), 2020.

[18] Stefan Schneider, Graham W. Taylor, Stefan C. Kremer, "Deep Learning Object Detection Methods for Ecological Camera Trap Data," Arxiv, 2018.

[19] Atri Saxena, Deepak Kumar Gupta, Samayveer Singh, "An Animal Detection and Collision Avoidance System Using Deep Learning," SpringerLink, 2020.

[20] Lee, S., Song, Y., &Kil, S. H. (2021). Feasibility analyses of real-time detection of wildlife using UAV-derived thermal and RGB images. Remote Sensing, 13(11), 2169.

[21] Wijeyakulasuriya, D. A., Eisenhauer, E. W., Shaby, B. A., & Hanks, E. M. (2020). Machine learning for modelling animal movement. PloS one, 15(7), e0235750.

[22] Santhanam, S., Panigrahi, S. S., Kashyap, S. K., & Duriseti, B. K. (2021, November). Animal Detection for Road Safety Using Deep Learning. In 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA) (pp. 1-5). IEEE.

[23] Nguyen, H., Maclagan, S. J., Nguyen, T. D., Nguyen, T., Flemons, P., Andrews, K., ... & Phung, D. (2017, October). Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring. In 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 40-49). IEEE.

[24] Trnovszky, T., Kamencay, P., Orjesek, R., Benco, M., & Sykora, P. (2017). Animal recognition system based on convolutional neural network. Advances in Electrical and Electronic Engineering, 15(3), 517-525.

[25] Chandrakar, R., Raja, R., & Miri, R. (2021). Animal detection is based on deep convolutional neural networks with genetic segmentation. Multimedia Tools and Applications, 1-14.

[26] Manohar, N., Sharath Kumar, Y. H., Kumar, G. H., & Rani, R. (2019). Deep learning approach for classification of animal videos. In Data Analytics and Learning: Proceedings of DAL 2018 (pp. 421-431). Springer Singapore.

[27] Mauro dos Santos de Arruda, Gabriel Spadon, Wesley Nunes Goncalves, & Bruno Brandoli Machado, "Recognition of Endangered Pantanal Animal Species using Deep Learning Methods," IJCNN, 2018.

[28] Mohammad Sadegh Norouzzadeha, Anh Nguyen, Margaret Kosmalac, Alexandra Swanson, Meredith S. Palmer, Craig Packer, and Jeff Clunea, "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," PNAS, 2018.

[29] Hung Nguyen, Sarah J. Maclagan, Tu Dinh Nguyen, Thin Nguyen, Paul Flemons, Kylie Andrews, Euan G. Ritchie, and Dinh Phung, "Animal Recognition and Identification with Deep Convolutional Neural Networks for Automated Wildlife Monitoring," Deakin University, Geelong, Australia, 2017.

[30] F. Schindler and V. Steinhage, "Identification of animals and recognition of their actions in wildlife videos using deep learning techniques," Ecological Informatics, vol. 61, p. 101215, Mar. 2021, doi: https://doi.org/10.1016/j.ecoinf.2021.101215.

[31] T. D. Pereira et al., "Fast animal pose estimation using deep neural networks," Nature Methods, vol. 16, no. 1, pp. 117–125, Dec. 2018, doi: https://doi.org/10.1038/s41592-018-0234-5.

[32] M. Zuerl et al., "PolarBearVidID: A Video-Based Re-Identification Benchmark Dataset for Polar Bears," Animals, vol. 13, no. 5, p. 801, Jan. 2023, doi: https://doi.org/10.3390/ani13050801.

[33] Amanda Tydén and Sara Olsson, "Edge Machine Learning for Animal Detection, Classification, and Tracking", 2020.

[34] R. Chen, R. Little, L. Mihaylova, R. Delahay, and R. Cox, "Wildlife surveillance using deep learning methods," Ecology and Evolution, vol. 9, no. 17, pp. 9453–9466, Aug. 2019, DOI https://doi.org/10.1002/ece3.5410.

[35] Jackulin Mahariba, A. U. (2022). An efficient automatic accident detection system using inertial measurement through machine learning techniques for powered two wheelers. Expert Systems With Applications, 0957.

[36] Ahmed Yaseer, H. C. (2021). A Review of Sensors and Machine Learning in Animal Farming. 11th IEEE International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (p. 21). Jiaxing, China: EEE Xplore.

[37] Dario Augusto Borges Oliveira, L. G. (6 September 2021). A review of deep learning algorithms for computer vision systems in livestock. Livestock Science,1-15.

[38] Heegon Kim, J. S. (2015). Automatic Identification of a Coughing Animal using Audio and Video Data. ISCC 2015 (p. http://pos.sissa.it/). Guangzhou, China: ISCC.

[39] Jun Bao, Q. X. (2022). Artificial intelligence in animal farming: A systematic literature review. Journal of Cleaner Production, 0959.

[40] Md Ekramul Hossain, M. A. (2022). A systematic review of machine learning techniques for cattle identification: Datasets, methods and future directions. Artificial Intelligence in Agriculture, 40.

[41] Md Sultan Mahmuda, A. Z. (2021). A systematic literature review on deep learning applications for precision cattle farming. Computers and Electronics in Agriculture, 0168.

[42] Prashanth C. Ravoor, S. T. (2020). Deep Learning Methods for Multi-Species Animal Re-identification and Tracking – a Survey. Computer Science Review,0137.

[43] Qiumei Yang, D. X. (2020). A review of video-based pig behavior recognition. Applied Animal Behaviour Science,1-7.

[44] Rodrigo García, J. A. (2020). A systematic literature review on the use of machine learning in precision livestock farming. Computers and Electronics in Agriculture,0168.

[45] S Jeevitha, D. V. (May 2020). A Review of Animal Intrusion Detection System. International Journal of Engineering Research & Technology,129-1221.

[46] Shi Dong, P. W. (2021). A survey on deep learning and its applications. Computer Science Review, https://doi.org/10.1016/j.cosrev.2021.100379.

[47] Verma, A. D. (2020). Convolutional neural network: a review of models, methodologies and applications to object detection. Progress in Artificial Intelligence, Progress in Artificial Intelligence.

[48] Vigneshwaran Palanisamy, N. R. (2021). Detection of wildlife animals using deep learning approaches: A Systematic review. 21st International Conference on Advances in ICT for Emerging Regions (ICTer 2021) (pp. 153-158). India: IEEE Explore.

[49] Vipal Kumar Sharma, R. N. (11 September 2020). A comprehensive and systematic look up into deep learning based object detection techniques: A review. Computer Science Review,1574-0137.

[50] Weinstein, B. G. (2017). A computer vision for animal ecology. Journal of Animals Ecology, 533-545.

[51] A Literature Research Review on Animal Intrusion Detection and Repellent Systems. (2021). L. Ashok Kumar, R. Neelaveni, M. Kathiresh, P. Sweety Jose, N. Archana, S. Saravanakumar, 227.

[52] Ferrante, G. S., Rodrigues, F. M., Andrade, F. R., Goularte, R., & Meneguette, R. I. (2021). Understanding the state of the Art in Animal detection and classification using computer vision technologies. 2021 IEEE International Conference on Big Data (Big Data) (p. https://doi.org/10.1109/BigData52589.2021.9672049). Orlando, FL, USA: IEEE Xplore.

[53] M.Sowmya, D. M. (2021). A Review On Animal Detection Using Different Detection Techniques. Turkish Online Journal of Qualitative Inquiry (TOJQI), 8249 - 8254.

[54] Sreedevi C K, S. E. (2019). Automated Wildlife Monitoring Using Deep Learning. In Proceedings of the International Conference on Systems, Energy & Environment (p. 7). Kerala: SSRN.

[55] L. Ashok Kumar, R. N. (2021). A Literature Research Review on Animal Intrusion Detection and Repellent Systems. ICCAP 2021, (p. 227). Chennai, India.

[56] Manasa Kommineni, M. L. (2022). Agricultural Farms Utilizing Computer Vision (AI) And Machine Learning Techniques For Animal Detection And Alarm Systems. JOURNAL OF PHARMACEUTICAL NEGATIVE RESULTS, https://doi.org/10.47750/pnr.2022.13.S09.411.

Studying the Security and Privacy Issues of Big Data in the Saudi Medical Sector

Ramy Elnaghy, Hazem M. El-Bakry
Faculty of Computer and Information Sciences, Mansoura University, Mansoura 35516, Egypt

Abstract—In today's era of Big Data, with the integration of data from various systems, devices, and machines used by healthcare service providers, health insurance companies, and their sub-sectors, maintaining privacy and security has become crucial. It is important to uphold the confidentiality and security of data exchanged between data service providers and insurance companies as required by law. The purpose of this paper is to focus on addressing the security and privacy issues associated with healthcare data, particularly concerning medical data in both in- transit and at-rest modes. We aim to provide a proposed solution to enhance data security and maximize privacy protection.

Keywords—Security; privacy; healthcare; medical data; big data

I. Introduction

The healthcare industry has made significant strides in recent years with the adoption of electronic patient records and the digitization of healthcare workflows. However, this transformation has resulted in an explosion of clinical data, which is characterized as big data. The abundance of clinical data has created numerous opportunities for healthcare organizations to leverage data analytics, clinical decision support, and disease surveillance to optimize treatment and improve patient outcomes. Nonetheless, with the benefits of big data come numerous challenges, particularly regarding information security and privacy. Healthcare organizations must adopt proactive measures to safeguard sensitive information and prevent security breaches and other incidents. This paper aims to address the security and privacy concerns associated with medical data, particularly in in-transit and at-rest modes. We propose a solution that enhances data security and maximizes privacy protection. The adoption of big data technologies in healthcare has created numerous opportunities for clinical decision support, disease surveillance, population health management, and treatment optimization. However, with the increased use of electronic patient records and digitization of healthcare workflows come significant security and privacy risks. Medical data is highly sensitive, and its confidentiality and privacy must always be protected. Healthcare organizations must take a proactive, multi-layered approach to data security, including encryption, access control, network security, and employee training. Data encryption is a highly effective way to protect medical information. Encryption uses algorithms to convert plain text into cipher text, adding a layer of security to ensure that only authorized parties can access the data. In-transit encryption is used to protect data that is transmitted over a network or the internet, such as emails or data transfers between healthcare providers. At-rest encryption, on the other hand, is used to protect data that is stored on electronic devices, such as servers or hard drives. To implement data encryption, healthcare organizations must first identify which data needs to be protected and where it is stored. Then, appropriate encryption algorithms and methods must be selected and applied to the data. The encryption keys must be securely managed and stored to ensure that only authorized parties can access the data. While data encryption is an effective security measure, it is not a complete solution. Healthcare organizations must also implement access controls to limit access to sensitive information to authorized personnel only. Network security measures such as firewalls, intrusion detection, and prevention systems can help protect against cyberattacks and data breaches. Employee training is also essential to ensure that all staff members understand their responsibilities in safeguarding patient data and are aware of potential security threats. In summary, healthcare organizations must adopt a comprehensive, multi-layered approach to data security that includes encryption, access control, and network security, through utilizing big data and modern technologies such as blockchain [1].

II. Literature Review

Soumya et al. in [2], reviews the security issues associated with big data in Internet of Things (IoT) and cloud computing. The authors discuss the challenges of securing big data in IoT and cloud computing, including issues related to data privacy, data integrity, authentication, and access control. They also discuss various security solutions that have been proposed to address these challenges. Parsa et al. in [3] proposes a security management framework for big data in smart healthcare. The authors identify various security challenges in smart healthcare, including issues related to data privacy, data integrity, and data availability. They then propose a framework that includes various security measures, such as access control, encryption, and monitoring, to address these challenges. Hayat et al. in [4] presents a security model for the big healthcare data lifecycle. The authors discuss the security challenges associated with healthcare data, including issues related to data privacy, data integrity, and data confidentiality. They then propose a security model that includes various security measures, such as access control, data encryption, and data backup, to address these challenges. Aqeel et al. in [5] reviews the security and privacy issues associated with big data in healthcare applications. The authors discuss the challenges of securing big data in health-care, including issues related to data privacy, data integrity, and data confidentiality. They also discuss various security and privacy solutions that have been proposed to address these challenges.

Musfira et al. in [6] surveys various big data security solutions that have been proposed to address security issues in healthcare. The authors discuss various security solutions, such as access control, data encryption, and data backup, and evaluate their effectiveness in addressing the security challenges associated with big data in healthcare. Isabel et al. in [7] analyzes the security issues associated with big data in healthcare. The authors discuss the challenges of securing big data in healthcare, including issues related to data privacy, data integrity, and data confidentiality. They also discuss various security solutions that have been proposed to address these challenges. Saraladevi et al. in [8] studies the security issues associated with big data and Hadoop. The authors discuss the challenges of securing big data in Hadoop, including issues related to data privacy, data integrity and data confidentiality. They also discuss various security measures that can be implemented to address these challenges. ABHISHEK et al. in [9] analyzes the issues related to healthcare data integrity. The authors discuss the challenges of maintaining data integrity in healthcare, including issues related to data accuracy and data consistency. They also propose various measures to ensure data integrity, such as data validation and data verification.

Kanika et al. in [10] proposes an encryption approach based on the Rivest-Shamir-Adleman (RSA) algorithm to preserve the confidentiality of big data. The authors discuss the challenges of securing big data and propose an encryption approach based on RSA to ensure data confidentiality. Mustafa et al. in [11] presents a systematic review of privacy-preserving healthcare data sharing on blockchain. The authors discuss the challenges of ensuring privacy in healthcare data sharing and evaluate various blockchain-based solutions that have been proposed to address these challenges. Peng Xi et al. in [12] reviews the blockchain-based solutions that have been proposed to secure healthcare data sharing. The authors discuss various blockchain-based solutions, such as smart contracts and permissioned blockchain, and evaluate their effectiveness in ensuring data security and privacy. Mehak et al. in [13] identified potential vulnerabilities and attacks that could compromise the security of Hadoop-based big data systems. The purpose of the paper is to highlight these issues and provide recommendations for mitigating them. Mohan et al. in [14] proposed an efficient and secure big data storage solution using triple data encryption standard (Triple Data Encryption Standard (3DES)) in a cloud environment. The purpose of the paper is to provide a secure storage solution for big data in the cloud. Rakib et al. in [15] presented privacy- preserving k-nearest neighbors (KNN) training scheme over blockchain-based encrypted health data. The purpose of the paper is to propose a secure and privacy-preserving KNN training mechanism for health data in a blockchain-based system.

Parvathaneni et al. in [16] proposed a blockchain-based solution for secure healthcare data communication among non-terminal nodes in the Internet of Things (IoT) architecture in 5G networks. The purpose of the paper is to provide a secure and efficient data communication mechanism for healthcare IoT systems. Rafik et al. in [17] proposed fully homomorphic en- cryption (Fully Homomorphic Encryption (FHE)) algorithms for secure big data analysis. The purpose of the paper is to provide a solution for secure big data analysis in a privacy- preserving manner using FHE. Jose et al. in [18] discussed the security and privacy issues of big data in general. The purpose of the paper is to provide an overview of the security and privacy challenges associated with big data. Gousiya et al. in [19] proposed a sandbox security model for the Hadoop file system. The purpose of the paper is to provide a security mechanism for the Hadoop file system, which enables secure execution of untrusted code. Iroju et al. in [20] This paper provides an overview of the prospects, challenges, and solutions related to big data in healthcare. It discusses the potential of big data to improve healthcare outcomes and reduce costs, as well as the challenges related to data privacy, security, and integration. The paper also discusses some of the current solutions to these challenges, such as the use of big data analytics and cloud computing.

MATTURDI et al. in [21 provides a comprehensive review of the security and privacy challenges of big data, with a focus on data protection, access control, and privacy- preserving techniques. The paper also discusses some of the current solutions to these challenges, such as the use of encryption, authentication, and authorization mechanisms, as well as data anonymization and data masking techniques. Karim et al. in [22] provides a review of the security and privacy challenges of big data in healthcare, with a focus on the protection of patient data. The paper discusses the importance of ensuring the confidentiality, integrity, and availability of healthcare data, as well as the challenges related to data privacy, security, and interoperability. The paper also discusses some of the current solutions to these challenges, such as the use of encryption, access control, and data anonymization techniques. Gunasekaran et al. in [22 focuses on the security challenges of big data in the healthcare industry 4.0, which includes the integration of advanced technologies such as IoT, Artificial Intelligence (AI), and cloud computing. The paper discusses the importance of ensuring the security and privacy of healthcare data in this new era, and proposes a framework for big data security intelligence that includes threat intelligence, security analytics, and incident response.

III. OBJECTIVES

The purpose of this paper is to address the security and privacy challenges encountered in the Saudi healthcare and medical insurance sector during the storage and transfer of big data. Specifically, this paper aims to achieve the following:

1) Identify potential security and privacy risks associated with storing and transferring big data. This includes identifying the risks of unauthorized access to sensitive information, data breaches, cyber-attacks, hacking, insider threats, and data loss.

2) Analyze current practices and technologies used for securing big data in the healthcare and medical insurance sector. This analysis will provide an understanding of the existing security and privacy measures implemented by the industry and highlight any gaps or deficiencies in current practices.

3) Propose technical solutions by applying best practices and techniques for addressing the security and privacy challenges associated with storing and transferring big data.

The proposed solutions aim to maximize the security and privacy of medical information.

This paper proposes a multi-layered approach to address the security and privacy risks associated with storing and transferring big data in the healthcare and medical insurance sectors. The proposed solution includes implementing encryption techniques, access control measures, and security protocols to minimize the risk of unauthorized access to sensitive information. Additionally, the solution involves the use of advanced technologies such as blockchain to ensure data integrity and prevent data tampering.

The implementation of the proposed solution can provide many benefits, including better protection of patient information, improved data accuracy, and increased trust and confidence among stakeholders. Moreover, it can help healthcare organizations comply with relevant regulations and standards, such as the General Data Protection Regulation (General Data Protection Regulation (GDPR)) and the Health Insurance Portability and Accountability Act (Health Insurance Portability and Accountability Act (HIPAA)).

IV. Methodology

To ensure the security and privacy of big data, it is essential to adopt a comprehensive approach that covers all aspects of data protection. In this paper, we propose a solution that combines multiple techniques to address the various vulnerabilities that could compromise healthcare security and privacy using big data.

The proposed solution will encompass techniques such as data masking, encryption, access controls, and data anonymization. By combining these techniques, the proposed solution provides a comprehensive approach to ensure data security and privacy, particularly in healthcare organizations when processing medical insurance data for insurance companies or claim transaction operators. The implementation of these techniques can significantly reduce the risks of data breaches, unauthorized access, and misuse of data, thereby enhancing data security and privacy.

A. Medical Insurance Transaction Ecosystem

To better understand the techniques and solutions involved, it is important to first gain insight into the transaction system of medical insurance. This includes understanding how data is ingested, stored, and processed.

The claim processing system typically comprises three main domains or parts, as illustrated in Fig. 1.

1) Healthcare Provider Integration Tier: A collection of tools and configurations are utilized to extract data from Hospital Management System (HMS) or Practice Management Application (PMA) systems and safeguard the original data before transmitting it to the data processing stage.

2) Data Processing Tier: During this tier, the data Extract, Transform, Load (ETL) process will occur, followed by data analysis. The data will then be structured and prepared for the subsequent phase, which involves transmitting the final claims to the insurance companies as shown in Fig. 2.

3) Insurance Company Integration Tier: During this stage, the integration hub component will direct the data to the appropriate destination based on the configurations and the targeted beneficiary or consumers.

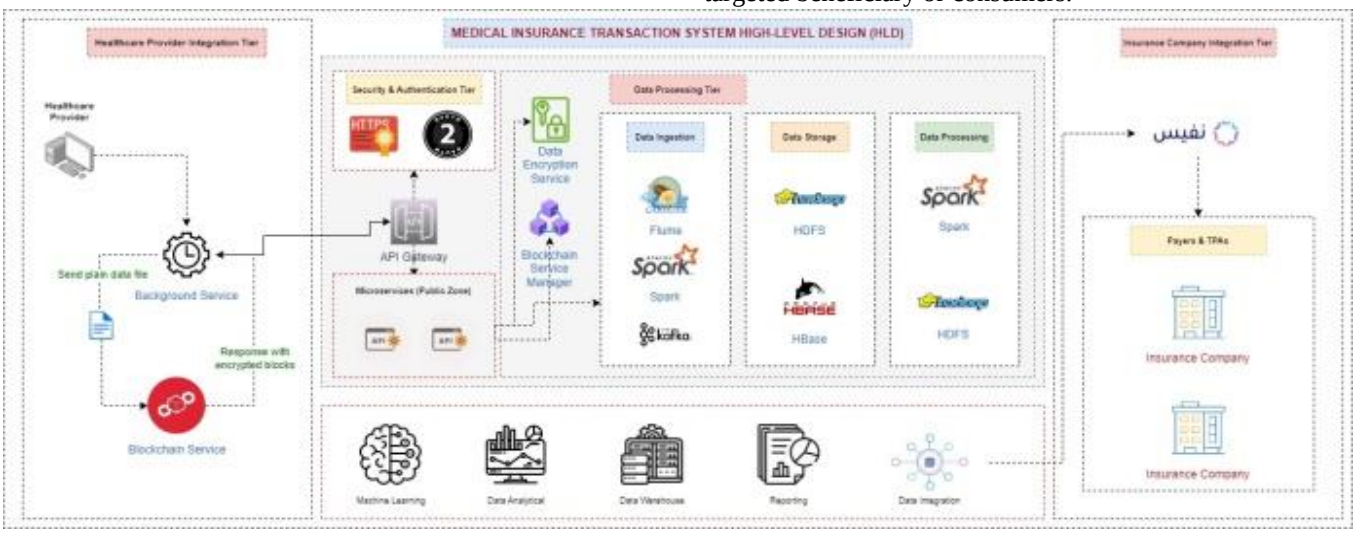


Fig. 1. Claim processing system.

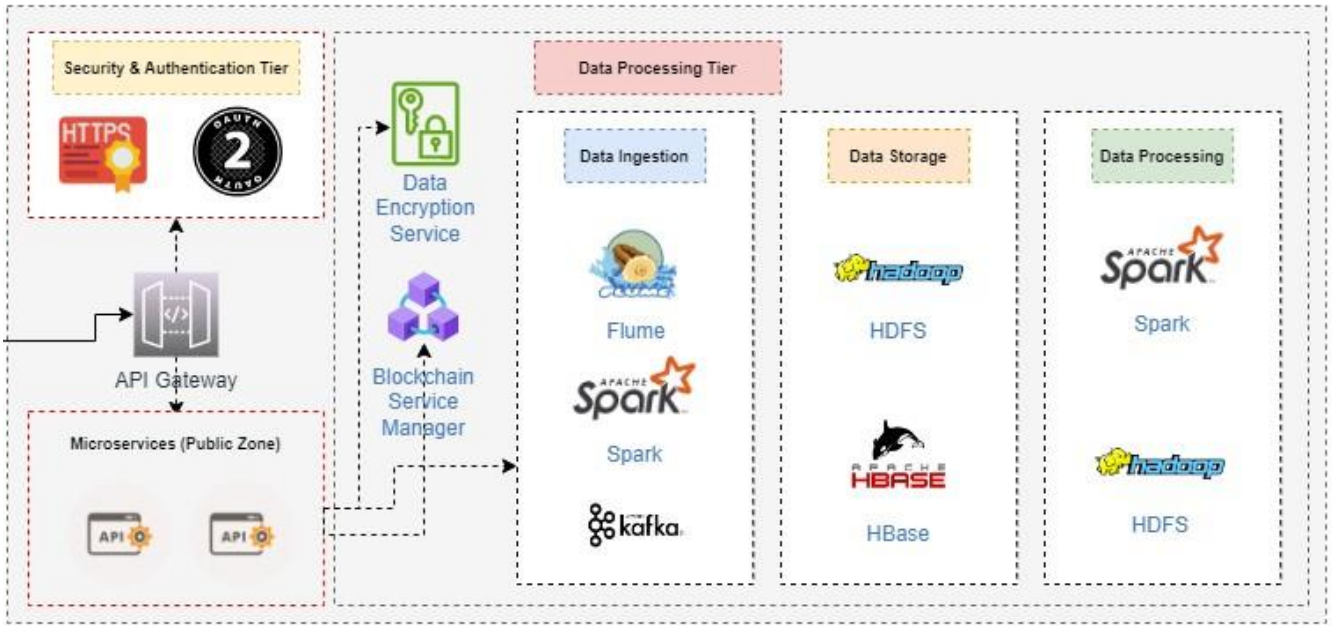


Fig. 2. Data processing tier.

B. A Common Data Processing Pipeline in the Medical Insurance Domain Includes the following:

1) Data Ingestion: The initial stage of the data processing pipeline involves collecting data from diverse sources, such as electronic health records and claims data, and ingesting it into a big data storage platform, such as Hadoop's Hadoop Distributed File System (HDFS) or other distributed file systems.

2) Data Cleaning: After data ingestion, cleaning and preprocessing are necessary to remove inconsistencies, missing data, or errors, which can involve various techniques like deduplication, normalization, and data imputation.

3) Data Transformation: Here, raw data is transformed into a format suitable for processing and analysis. This involves tasks such as data format conversion, data aggregation at different granularities, and feature engineering.

4) Data Analysis: After data transformation, patterns and trends can be identified through analysis using techniques like statistical analysis, machine learning, and predictive modeling.

5) Data Visualization: Visual aids such as dashboards, charts, and graphs can be utilized to present the analysis results. This helps in conveying the data insights to stakeholders and decision-makers.

6) Data Storage: After completing the analysis, the processed data can be securely stored in scalable data stores like Hadoop's HDFS or other distributed file systems. This stored data can be further transformed and analyzed in the future as required.

C. Compliance with Regulations

Different regulations and standards, such as HIPAA and GDPR, govern the storage and transfer of medical information that we need to comply with.

1) The Saudi Arabian Standards Organization (The Saudi Arabian Standards Organization (SASO)): SASO sets standards for consumer products and services, including those related to data protection.

2) The Saudi Arabian Monetary Authority (The Saudi Arabian Monetary Authority (SAMA)): SAMA regulates financial institutions in Saudi Arabia, including insurance companies, and has issued regulations related to data protection for the financial sector.

3) HIPAA: This law requires the establishment of national standards for electronic health care transactions [22].

4) The Privacy Law: This law was enacted in 2019 and governs the collection, processing, use, and disclosure of personal data in Saudi Arabia. The law applies to both government and private sector entities and requires organizations to obtain consent for the collection and use of personal data and to implement appropriate security measures to protect personal data.

5) The Cybercrime Law: This law criminalizes a range of activities related to cybercrime, including unauthorized access to computer systems and the theft of personal data.

Data protection regulations and laws in some countries as mentioned in study [22] address the growing thicket of applicable data protection legislation.

D. Threats and Risks

The medical insurance sector deals with vast amounts of sensitive data, including patient medical records, financial information, and insurance claims. Storing and transferring this big data poses significant security and privacy risks [23], such as data breaches, loss, and unauthorized access [24].

1) Unauthorized access: Malicious individuals may gain access to medical information in big data, leading to identity theft, fraud, and other malicious activities.

2) Data breaches: Compromised systems or networks can result in the loss or theft of sensitive medical information, which can be exploited for malicious purposes.

3) Insider threats: Misuse or disclosure of medical information by employees or contractors can lead to breaches in confidentiality or privacy.

4) Cyberattacks: Attacks like denial-of-service or malware can compromise the availability or integrity of medical information in big data.

5) Data loss: Accidental deletion or system failure can result in the permanent loss of medical information.

6) Inadequate security measures: Weak passwords or lack of encryption can make medical information in big data vulnerable to attacks and breaches.

7) Lack of regulatory compliance: Noncompliance with regulations and standards like HIPAA or GDPR can lead to legal and financial penalties.

E. Security Measures and Techniques

Various technologies are in use for protecting the security and privacy of healthcare data. The most widely used technologies are [22]:

1) Access control: Restricts medical information access based on user roles and privileges.

2) Encryption: Encodes medical information to prevent unauthorized access.

3) Data masking: Replaces sensitive information with fake values to protect privacy while allowing for research.

4) Data minimization: Reduces the amount of sensitive information stored and transferred.

5) Data backups: Stores copies of medical information in a separate location.

6) Penetration testing: Simulates cyberattacks to identify system vulnerabilities.

7) Security audits: Assessments of security measures and practices to ensure compliance with regulations and standards.

F. Privacy-preserving Techniques

1) Differential privacy: Adds noise to data to protect the privacy and prevent re-identification attacks.

2) Anonymization: Removes identifying information from medical data to protect privacy while allowing data analysis [24].

3) Pseudonymization: Replaces identifying information with pseudonyms to protect privacy while allowing data analysis [25].

4) Secure multi-party computation: Allows multiple parties to compute medical data without revealing it to each other, enabling collaborative analysis while protecting privacy.

5) Homomorphic encryption: Enables computation on encrypted data without decrypting it, protecting privacy while allowing data analysis [15].

6) Secure data sharing: Enables secure sharing of medical data between parties with appropriate permissions, protecting privacy while allowing data analysis [9].

G. Proposal

Several existing techniques have been employed to secure healthcare data, such as access controls, data encryption, and anonymization. Studies have shown that using blockchain can effectively limit the risk of data integrity by restricting unauthorized access to sensitive data [12] (Peng et al., 2022). Similarly, the use of data encryption is an effective method for maintaining data confidentiality and integrity [10] (ABHISHEK et al., 2019). Furthermore, anonymization, column-based encryption, and two-way authentication techniques are effective in preserving patient privacy (Musfira et al., 2018). These results demonstrate the value of utilizing such techniques to enhance the security and privacy of healthcare data as shown in Table I.

TABLE I. SECURITY AND PRIVACY TECHNIQUES

Ref	Year	Method / Technique
[21]	2004	Using integrated Rule-Oriented Data (Integrated Rule-Oriented Data (iRODS)) is proposed to be the solution to ensure security and privacy in big data.
[8]	2015	Highlighting the big data security issues and how to secure the Hadoop File System (HDFS) by using Kerberos, Algorithm, and Name node.
[22]	2017	The proposed system uses key management security mechanisms to protect big data.
[4]	2018	Discussed the threats and countermeasures of the healthcare data life cycle and its suggested defense.
[6]	2018	Using the proposed security framework uses column-based encryption and two-way authentication for privacy and multi-biometric-based key generation.
[13]	2018	Using Kerberos authentication protocol, and attribute-based encryption.
[10]	2019	Using the RSA algorithm to encrypt sensitive medical data.
[2]	2020	Highlighting the cloud security threats, Fog computing, IoT security threats, and Blockchain security threats and addressing the defense mechanisms of different computing technologies.
[9]	2020	Discuss the blockchain as a data integrity technique.
[11]	2020	Examining blockchain-based techniques used to privacy-preserving healthcare data and discussing smart contracts and PKI as blockchain technologies.
[19]	2020	Using Sandbox security for Map Reduce jobs to scan the jar files and check if it contains any harmful code.
[3]	2021	A proposal for a security framework is being made that utilizes the logistic equation, hyperchaotic equation, and DNA encoding techniques to cipher and encrypt images. The encrypted images will then be stored in shares across multiple cloud-based servers for distributed storage.
[16]	2021	Discussing different encryption algorithms used in blockchain.
[12]	2022	Using blockchain and IOMT technologies to protect healthcare data.
[14]	2022	Using Triple Data Encryption Standard (TDES) to secure big data in the cloud environment.
[17]	2022	Focused on using homomorphic encryption technology to secure Big Data processing.

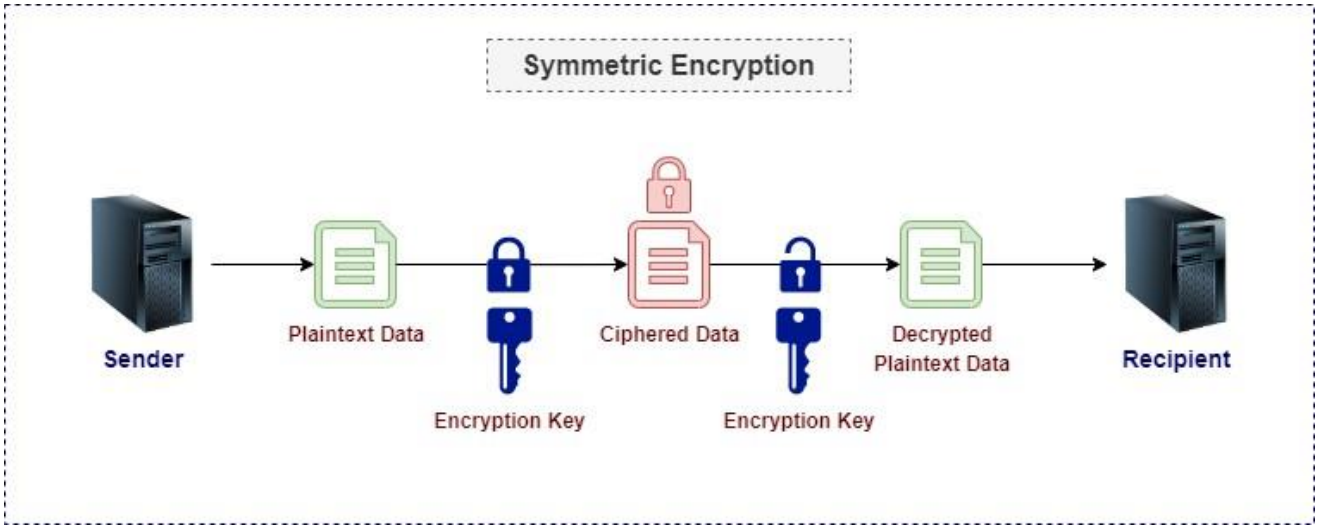


Fig. 3. Symmetric encryption.

Proposed Techniques: Ensuring the security of patient information is critical in the healthcare industry. The following suggested implementations can maximize the security and privacy of the patient's electronic medical records (Electronic Medical Records (EMR)), at-rest, and in-transit modes.

1) Data Encryption: Encryption is a method of encoding information so that only authorized parties can access it.

Symmetric encryption – Advanced Encryption Standard (Advanced Encryption Standard (AES)). Fig. 3 shows the symmetric encryption.

The RSA or Advanced Encryption Standard is one of the safest and most secure types of encryption. As a matter of fact, RSA is used by the United States Government to secure classified information as well as other hardware and software products. RSA, on the other hand, makes use of block ciphers that encrypt information one fixed-size chunk at a time. RSA has three key sizes: 128-bit, 192-bit, and 256-bit. The larger the key size, the stronger the encryption, therefore, we are going to use the 256-bit key as shown in the below algorithm [26].

Algorithm 1 GenerateAESKey(password)

```
function GENERATEAESKEY(password)
    sha256 ← SHA256.CREATE()
    passwordBytes←ENCOD-
ING.UTF8.GETBYTES(password)
    hash ← SHA256.COMPUTEHASH(passwordBytes)
    aes ← AES.CREATE()
    AES.KEYSIZE ← 256
    AES.KEY ← hash
    return AES.KEY
end function
```

Hashing SHA-256 Hashing is a one-way encryption technique that generates a unique fixed-length string of characters from a given input. There are several hashing algorithms, including MD5 and SHA-1, and SHA-256 is the most commonly used algorithm in blockchain technology, and it is recommended for its security and performance, we are going to use the SHA-256 as shown in the below algorithm [16].

Algorithm 2 CalculateHash(publicKey)

```
function CALCULATEHASH(publicKey)
    sha256 ← create SHA256 object
    inputString ← concatenate PreviousHash, Data and
publicKey using "-" separator
    inputBytes ← convert inputString to bytes using ASCII
encoding
    outputBytes ← compute hash of inputBytes using sha256
    return outputBytes as Base64 encoded string
end function
```

2) Data Encryption Flow stepsenumerate1 [steps,1]label=Step 0:

3) Hash the original data using the SHA-256 RSA algorithm: This generates a unique, fixed-length hash value that represents the original data.

Algorithm 3 EncryptAES256(plainText, publicKey)

```
function ENCRYPTAES256(plainText, publicKey)
    publicKeyBytes ← publicKey
    iv ← create new byte array of size 16
    plainBytes ← convert plainT ext to bytes using UTF-8
encoding
    aesAlg ← create new AES object
    aesAlg.KeySize ← 256
    aesAlg.Key ← publicKeyBytes
    aesAlg.IV ← iv
    aesAlg.M ode ← CBC
    encryptor ← create encryptor object using aesAlg Key and
IV
    memoryStream ← create new memory stream object
    cryptoStream ← create new crypto stream object using
memoryStream, encryptor and Write mode
    cryptoStream.W rite(plainBytes, 0, plainBytes.Length)
    cryptoStream.F lushF inalBlock()
    cipherBytes ← convert memoryStream content to byte
array
    return cipherBytes as Base64 encoded string
end function
```

4) Encrypt the original data using the RSA 256-bit public computed key: This ensures that the data is secure during transmission and can only be decrypted by someone who has access to the public computed key, following is the encryption algorithm.

Algorithm 4 DecryptAES256(cipherText, publicKey)

```
1: function DECRYPTAES256(cipherText, publicKey)
2:     privateKeyBytes ← publicKey
3:     iv ← create new byte array of size 16
4:     cipherBytes ← convert cipherText to bytes using Base64 decoding
5.     aesAlg ← create new AES object
6.     aesAlg.KeySize ← 256
7.     aesAlg.Key ← privateKeyBytes
8.     aesAlg.IV ← iv
9.     aesAlg.Mode ← CBC
10.    decryptor ← create decryptor object using aesAlg Key and IV
11.    memoryStream ← create new memory stream object using cipherBytes
12.    cryptoStream ← create new crypto stream object using memoryStream, decryptor and Read mode
13.    plainBytes ← create new byte array of size cipherBytes.Length
14:    decryptedByteCount ←cryptoStream.Read(plainBytes, 0, plainBytes.Length)
15:    return plainBytes as UTF-8 encoded string from index 0 to decryptedByteCount
16: end function
```

5) Concatenate the hash value with the encrypted data: This ensures that the hash value is bound to the encrypted data and cannot be altered without detection.

6) Add the concatenated data on a blockchain: This ensures that the data is secured by the distributed ledger and cannot be altered without detection.

7) Send the concatenated data: using a secured channel such as a Token-based Representational State Transfer (REST) Application Programming Interface (API), as shown in the process flow below.

8) Retrieve the concatenated data: from the blockchain: The data can be retrieved using a specific transaction ID or other identification methods.

9) Verify the integrity of the data: using the hash value: The hash value can be recalculated from the retrieved data and compared to the original hash value to ensure that the data has not been tampered with.

10) Decrypt the encrypted data: using the public computed key: This ensures that the original data can be recovered.

11) Convert the decrypted data back into the original format: This ensures that the data can be used in its original form. Using Secure Hash Algorithm (SHA)-256 and RSA 256-bit key pair with blockchain provides an additional layer of security and integrity to the data being transmitted.

12) Blockchain: Blockchain technology has the potential to revolutionize the healthcare industry by enabling secure and decentralized storage and sharing of patient data. Blockchain technology can help by enabling encrypted, distributed storage of data that can be securely shared for public health purposes [27]. In addition, blockchain networks can be used in the healthcare system to preserve and exchange patient data through hospitals, diagnostic laboratories, pharmacy firms, insurance companies, and other healthcare providers. This can help improve the security and integrity of healthcare data management. Blockchain-based healthcare data management system between multiple stakeholders (nodes) within a healthcare ecosystem [28].

Algorithm 5 DICOM Encryption Algorithm

```
1: function ENCRYPTDICOMFILE(inputFile, outputFile, key, iv)
2:     originalFileBytes ← read all bytes from the inputFile
3:     aes ← create new AES object
4:     aes.KeySize ← 256
5:     aes.Key ← convert key from Base64 string to byte array
6:     aes.IV ← convert iv from Base64 string to byte array
7:     aes.Padding ← PKCS7
8:     encryptor ← create encryptor object using aes
9:     encryptedStream ← create new memory stream object
10:    cryptoStream ← create new crypto stream object using encryptedStream, encryptor and Write mode
11:    cryptoStream.Write(originalFileBytes, 0, originalFileBytes.Length) cryptoStream.FlushFinalBlock()
13:    encryptedFileBytes ← convert encryptedStream to byte array
14:    write all bytes from encryptedFileBytes to the outputFile
15:    end function
```

13) Secure Sockets Layer (Secure Sockets Layer (SSL)) and Transport Layer Security (Transport Layer Security (TLS)): SSL and TLS are cryptographic protocols that provide secure communication over the internet. These protocols use a combination of symmetric and asymmetric encryption, along with digital certificates, to secure data transmission [22].

- Secure File Transfer Protocols: Medical patient files can contain a variety of file types depending on the type of information being recorded and the needs of the healthcare provider. Here we are focusing on the information that might be fetched from machines or third-party systems in the form of files:
- Medical images: Medical imaging files such as X-rays, CT scans, MRIs, and ultrasounds are often included in patient files.
- Lab results: Reports from laboratory tests such as blood tests, urine tests, and biopsies are commonly included in patient files.
- Consent forms: Documents that record patient consent for medical procedures, treatment plans, and research studies.
- DICOM files are created by medical imaging equipment, and they contain both image data and

information about the patient and the imaging procedure. This information includes patient demographics, imaging protocols, acquisition parameters, and image annotations [6].

Here we are going to secure the file transfer channel by using Secure File Transfer Protocol (SFTP) protocol and Hypertext Transfer Protocol (HTTP), to ensure that data is transferred securely. The file itself should be encrypted while transferring from the providers to the receivers (payers and claim transaction system), below the Digital Imaging and Communications (DICOM) encryption and decryption algorithm.

Algorithm 6 DICOM Decryption Algorithm

```
function DECRYPTDICOM-FILE(inputFile, outputFile, key, iv)
    encryptedFileBytes ← read all bytes from the inputFile
    aes ← create new AES object
    aes.KeySize ← 256
    aes.Key ← convert key from Base64 string to byte array
    aes.IV ← convert iv from Base64 string to byte array
    aes.Padding ← PKCS7
    decryptor ← create decryptor object using aes
    decryptedStream ← create new memory stream object
    cryptoStream ← create new crypto stream object using encryptedFileBytes, decryptor and Read mode
    cryptoStream.CopyTo(decryptedStream)
    decryptedFileBytes ← convert decryptedStream to byte array
    write all bytes from decryptedFileBytes to the output File
end function
```

14) Anonymization and Aggregation: Anonymization removes identifiable information from data to protect patient privacy, using techniques like hashing, tokenization, and generalization, with tools such as MapReduce and Pig from Hadoop assisting in this process. Aggregation combines and summarizes data, helping minimize identification risks by presenting broader statistics, as seen when insurance companies aggregate health claims by zip code or age range. While these methods reduce data breach risks and maintain data utility for research, there's still some re-identification risk with anonymized and aggregated data, necessitating additional safeguards depending on data sensitivity and re-identification risks [25].

15) Data Accessibility: Due to the sensitivity of patient information, it is crucial that access to such data is restricted and not available to everyone who has access to the system, including healthcare providers and insurance companies. Once a doctor is providing service to a patient, they are allowed access to the EMR, but the patient's medical and personal information should be safeguarded with a high level of security after the service is provided. If necessary, access to the patient's EMR can be requested with the patient's consent through an electronic approval process. The patient can grant access through one of the following options:

- OTP: One-Time Password (OTP) stands for "one-Time Password." An OTP is a unique code or password that is valid for a single use, typically to verify a user's identity or for securing an online transaction. OTP is often used as an additional layer of security in multi-factor authentication, where users are required to provide more than one form of authentication before being granted access to a system or performing a sensitive action.
    - OTP: can be delivered to the user via various methods, such as:
    - Short Message Service (SMS): The OTP is sent to the user's mobile phone as a text message.
    - Email: The OTP is sent to the user's email address.
    - Authenticator app: The user installs an authenticator app on their mobile device, which generates a new OTP every few seconds.
    - Voice call: The OTP is delivered to the user via an automated voice call.

Here once the patient receives the OTP, he must provide it to the requester within a limited period before it expires. If the OTP is entered correctly and within the valid period, the doctor or physician is granted access or permission to perform the requested action. If the OTP is not entered correctly or within the valid period, the doctor may be locked out of the system or required to request new access.

- EMR Authenticator: Here we can integrate with any of the identity authenticators that supports the Open Authentication (OATH) Time-Based One-Time Password (TOTP) standard such as Google Authenticator, or Microsoft Authenticator. Additionally, Microsoft Authenticator also supports push notifications, which can provide an additional layer of security and convenience for users.

16) Big Data (Hadoop Security)

- Authentication: Authentication is the process of verifying the identity of users and ensuring that they have the necessary permissions to access the Hadoop cluster. Hadoop supports several authentication mechanisms such as Kerberos and Lightweight Directory Access Protocol (LDAP) [13].
- Authorization: Authorization is the process of determining what users are allowed to do once they have been authenticated. Hadoop provides mechanisms for fine-grained access control, such as Access Control Lists (Access Control Lists (ACL)s) and Role-Based Access controls (Role-Based Access Controls (RBAC)).
- Encryption: Encryption is the process of encoding data so that it cannot be read by unauthorized users. Hadoop provides encryption mechanisms for data at rest and in transit [29]. For data at rest, Hadoop supports encryption at the file system level using tools like HDFS Transparent Encryption [30]. For data in transit,

Hadoop supports encryption using Secure Sockets Layer (SSL) and Transport Layer Security (TLS).

- Auditing: Auditing is the process of recording all user activities on the Hadoop cluster. Hadoop provides auditing tools such as Apache Ranger and Apache Atlas, which can be used to track user activity and detect any security breaches.
- Network security: Hadoop clusters should be deployed in a secure network environment, with firewalls and other security measures in place to prevent unauthorized access. Network security can also include isolating different parts of the cluster and ensuring that only authorized users have access to sensitive data.[2]

17) Data Monitoring and Auditing Healthcare facilities and insurance companies need robust monitoring and auditing systems to ensure patient care and data quality, safety, and security. These systems monitor user access, track changes to records, and detect unauthorized or suspicious activity, ensuring data is used only for authorized purposes by authorized individuals [15]. They provide real-time alerts on security threats, enabling immediate action to protect patients and their data. Monitoring and auditing are crucial for maintaining data confidentiality, integrity, and availability, and ensuring compliance with regulations and best practices in patient data security [15] [31-41].

V. Discussion

In conclusion, the security techniques discussed in this paper provide a comprehensive approach to addressing the security and privacy concerns associated with healthcare data. By using data encryption, blockchain technology, secure API, and Hypertext Transfer Protocol Secure (HTTPS)/SFTP protocols, we can ensure the security and integrity of medical data both at rest and in transit. Additionally, the use of Kerberos can help ensure big data security. To ensure compliance with HIPAA standards and other regulations, it is important to limit data access to authorized parties only. Furthermore, in order to maintain a balance between performance and security, multi-layered security approaches have been employed, including big data security. It is important to note that the complexity of data processing can lead to poor performance and high latency, which is why it is crucial to carefully design and implement security measures that do not compromise the performance of the system. Overall, this proposed ecosystem provides a robust and secure solution for managing healthcare data and ensuring the privacy and security of patients' sensitive information.

VI. Conclusion

In conclusion, safeguarding healthcare data security is crucial to ensure patient privacy and uphold the confidentiality of their information. While advanced technologies like cloud computing bring numerous benefits to the healthcare industry, they also give rise to new security challenges that demand effective management. In this paper, we have applied a range of techniques, such as a customized Blockchain encryption mechanism, a combination of asymmetric and symmetric algorithms for data encryption, hashing, secure network protocols like HTTPS and SFTP, protection of APIs and resources within a cloud environment, and the incorporation of HDFS security.

The security strategies discussed in this paper form a solid foundation for healthcare organizations to establish a secure infrastructure that adheres to pertinent regulations and safeguards sensitive data. This paper highlights a multi-layered security approach that balances performance and data processing without compromising robust protection against cyber threats. After applying these techniques and integrating them, a highly secure data streaming system was achieved. During testing, neither passive nor active attacks were able to penetrate the protected data, demonstrating the effectiveness of the combined security measures.

It is essential for healthcare organizations to prioritize data security and consistently evaluate and improve their security measures to keep pace with emerging risks. By embracing a proactive and comprehensive approach to security, which includes the techniques outlined above, healthcare organizations can ensure the safety and privacy of their patients' data, instilling trust and confidence among all stakeholders.

References

[1] X. Yan, S. Feng, Y. Tang, P. Yin, and D. Deng, "Blockchain-based verifiable and dynamic multi-keyword ranked searchable encryption scheme in cloud computing," J. Inf. Secur. Appl., vol. 71, no. 103353, p. 103353, Dec. 2022.

[2] K. Soumya and S. Nath Mishra, "Big data security issues from the perspective of IoT and cloud computing: A review," Recent Advances in Computer Science and Communications, vol. 12, no. 1, pp. 1–22, 2020.

[3] S. A. Parsa, G. M. Parah, and K. Bhat, "A security management framework for big data in smart healthcare," Big Data Research, vol. 25, 2021.

[4] H. Khaloufi, K. Abouelmehdi, A. Beni-hssane, and M. Saadi, "Security model for big healthcare data lifecycle," Procedia Comput. Sci., vol. 141, pp. 294–301, 2018.

[5] Aqeel-ur-Rehman, I. U. Khan, and S. u. Rehman, "A review on big data security and privacy in healthcare applications," in Big Data Management. Cham: Springer International Publishing, 2017, pp. 71–89.

[6] M. Siddique, M. A. Mirza, M. Ahmad, J. Chaudhry, and R. Islam, "A survey of big data security solutions in healthcare," in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, ser. Lecture notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Cham: Springer International Publishing, 2018, pp. 391–406.

[7] B. Isabel and M. Garc´ıa-Zapirain, "Analysis of security in big data related to healthcare," Journal of Digital Forensics, Security and Law, vol. 12, no. 3, 2017.

[8] B. Saraladevi, N. Pazhaniraja, P. V. Paul, M. S. S. Basha, and P. Dhavachelvan, "Big data and hadoop-a study in security perspective," Procedia Comput. Sci., vol. 50, pp. 596–601, 2015.

[9] K. Pandey, A. I. Khan, Y. B. Abushark, M. M. Alam, A. Agrawal, R. Kumar, and R. A. Khan, "Key issues in healthcare data integrity: Analysis and recommendations," IEEE Access, vol. 8, pp. 40 612–40 628, 2020.

[10] K. Sharma, A. Agrawal, D. Pandey, R. A. Khan, and S. K. Dinkar, "RSA based encryption approach for preserving confidentiality of big data," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 5, pp. 2088–2097, May 2022.

[11] M. Tanrıverdi, A Systematic Review of Privacy-Preserving Healthcare Data Sharing on Blockchain, 2020.

[12] P. Xi, X. Zhang, L. Wang, W. Liu, and S. Peng, "A review of blockchain-based secure sharing of healthcare data," Appl. Sci. (Basel), vol. 12, no. 15, p. 7912, Aug. 2022.

[13] M. Choudhary, A. Singh Yadav, D. Kumar Yadav, and V. Pawar, "A review on hadoop security issues," A Review on Hadoop Security Issues.

[14] M. N. Ramachandra, M. Srinivasa Rao, W. C. Lai, B. D. Parameshachari, J. Ananda Babu, and K. L. Hemalatha, "An efficient and secure big data storage in cloud environment by using triple data encryption standard," Big Data Cogn. Comput., vol. 6, no. 4, p. 101, Sep. 2022.

[15] R. U. Haque, A. S. M. T. Hasan, Q. Jiang, and Q. Qu, "Privacy-preserving k-nearest neighbors training over blockchain-based encrypted health data," Electronics (Basel), vol. 9, no. 12, p. 2096, Dec. 2020.

[16] P. N. Srinivasu, A. K. Bhoi, S. R. Nayak, M. R. Bhutta, and M. Woz´niak, "Blockchain technology for secured healthcare data communication among the non-terminal nodes in IoT architecture in 5G network," Electronics (Basel), vol. 10, no. 12, p. 1437, Jun. 2021.

[17] R. Hamza, A. Hassan, A. Ali, M. B. Bashir, S. M. Alqhtani, T. M. Tawfeeg, and A. Yousif, "Towards secure big data analysis via fully homomorphic encryption algorithms, Entropy (Basel), vol. 24, no. 4,p. 519, Apr. 2022.

[18] J. Moura and C. Serra˜o, "Security and privacy issues of big data," in Cloud Security. IGI Global, 2019, pp. 1598–1630.

[19] G. Begum, S. Z. U. Huq, and A. P. S. Kumar, "Sandbox security model for hadoop file system," J. Big Data, vol. 7, no. 1, Dec. 2020.

[20] Olaronke and O. Oluwaseun, "Big data in healthcare: Prospects, challenges and resolutions," in 2016 Future Technologies Conference (FTC). IEEE, Dec. 2016.

[21] Matturdi, X. Zhou, S. Li, and F. Lin, "Big data security and privacy: A review," China Commun., vol. 11, no. 14, pp. 135–145, 2014.

[22] K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi, and M. Saadi, "Big data security and privacy in healthcare: A review," Procedia Comput. Sci., vol. 113, pp. 73–80, 2017.

[23] H. Patil and R. Kupwade, "Big data security and privacy issues in healthcare," in 2014 IEEE international congress on big data. IEEE, 2014, pp. 762–765.

[24] Y. Himeur, S. S. Sohail, F. Bensaali, A. Amira, and M. Alazab, "Latest trends of security and privacy in recommender systems: A comprehensive review and future perspectives," Comput. Secur., vol. 118, no. 102746, p. 102746, Jul. 2022.

[25] M. Paul, L. Maglaras, M. A. Ferrag, and I. Almomani, "Digitization of healthcare sector: A study on privacy and security concerns," ICT Express, Feb. 2023.

[26] P. Matta, M. Arora, and D. Sharma, "A comparative survey on data encryption techniques: Big data perspective," Mater. Today, vol. 46, pp. 11 035–11 039, 2021.

[27] L. Guo, H. Xie, and Y. Li, "Data encryption based blockchain and privacy preserving mechanisms towards big data," J. Vis. Commun. Image Represent. vol. 70, no. 102741, p. 102741, Jul. 2020.

[28] W. Y. Ng, T. E. Tan, P. V. H. Movva, A. H. S. Fang, K. K. Yeo, D. Ho, F. S. S. Foo, Z. Xiao, K. Sun, T. Y. Wong, A. T. H. Sia, and D. S. W. Ting, "Blockchain applications in health care for covid 19 and beyond: a systematic review," Lancet Digit. Health, vol. 3, no. 12, Dec. 2021.

[29] S. Gattoju and N. Vadlamani, "A survey on security of the hadoop framework in the environment of BigData," J. Phys. Conf. Ser., vol. 2089, no. 1, p. 012031, Nov. 2021.

[30] W. Rajeh, "Hadoop distributed file system security challenges and examination of unauthorized access issue," J. Inf. Secur., vol. 13, no. 02, pp. 23–42, 2022.

[31] J. J. Hathaliya and S. Tanwar, "An exhaustive survey on security and privacy issues in healthcare 4.0," Comput. Commun., vol. 153, pp. 311–335, Mar. 2020.

[32] H. M. El-Bakry, "An Efficient Algorithm for Pattern Detection using Combined Classifiers and Data Fusion," Information Fusion Journal, vol. 11, issue 2, April 2010, pp. 133-148.

[33] H. M. El-Bakry, and Nikos Mastorakis "New Fast Normalized Neural Networks for Pattern Detection," Image and Vision Computing Journal, vol. 25, issue 11, 2007, pp. 1767-1784.

[34] H. M. El-Bakry, and Nikos Mastorakis, "A New Fast Forecasting Technique using High Speed Neural Networks," WSEAS Transactions on Signal Processing, vol. 4, Issue 10, Oct. 2008, pp. 573-595.

[35] H. M. El-Bakry, "Fast Virus Detection by using High Speed Time Delay Neural Networks," Journal of Computer Virology, vol.6, no.2, 2010, pp.115-122.

[36] H. M. El-Bakry, and Nikos Mastorakis, "Realization of E-University for Distance Learning," WSEAS Transactions on Computers, vol. 8, issue 1, Jan. 2009, pp. 48-62.

[37] N. El-Rashidy, L. Alarabi, H. M. El-Bakry, S. Abdelrazek, T. ABUHMED, F. Ali and S. El-Sappagh "Sepsis Prediction in Intensive Care Unit based on genetic feature optimization and stacked deep ensemble learning," Neural Computing and Applications, vol. 34, pp. 3603–3632, 2022.

[38] H. El-Bakry, "Face Detection Using Neural Networks and Image Decomposition," Proc. of INNS-IEEE International Joint Conference on Neural Networks, 12-17 May, 2002, Honolulu, Hawaii, USA.

[39] H. El-Bakry, "Fast Face Detection Using Neural Networks and Image Decomposition," Proc. of the 6th International Computer Science Conference, AMT 2001, Hong Kong, China, December 18-20, 2001, pp.205-215.

[40] H. M. El-Bakry and M. Hamada, "A New Implementation for High Speed Neural Networks in Frequency Space," Lecture Notes in Artificial Intelligence, Springer, KES 2008, Part I, LNAI 5177, pp. 33-40.

[41] H. M. El-Bakry, and Q. Zhao, "Fast Time Delay Neural Networks," International Journal of Neural Systems, vol. 15, no.6, December 2005, pp.445-455.

A Novel Deep Learning-Assisted SVD-based Method for Medical Image Watermarking

Saima Kanwal, Feng Tao*, Rizwan Taj
School of Computer and Communication, Lanzhou University of Technology, Lanzhou, China

***Abstract*—In the present era, the administration of medical images faces various security challenges that necessitate the authentication of image source and origin for accurate patient identification. With the increasing exchange of medical images between hospitals to facilitate informed decision-making, the adoption of digital watermarking techniques has emerged as an efficient solution to address the imperceptibility and robustness requirements in medical imaging watermarking. This research work introduces a technically advanced approach that combines singular value decomposition (SVD) watermarking with deep learning segmentation models to enhance the security of medical image sharing and transfer. The primary objective is to seamlessly integrate the watermark while minimizing distortion to preserve critical medical information within the image. The proposed methodology involves utilizing a ResNet-based U-Net segmentation model to segment X-Ray radiographs into the Region of Interest (ROI) and the Region of Non-Interest (RONI). The watermark data is then encoded into the ROI using singular value decomposition. Subsequently, the ROI and RONI are merged to reconstruct the complete image, preserving its original identity. Additionally, XOR encryption is applied to the watermarked image to enhance data integrity and copyright protection. On the other side of the methodology, the reconstructed image is once again separated into ROI and RONI. The ROI is decoded to recover the original transferred content. To assess the efficacy of the proposed method, a publicly available X-Ray radiograph dataset is employed, and evaluation metrics demonstrate an impressive segmentation accuracy of 98.27%. The proposed approach ensures information integrity, patient confidentiality during data sharing, and robustness against various conventional attacks, demonstrating its effectiveness in the field of medical image watermarking.**

***Keywords*—Singular value decomposition; medical image watermarking; digital watermarking; deep learning**

I. INTRODUCTION

With the introduction of 5G networks as an example, communication technologies have advanced quickly, dramatically altering many facets of daily life [1]. This shift in paradigm has been facilitated by simultaneous significant progressions in the fields of data analytics, computing in the cloud, and online storage. Emerging within the framework of the Internet of Medical Things (IoMT), novel approaches for both diagnosis and treatment have surfaced. These encompass telemedicine, web medicine, systematic diagnosis, and smart medicine. These cutting-edge trends are accompanied by highly advanced diagnostic and therapeutic tools, state-of-the-art sensors for medical, immersive virtual reality technologies, and complex artificial surgical methods [2]. Advancements in healthcare tech merge for better diagnosis & treatment, reshaping the ecosystem, empowering doctors, and enhancing overall care. Medical imaging transforms through e-diagnosis workflows, forming core modern healthcare structures. Hospital systems utilize HIS and advanced imaging platforms for seamless management of varied digital images (X-ray, Ultrasound, MR, and CT) [3]. PACS in hospitals securely store and retrieve these images, while DICOM standardizes their acquisition, transmission, storage, and exchange with patient data. By synergistically leveraging the functionalities of HIS, medical imaging platforms, PACS, and adhering to the DICOM standard, healthcare professionals are empowered with swift access to an extensive collection of high-fidelity digital images [4]. Seamless integration ensures precise diagnosis, treatment planning, and overall care. Interchange of medical images aids diagnosis, therapy, education, and consultations within established PACS workflows. Yet, external transmission to third parties poses potential risks of unauthorized image alterations, impacting accuracy and potentially endangering patients' lives [5]. As a result, maintaining the integrity and authenticity of medical images has become of utmost importance, calling for the implementation of strong controls and strict protections both within internal systems and during the transmission and exchange processes with external systems. A secure and reliable framework for the lifecycle management of medical images must be established, which necessitates the adoption of cutting-edge methods and technology.

Several strategies are used to handle the challenges associated with copyright protection, the confidentiality of healthcare images, and the diagnostic and personal information of patients. For these goals, methods like digital picture watermarking, cryptography, and steganography are frequently used. One of these techniques, digital picture watermarking, has attracted a lot of attention from researchers due to its unique characteristics that might offer solutions for protecting copyright and safeguarding medical data. Digital image watermarking makes it possible to incorporate subtle or noticeable watermarks into medical photographs, allowing copyright enforcement, integrity verification, and authentication [6]. Watermarking conceals metadata in images, safeguarding patient privacy and ensuring traceability. It operates in spatial and frequency domains: spatially by adjusting pixel grey levels and frequency-wise via methods like Discrete Fourier Transform (DFT) or Discrete Cosine Transform (DCT) to embed the watermark into the image [7]. The secret information is then incorporated by tampering with the frequency coefficients after conversion. The amplitudes and phases of the various frequency components that are present in the image are represented by these coefficients. By

*Corresponding Author.

changing these coefficients, the watermark is smoothly included into the image's representation in the frequency domain. Based on how the embedded watermark is extracted, watermarking methods are further divided into different categories. For non-blind watermarking, the source image and the watermarked image are both necessary in order to obtain the watermark. As a result, both the original host image and the matching watermark image must be accessible during the extraction process [8]. In contrast, semi-blind watermarking entails using both the watermarked image and the secret key throughout the extraction process. The watermark is included in the watermarked image and the secret key acts as a cryptographic parameter. Finally, only the secret key is required to extract the watermark in blind watermarking techniques. Both types of watermarking techniques, robust watermarking and fragile watermarking, each serve a particular function in the context of data protection and verification [9]. Watermarking aims for resilient data embedding, crucial for protecting against tampering during transfer, especially for copyright protection. Resilience determines resistance to outsider attacks or unintentional distortions during transmission. Fragile watermarking, on the other hand, detects alterations in watermarked content, ensuring integrity and authenticity. It verifies material reliability by identifying inconsistencies, serving as a vital tool for content authentication and integrity checks [10]. When discussing watermarking, as we are dealing with images, the terms "frequency domain" and "spatial domain" in the context of imaging apply to various images' representations [11]. In imaging, the term "spatial domain" describes how a picture is represented as a grid of pixels, where each pixel holds information about color or intensity and correlates to a particular place. On the other hand, the image is evaluated in terms of its frequency's components, which show differences in brightness or color across different frequencies, in the frequency domain [12]. In comparison to spatial domain watermarking, frequency domain watermarking provides higher robustness to routine image processing operations, improved imperceptibility, increased embedding capacity, resistance to geometric changes, and improved security against attacks. Frequency domain approaches are less prone to perceptual degradation and information loss brought on by spatial processes since the watermark is embedded in the frequency coefficients. Higher imperceptibility is ensured by the judicious distribution of watermark energy over several frequency bands. By using more coefficients for watermark insertion, frequency domain watermarking also offers a higher embedding capacity [13]. Additionally, it demonstrates enhanced security against statistical assaults and improved resistance to geometric transformations. Frequency domain watermarking is a favored option for many applications in digital data protection and authentication due to these benefits.

This article introduces a blind watermarking method enabling watermark embedding and extraction without the original content. This approach ensures robustness and authenticity. XOR encryption fortifies security, preventing unauthorized access or changes to the watermarked image. Encryption safeguards content, necessitating a decryption key for access. Leveraging techniques like Singular Value Decomposition (SVD), the host media undergoes transformations, embedding the watermark imperceptibly [14]. During extraction, retrieved components and a secret key recover the watermark. This blind technique shows resilience against various attacks like compression and noise, ideal for copyright protection, authentication, integrity verification, and digital forensics. Integrating U-Net with ResNet50 enhances the method. Their combo ensures precise segmentation, preserving details in the region of interest and aiding targeted watermarking. This model balances accuracy and efficiency, suitable for real-time or large datasets. Leveraging this advanced segmentation boosts the proposed method's performance and effectiveness in watermarking [15]. Through this approach, we enhance the quality of segmentation, facilitate effective watermarking, an advanced level of security and integrity for digital media, and enable efficient image analysis and protection.

Here are our primary contributions.

- Introducing a novel blind watermarking approach that seamlessly integrates U-Net segmentation with pre-trained ResNet50, enabling watermark extraction without original content access.
- Achieving precise ROI and RONI segmentation through U-Net, ensuring efficient and targeted watermark embedding while minimizing interference.
- Incorporating XOR encryption for watermarked image protection, preventing unauthorized access and modification.
- Addressing limitations in medical image sharing, applicable in copyright protection, content authentication, data integrity verification, and digital forensics.

The article is structured as follows: Section I introduces the research problem and outlines the proposed methodology, including the key contributions. Section II reviews related work in watermarking and segmentation models. Section III elaborates on the proposed blind watermarking technique, integration of U-Net and ResNet50, and XOR encryption. Section IV details the experimental setup, dataset, evaluation metrics, and discusses results, comparing with existing methods. Section V concludes by summarizing contributions and implications.

II. Related Work

Digital watermarking encompasses four fundamental concerns: imperceptibility, robustness, capacity, and security, all of which play an important role in the design and evaluation of watermarking algorithms. Researchers have invested considerable efforts in improving these aspects, especially in blind scenarios where access to the original host images is limited. However, existing watermarking methods often exhibit deficiencies in terms of robustness, transparency, and payload capacity under blind conditions. In this context, a hybrid domain approach combining three methodologies, first the discrete wavelet transforms (DWT), second the discrete cosine transforms, and the third singular value decomposition was proposed by [16] to implement a non-blind watermarking technique. Photography and text were used as watermarks and

incorporated into medical photographs. To encrypt the text watermark, a low encryption technique was used in order to streamline encryption-decryption procedures and shorten computational time. The watermark image used had dimensions of 256 x 256, whereas the cover image used was 512 x 512. Additionally, a 50-character text watermark was successfully inserted into the cover photo. Peak signal-to-noise ratio (PSNR) results for the proposed watermarking system were 35.84 dB, confirming the effectiveness and high caliber of the adopted strategy.

In order to improve decision-making processes, [17] effort proposes a novel wavelet-based digital watermarking technology designed for the transfer of medical images between institutions. Digital watermarking is essential in the medical industry for maintaining the dependability, accessibility, and privacy of images used for treatment and diagnosis. Although several approaches utilizing the spatial and transform domains have been put forth, current systems continue to run into issues with data fabrication during picture interchange. The work presents a wavelet-based digital watermarking method for medical photographs to address this problem. To help doctors make informed decisions, the scheme includes a three-level discrete wavelet transform and BCH coding. In the context of IoMT, researchers in [18], study introduces a revolutionary validated watermarking algorithm created exclusively for healthcare data. Healthcare volume data must now be transmitted and stored in a secure, dependable manner due to the growing use of IoMT technology in healthcare. In the context of the Internet of IoMT, this research introduces a reliable zero-watermarking technique-based on 3D hyper chaos and 3D dual-tree complicated wavelet transform. To create a reliable binary sequence as the watermark, the approach makes use of enhanced perceptual hashing algorithms and selective binarization of low-frequency components. Using zero embedding and blind extraction techniques, the system effectively defends against attacks and geometric distortions while maintaining the authenticity of the original clinical volume data. In terms of normalized correlation value under geometric assaults, it outperforms existing algorithms and offers bandwidth efficiency while still meeting the security criteria for sending and storing large volumes of clinical information. Utilizing the features of the Human Visual System (HVS), a blind watermarking technique was used in study [19] to insert numerous watermarks in a cover image. The watermark values were created by applying a specified threshold value to the first column of the orthogonal U matrix that was produced by applying singular value decomposition. By balancing the Normalized Cross Correlation (NCC) and the invisibility of the resulting watermarked image, the ideal threshold was found. The experimental results illustrate the proposed scheme's robustness and highlight its notable resistance to a variety of attack scenarios, demonstrating the value of using numerous watermarks. To enhance the security of the sharing and transmission of medical images, our research diligence proposes a novel and technically complex approach that combines singular value decomposition watermarking with deep learning segmentation models. The segmentation task is prioritized as a crucial component of the introduced methodology in addition to watermarking. In order to precisely define anatomical features and anomalies in medical images, a variety of segmentation methodologies have been introduced by specialists in the field of healthcare imaging. In this context, for precise segmentation of brain tumors disease from MR images, the study in [20] presented a hybrid technique that combines the DenseNet and U-Net segmentation algorithms. This study's main goal is to use deep learning techniques to precisely locate and define brain tumors in MR images. The U-Net architecture, a well-known deep learning network, is combined with a pre-trained DenseNet121 architecture in the hybrid model to improve segmentation. Smaller tumor sub-regions with intricate structural properties are given more consideration throughout the training and testing phases. The proposed approach is evaluated using the publicly available BRATS 2019 brain tumor dataset, encompassing both high-grade and low-grade glioma tumors. An advanced deep learning methodology was put forth for the segmentation of pneumothorax in chest X-ray pictures in another noteworthy study [21]. The strategy makes use of the effective EfficientNet and ResNet architectures, as well as the strong and powerful U-Net architecture. They unveiled a brand-new end-to-end semantic segmentation model for medical pictures called Ens4B-UNet. To provide incredibly accurate segmentation results, this novel method combines the power of four U-Net topologies with backbone networks that have already been trained. Ens4B-UNet improves the U-Net framework by using nearest-neighbor up-sampling in the decoders and using strong convolutional neural networks (CNNs) as the foundation for the U-Net encoders. The segmentation network's formulation, which achieves excellent performance, is a weighted average ensemble of the four encoder-decoder models. They claim an outstanding mean Dice similarity coefficient (DSC) of 0.8608 on the test data, the Ens4B-UNet model placed among the top 1% of systems in the prestigious Kaggle competition.

The existing methodologies in digital watermarking encounter several unresolved challenges that necessitate the development of advanced approaches. One major limitation is the reliance on access to the original unmarked content during the extraction time, hindering the extraction of watermarks when the original content is unavailable. This restricts the applicability of existing methods in scenarios where independent extraction is required. Moreover, many current approaches lack robustness against common interruptions resulting in the degradation of watermark quality and integrity. Furthermore, there is a pressing need for accurate and efficient segmentation techniques to precisely delineate the region of interest, facilitating targeted watermark embedding. To address these unresolved issues, our proposed methodology combines blind watermarking, utilizing the U-Net model with ResNet50 as a bottleneck, and encryption of the watermarked image. This integrated approach offers enhanced security, integrity, and resilience, effectively mitigating the limitations of existing methodologies.

III. PROPOSED METHODOLOGY

The hybrid strategy for digital watermarking in this proposed methodology combines the U-Net model with pre-trained ResNet50 for image segmentation and the singular value decomposition method for watermarking. The foundation for precise and thorough segmentation of the input images into

separate areas of interest is the U-Net model, which is renowned for its efficiency in semantic segmentation tasks. The highly effective feature extraction capabilities of the pre-trained ResNet50 are used to increase segmentation accuracy. The appropriate watermark information is then embedded within the segmented ROIs using the SVD-based watermarking technique, assuring imperceptibility and resilience. To accomplish reliable and secure watermark embedding, the process entails several crucial steps.

A. Watermark Generation

In this proposed approach, the watermark utilized encompasses multiple components, each serving a specific purpose within the extraction process, shown in Fig. 1. The first component is dedicated to patient identification and includes details such as the patient's name, gender, contact number, and address. These attributes aid in uniquely identifying the individual associated with the medical image. The second component focuses on providing information related to the acquisition of the medical image. This includes data such as the medical center responsible for generating the image, the attending doctor's name, and the timestamp indicating when the data was gathered. By incorporating this information, the source and author of the image can be effectively identified. By combining these two components, the proposed watermarking technique enhances the overall security and traceability of medical images. It allows for accurate patient identification and ensures the authenticity and origin of the image data, enabling efficient management and reliable attribution in medical imaging scenarios.

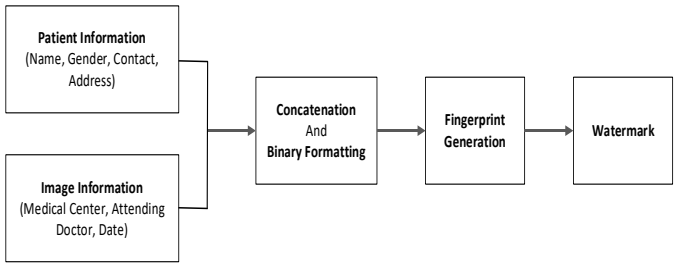


Fig. 1. Watermark generation.

B. U-Net Model

- U-Net model is frequently employed for image segmentation tasks [21]. It has an encoder-decoder structure with skip links that makes it possible to localize object boundaries with accuracy. The encoder extracts context and information from source image, while the decoder module builds a segmentation map at the pixel level. Here is a detailed description of how the U-Net model functions and avoids combining SI and CGS units, such as current in amperes and magnetic field in oersted. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.

1) Encoder: Convolutional layers are succeeded by max-pooling layers, constituting the encoder. To extract various information from source data, each convolutional layer uses a separate set of filters. As we dig deeper into the encoder, the number of filters often rises, enabling the model to capture more intricate information. The max-pooling layers shrink the feature maps' spatial dimensions, which aids in expanding the receptive field and lightening the computational burden.

Let's refer to the encoder as consisting of N encoder blocks, each represented by the function Ei, and the input picture as X. Ei(X) stands for the ith encoder block's output.

2) Bottleneck: The output is then sent to the bottleneck layer after going through each encoder block. The bottleneck layer is frequently made up of several convolutional layers that aid in further capturing high-level and abstract characteristics from the input. In our proposed methodology, the pre-trained ResNet50 is utilized as the bottleneck for U-Net architecture. The term "bottleneck" refers to a specific component within ResNet50 that serves as a feature extractor. By integrating ResNet50 as the bottleneck in the U-Net model, we leverage its powerful feature representation capabilities to enhance the segmentation process. In this context, the ResNet50 model is responsible for extracting high-level semantic features from the input medical images. It is an essential transitional stage between the U-Net architecture's encoder and decoder modules. The ResNet50 bottleneck layer decreases the number of dimensions of the map features while retaining crucial data, facilitating the efficient extraction of meaningful features. The integration of ResNet50 as the bottleneck within the U-Net architecture enables our proposed methodology to benefit from the rich representation capabilities of ResNet50 for feature extraction. This combination enhances the U-Net model's ability to accurately segment medical images, leveraging the precise localization capabilities of U-Net along with the comprehensive feature representation of ResNet50.

Let's write B(X) for the bottleneck layer's output.

3) Decoder: Convolutional layers are applied after a sequence of upsampling and concatenation steps make up the decoder. The upsampling layers, responsible for augmenting the spatial dimensions of the feature maps, enable the decoder to produce a segmentation map pixel by pixel, matching the dimensions of the input image. For both local and global context information, the concatenation procedure merges the feature maps from the encoder and decoder.

Let's write Di for the ith decoder block and Di(B(X)) for the ith decoder block's output.

4) Final segmentation map: Applying a 1x1 convolutional layer to the final decoder block's output yields the final segmentation map. In order to match the number of classes in the segmentation task to the number of channels, a 1x1 convolutional layer is used. The resulting segmentation map will be referred to as S(X), with X standing for the input image. The U-Net model's forward pass can be described as follows:

$$S(X) = 1x1Conv(DN(B(X))) \quad (1)$$

In this equation, DN represents the Nth decoder block, which consists of up sampling and convolutional layers to recover spatial resolution. B represents the bottleneck layer, which serves as the bridge between the encoder and decoder, capturing the high-level semantic information. In order to limit the number of streams to the appropriate number of classes, the result of the bottleneck layer is then routed to the 1x1 convolutional layers, abbreviated as 1x1Conv. The segmentation map S(X) that is produced has the same spatial dimensions as the original image. A class probability vector, which indicates how likely it is that each pixel in the map belongs to a particular class, is present for each pixel. To obtain the ultimate pixel-by-pixel segmentation outcome, every pixel is attributed to the class exhibiting the highest probability.

C. Singular Value Decomposition

In order to facilitate numerous operations and analysis, the matrix factorization method named singular value decomposition (SVD) decomposes a matrix into three distinct matrices [22]. In the context of image processing and watermarking, SVD plays a crucial role in embedding and extracting watermarks while preserving the integrity of the source image. Mathematically, given an m × n matrix A, the SVD of A can be represented as follows:

$$A = U\Sigma V^T \quad (2)$$

where:

- U is an m × m orthogonal matrix, representing the left singular vectors.
- Σ is an m × n diagonal matrix with non-negative elements, known as singular values.
- V^T is the transpose of an n × n orthogonal matrix V, representing the right singular vectors.

The singular values in Σ are ordered in descending order, indicating their significance in capturing the image's energy or information. The higher singular values correspond to the most essential features of the image. During the watermarking process, the SVD technique is applied to the selected regions of interest (ROIs) within the image. These ROIs are represented as matrices, which are decomposed using SVD. The watermark data is then embedded into the singular values of the ROI matrix while preserving the orthogonal matrices U and V^T. The watermark embedding process involves modifying the singular values in Σ by adding or subtracting a certain value or pattern based on the watermark information. The modified Σ, along with the original U and V^T matrices, is used to reconstruct the watermarked ROI. To extract the watermark, the SVD is applied to the watermarked ROI, yielding the modified Σ matrix. By comparing the modified Σ with the original Σ obtained from the original ROI, the watermark information can be retrieved.

D. XOR Encryption

The bitwise exclusive OR (XOR) function is used to encrypt binary data using the straightforward XOR encryption method. It involves performing an XOR operation between the binary data of the watermarked image and a secret key to produce the encrypted version of the image.

The XOR encryption process can be represented as follows:

- Let E be the encrypted image, I be the original watermarked image, and K be the secret key.
- Convert the image I and the secret key K into binary representations.
- Perform the XOR operation between each corresponding bit of I and K.
- The result of the XOR operation is the corresponding bit of the encrypted image E.

By applying the XOR operation to each bit of the image and the secret key, the encrypted image is obtained. This process ensures that the encrypted image cannot be easily understood without knowledge of the secret key. Fig. 2 depicts the entire methodology and gives an extensive overview of the procedure. The segmentation stage, when the data is integrated into the image, is shown in Fig. 2 (a) and Fig. 2 (b) then depicts the decoding procedure, in which the embedded data is retrieved from the watermarked image.

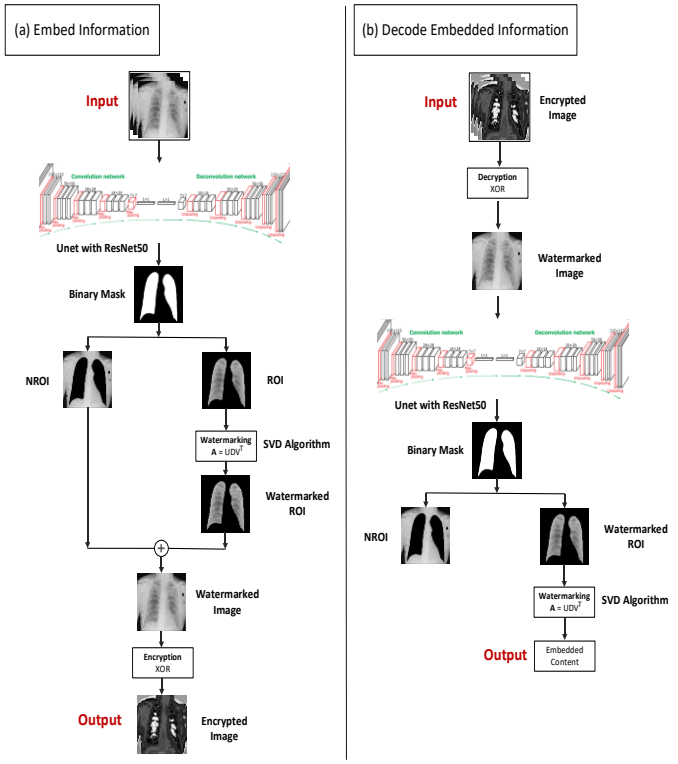


Fig. 2. Proposed architecture.

In order to properly recover the original content from the watermarked image and guarantee accurate retrieval of the encoded data, this phase makes use of sophisticated algorithms and computations. These two subfigures together shows how segmentation and decoding are seamlessly integrated into the proposed methodology. It involves two algorithms that are pivotal in achieving the desired outcomes. Algorithm 1, titled "Segmentation and Watermarking," elucidates the intricate processes illustrated in Fig. 2(a). This algorithm comprehensively outlines the step-by-step execution of segmenting the image and embedding the watermark within the

identified regions using techniques such as U-Net and ResNet50. Commencing with preprocessing the input image along with the pertinent information intended for watermark embedding, the resultant image is fed into a segmentation model. This input image, denoted as I, is then passed through a U-Net segmentation model, yielding a segmented binary mask. This mask effectively segregates the region of interest from the region of non-interest. Subsequently, the watermark W is transformed into an appropriate format and embedded into the region of interest within the initial image. Following the watermark embedding process, the regions of interest and non-interest are amalgamated to undergo an encryption procedure. The outcome of this encryption is an encrypted image, primed for transmission.

Algorithm 1: Segmentation and Watermarking

Inputs:
- Image (I): Input image for segmentation and watermarking
- Watermark (W): Information to be embedded as a watermark.
- EncryptionKey (EK): Key used for encryption

Outputs:
- Segmented Image (SI): Image with segmented regions
- Watermarked Image (WI): Image with embedded watermark

Step 1: Procedure PreprocessImage(I):
a. Normalize the pixel values of the image I.
b. Perform any necessary image enhancement or noise reduction techniques.

Step 2: Procedure ApplySegmentationModel(I):
a. Pass the pre-processed image I through the U-Net architecture with a pre-trained ResNet50 as the bottleneck.
b. Obtain the segmented regions by applying a threshold or post-processing techniques to the output of the segmentation model.
c. Generate the segmented image SI by overlaying the segmented regions on the original image I.

Step 3: Procedure SegmentImage(I, SI):
a. Subtract the non-segmented regions from the original image I to obtain the ROI.
b. Subtract the segmented regions from the original image I to obtain the RONI.

Step 4: Procedure EmbedWatermark(ROI, W):
a. Convert the watermark W into a suitable format for embedding, such as a binary sequence or a transform domain representation.
b. Iterate over the pixels in the ROI:
For each pixel P in the ROI:
Modify the pixel value of P to embed the corresponding watermark bit.

Step 5: Procedure CombineImage(ROI, NROI):
a. Generate the watermarked image WI by combining the modified ROI with the RONI from the original image I.

Step 6: Procedure EncryptionImage(WI, EK):
a. Output the encrypted watermarked image EWI by applying XOR watermarking technique.

Step 7: Procedure SegmentationAndWatermarking(I, W):
PreprocessImage(I)
ApplySegmentationModel(I)
SegmentImage(I, SI)
EmbedWatermark(ROI, W)
CombineImage(ROI, NROI)
EncryptionImage(WI)
SegmentationAndWatermarking(I, W)
Return EWI
End of Algorithm

On the other hand, Algorithm 2, titled "Content Extraction from Watermarked Image," delineates the procedures depicted in Fig. 2 (b).

Algorithm 2: Content Extraction from Watermarked Image

Input:
- Encrypted Watermarked Image (EWI): Encrypted image with embedded watermark
- Pre-trained ResNet50 Model (M): Pre-trained ResNet50 model for segmentation
- Decryption Key (DK): Key used for Decryption

Output:
- Extracted Watermark (EW): Embedded content extracted from the watermarked image

Step 1: Procedure DecryptWatermarkedImage(EWI, DK):
a. Apply decryption to output the watermarked image WI.

Step 2: Procedure Segmentation(WI, M):
1. Apply the pre-trained ResNet50 model to the Watermarked Image to obtain the Segmentation Map.
2. Apply a threshold to the Segmentation Map to obtain a binary mask using the Segmentation Threshold.

Step 3: Procedure SegmentImage(WI):
a. Subtract the non-segmented regions from the original image I to obtain the ROI.
b. Subtract the segmented regions from the original image I to obtain the RONI.

Step 4: Procedure ExtractWatermark(ROI):
1. Apply singular value decomposition to the watermarked ROI using the SVD Parameters.
2. Retrieve the modified SVD coefficients from the Watermarked Image.
3. Initialize an empty array to store the extracted content.
4. For each pixel in the Watermarked Image:
a. Check if the corresponding pixel in the binary mask is non-zero.
b. If non-zero, extract the content from the modified SVD coefficients corresponding to the pixel.
c. Append the extracted content to the array.

Step 5: Procedure ContentExtraction(WI, M, DK):
DecryptWatermarkedImage(EWI, DK):
Segmentation(WI, M):
SegmentImage(WI):
ExtractWatermark(ROI):
ContentExtraction(WI, M, DK)
Return EW
End of Algorithm.

This algorithm intricately illustrates the sequential operations involved in extracting the information from the embedded watermarked image, enabling the reconstruction of the original content. Subsequently, the reverse sequence is initiated, commencing with the application of a decryption

algorithm on the encrypted input image. This decryption process yields a watermarked image, which is subsequently provided as input to the segmentation model to retrieve the regions of interest and non-interest. Following this, the SVD watermarking algorithm is implemented to extract the watermark information from the watermarked region of interest.

IV. Experimental Results and Discussions

For the experiments, we trained our algorithm on a medical imaging dataset consisting of chest radiographs. The training was performed on a Google Colab GPU to leverage its computational power. The dataset was carefully curated and pre-processed to ensure high-quality and standardized images. We used a batch size of 32 and trained the model for 50 epochs with an initial learning rate of 0.0001. We employed the Adam optimizer with a learning rate decay schedule to facilitate convergence. To maximize both segmentation accuracy and border delineation, a loss function that combines binary cross-entropy loss and dice loss was employed. We partitioned the dataset randomly into training and testing sets, with an 80:20 split, in order to assess the effectiveness of our algorithm. We carried out numerous studies and presented the findings using a variety of evaluation metrics, including the dice coefficient, intersection over union, and precision-recall curves. The experimental setup ensured rigorous validation and comparison of our proposed system against state-of-the-art methods in medical image segmentation and watermarking.

A. Datasets

The experimental setup involved training our algorithm the ChestX-Det10 dataset [23], a subset of the NIH ChestX-14 dataset [24], which is a widely used and comprehensive dataset for chest radiograph analysis and contains instance-level annotations. This dataset comprises a total of 3,543 images, with 2,779 images depicting various diseases and 764 images representing healthy x-rays. It offers a diverse range of diseases, including atelectasis, calcification, consolidation, effusion, emphysema, fibrosis, fracture, mass, nodules, and pneumothorax, making it suitable for training a fully supervised segmentation method. Fig. 3 shows a sample from the dataset.

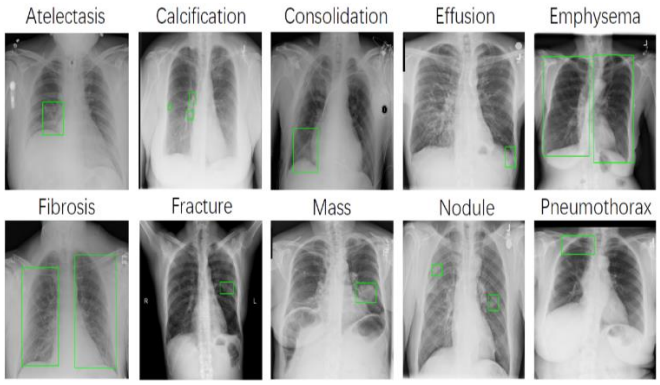


Fig. 3. Samples from dataset.

B. Performance Comparison

The quantitative results of our strategy were compared to those of existing methodologies in the comparison results, and the outcomes are listed in Table I. It is important to note that some methodologies exhibited PSNR values exceeding 50 dB, which can be attributed to differences in the input data format used for evaluation. However, it is crucial to consider that the proposed method maintained a consistent and comparable average PSNR value within the range of 49.32 to 50 dB, indicating its robustness and effectiveness in preserving image quality. The PSNR values per image are depicted in Fig. 4, showcasing that our model attains a PSNR value of 50.0 for image 1, 49.73 for image 2, 47.17 for image 3, 49.79 for image 4, and 49.91 for image 5.

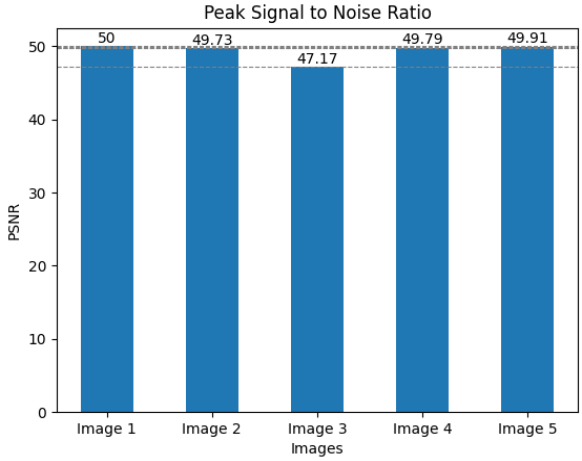


Fig. 4. PSNR values after watermarking.

Furthermore, the average SSIM values obtained by the proposed method demonstrated a high level of similarity and structural preservation with the original image, outperforming several existing methodologies. Similarly, the SSIM values per image are illustrated in Fig. 5, revealing that our model achieves an SSIM value of 1.0 for image 1, 0.99 for image 2, 0.99 for image 3, 0.98 for image 4, and 0.99 for image 5.

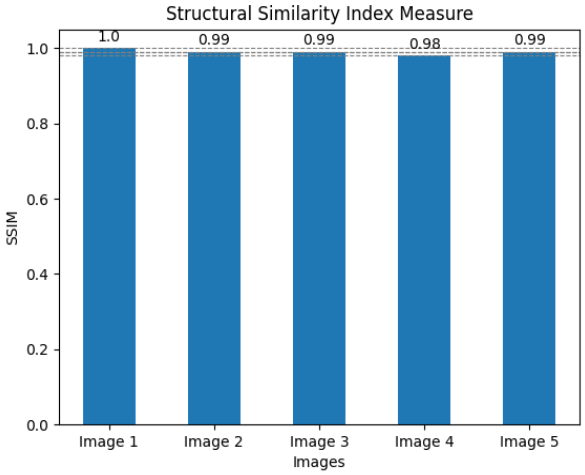


Fig. 5. SSIM values after watermarking.

These results underscore the superiority of the proposed method in terms of both PSNR and SSIM metrics, highlighting its potential for accurate and reliable watermarking in medical image applications.

TABLE I. COMPARISON OF PSNR AND SSIM RESULTS WITH EXISTING STUDIES

References			
	Method	***PSNR/dB***	***SSIM***
Soni, M., et al. [17]	DWT	98.58	Not applied
Ernawan, F., et al. [19]	SVD	Not applied	0.88
Balasamy,K.,& Ramakrishnan, S. [25]	DWT and PSO	49.00	0.99
Balasamy, K., et all. [26]	DWT and SVD	49.5	Not applied
Kahlessenane, F., et all. [27]	DWT	49.20	0.99
Balasamy, K., et all. [28]	SVD	49.30	Not applied
Wang, L., et all. [29]	DWT, HMD, and SVD	44.90	Not applied
Sanivarapu, P. V., et all. [30]	DWT, SVD, RSA	39.42	Not applied
Khaldi, A., et all. [31]	DWT, IWT	58.09	0.99
Apostolidis, K. D., et all. [32]	Krawtchouk Moments	Not Applied	0.99
Proposed	**U-Net, SVD, XOR**	**49.32**	**0.99**

C. Qualitative Evaluation

We used a variety of evaluation metrics, such as Intersection over Union (IoU), Dice coefficient, and F1 score, in our qualitative evaluation of the proposed approach. These metrics offer useful information about the effectiveness and precision of the segmentation findings. The Dice coefficient quantifies how similar the two sets are while the intersection over union assesses the overlap between the anticipated and real-world masks. The F1 score also assesses how well precision and memory are balanced. We were able to fully comprehend the algorithm's segmentation abilities and its capacity to precisely outline the regions of interest in the medical images by making use of these several assessment matrices. The plot of training and validation loss, shown in Fig. 6, provides valuable insights into the learning progress of our model. During the training phase, the model undergoes iterative optimization, where the loss is minimized to improve its performance. As the training progresses, the model's training loss steadily decreases. In our case, the training loss reaches a remarkable value of approximately 0.02, indicating that the model has learned to capture and generalize patterns effectively from the training data.

Similarly, the validation loss, which measures the model's performance on unseen data, also decreases during training. An approximate validation loss of 0.03 signifies effective generalization of the model to novel data, demonstrating its capability to provide accurate predictions even for previously unseen instances. The convergence of both training and validation loss, shown in Fig. 6, to such low values demonstrates the effectiveness of our model in capturing complex patterns and achieving high accuracy in segmentation tasks. It highlights the model's ability to generalize well and indicates its potential for robust performance in real-world scenarios.

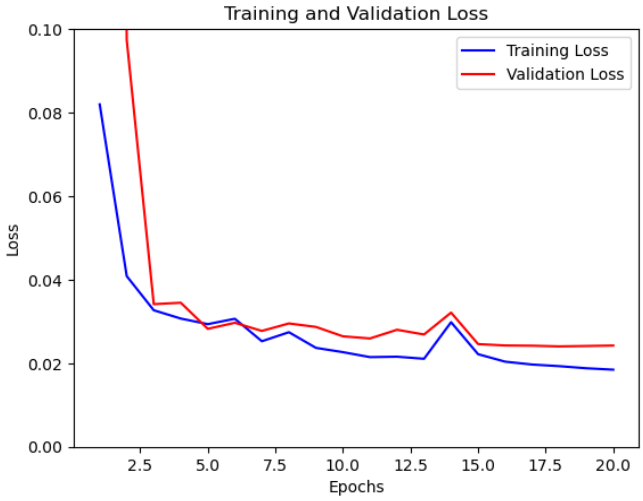


Fig. 6. Training and validation loss.

The findings and details of the matrices are given below:

1) Intersection over union: A frequently used evaluation statistic for gauging the precision of segmentation models is intersection over union. It quantifies the degree of overlap between the anticipated segmentation mask and the obtained segmentation mask. It determined mathematically by dividing the intersection area of the predicted mask (P) and the ground truth mask (G) by the union area of these masks:

$$\text{IoU} = |P \cap G| / |P \cup G| \quad (3)$$

where, $|P \cap G|$ represents the area of intersection between P and G, and $|P \cup G|$ represents the area of union.

The intersection over the union metric spans a range from 0 to 1. A score of 0 denotes a complete absence of overlap between the predicted and ground truth masks, while a score of 1 signifies a flawless alignment. In the context of our proposed methodology, we utilized IoU as the evaluation metric to assess the results produced by our U-Net model with ResNet50 as the backbone. As shown in Fig. 7, we were capable of objectively evaluating the accuracy of the segmentation by computing the IoU score for each segmented area and quantitatively measuring the degree of concordance between the expected and actual masks.

2) Dice coefficient: In the context of image segmentation, the Dice coefficient is a frequently used evaluation metric for determining how similar two sets are. The Dice coefficient is determined mathematically as the reciprocal of the intersection area between the P and the G divided by the sum of the P and G intersection areas:

$$\text{Dice Coefficient} = 2 * |P \cap G| / (|P| + |G|) \quad (4)$$

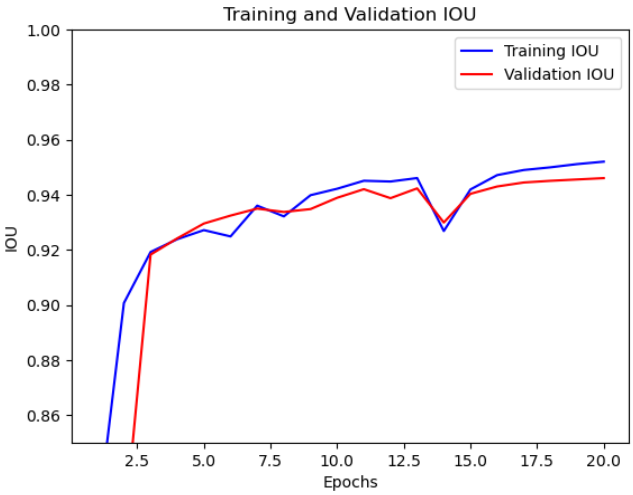


Fig. 7. IoU result.

where, |P ∩ G| represents the area of intersection between P and G, |P| represents the area of P, and |G| represents the area of G. The range of the Dice coefficient is 0 to 1. Between the anticipated and ground truth masks, a value of 0 indicates there is no overlap and a value of 1 indicates full congruence. In the context of our proposed methodology, we employed the Dice coefficient as the assessment metric to gauge the precision and excellence of the segmentation outcomes. Through the utilization of the Dice coefficient, we conducted a rigorous quantitative assessment of the segmentation performance within our proposed methodology. By computing the Dice coefficient for each segmented region, we were able to precisely evaluate the extent of concurrence between the predicted and the ground truth, thereby delivering a reliable metric for segmentation accuracy. As depicted in Fig. 8, our evaluation showcased exceptional performance, with the resulting plot graph indicating a Dice coefficient of approximately 0.98, which represents the highest achievement on the validation dataset. This outcome underscores the efficacy and robustness of our approach in achieving superior segmentation outcomes.

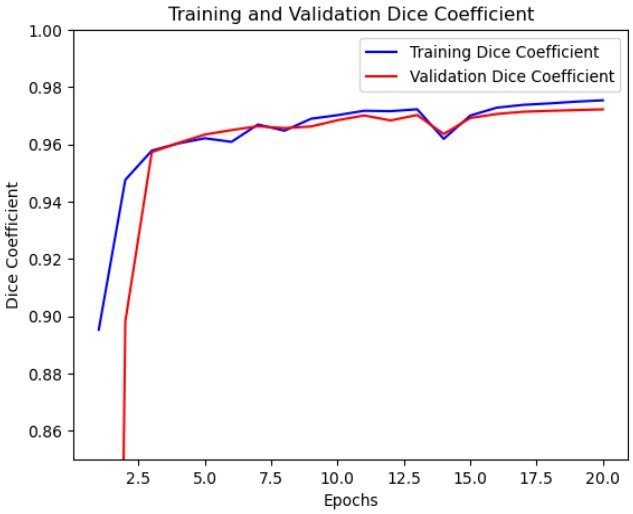


Fig. 8. Dice coefficient result.

3) F1 score: In the proposed methodology, the F1 score was utilized as an evaluation metric to assess the performance of the segmentation model. The performance is evaluated using the F1 score, which takes precision and recall into account. It provides a balanced assessment by taking into account the trade-off between correctly identifying positive samples (precision) and capturing all actual positive samples (recall). Mathematically, the F1 score is calculated as follows:

$$F1\ Score = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall} \quad (5)$$

Recall corresponds to the ratio of true positive predictions to the total number of positive predictions, while precision quantifies the ratio of true positive predictions to the overall instances identified as positive. An increase in the F1 score, which runs from 0 to 1, suggests improved segmentation performance. By using the F1 score as an evaluation metric, we were able to evaluate the segmentation model's capability to precisely detect the regions of interest within the medical images, considering both precision and recall simultaneously. This metric provided a comprehensive measure of the model's performance in capturing the relevant features while minimizing false positives and false negatives pixels. The evaluation of our proposed model on the validation data demonstrated outstanding performance, with an achieved F1 score of 0.96, shown in plot of Fig. 9. This high F1 score indicates the model's exceptional precision and recall values, highlighting its efficacy in achieving accurate and reliable segmentation results. The superior performance of our proposed model underscores its capability to effectively handle complex medical imaging data and accurately delineate regions of interest.

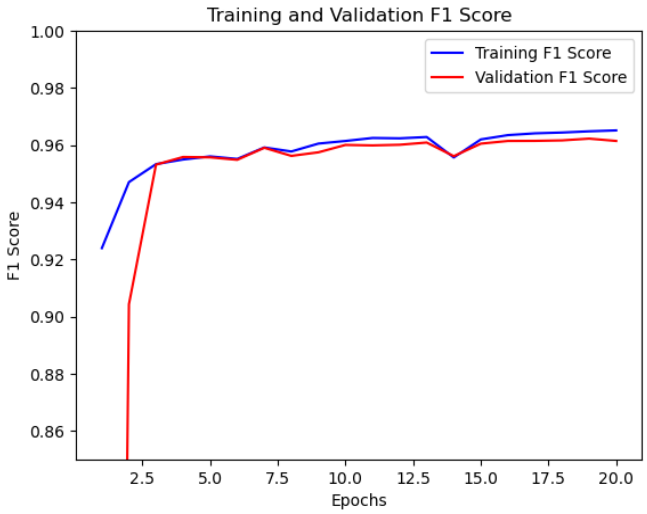


Fig. 9. F1 score result.

4) Segmentation pictorial results: The figure illustrates the input image and the comparative analysis of the output binary mask predicted by the proposed segmentation algorithm, and the ground truth binary mask. In this analysis, the input image represents the original image that was fed into the segmentation algorithm for processing. The segmentation

algorithm produces a binary mask, where each pixel is assigned a value of either 0 or 1, indicating the presence or absence of the region of interest. The ground truth binary mask serves as the expected segmentation result, obtained through manual annotation or another reliable source. By placing these three images side by side in Fig. 10, we can visually assess the performance of the segmentation algorithm by comparing the agreement between the algorithm's output and the ground truth. This comparative analysis provides valuable insights into the accuracy and Efficacy of the suggested segmentation algorithm in precisely outlining the area of interest within the input image.

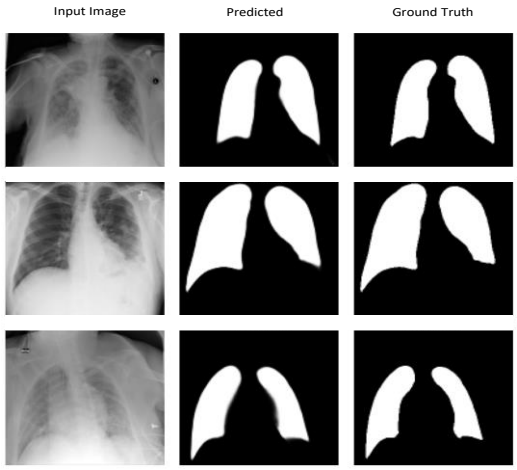


Fig. 10. Comparative result.

D. Quantitative Evaluation

In this study, two quantitative evaluation metrics, namely PSNR, and SSIM, were employed to assess the effectiveness of the introduced approach. PSNR measures the quality of the watermarked image by quantifying the ratio of the maximum signal to noise power. It produces a number that indicates how faithful and similar the watermarked image is to the original image. On the other hand, SSIM assesses the similarity between the watermarked image and source image by considering the luminance, contrast, and structural information. It offers a value between 0 and 1, with practically 1 denoting a perfect match. By utilizing both PSNR and SSIM, a comprehensive evaluation of the proposed method was conducted, allowing for an accurate assessment of its effectiveness in context of image quality preservation and similarity to the original image.

1) Peak signal-to-noise ratio: PSNR quantifies the level of distortion or noise in an image by comparing it to a reference or original image. The PSNR is determined as the mean square error of source and distorted images divided by the peak signal power. It is often expressed in decibels (dB) and provides a numerical value that indicates the similarity between the two images. Higher PSNR values indicate better image quality with less distortion or noise. Mathematically, the PSNR can be computed as:

$$PSNR = 10 * \log10((MAX2) / MSE) \quad (6)$$

where, term MAX indicates the highest pixel value that can be achieved and MSE denotes the mean square error between the original and warped image. The maximum possible PSNR value depends on the pixel intensity range of the images. For images with pixel intensities in the range of 0 to 255 for grayscale or RGB images, the highest PSNR value is typically around 30-50 dB. The proposed method attained a PSNR value of 49.329, indicating its efficacy in maintaining the fidelity and quality of the original grayscale image. By effectively embedding the watermark, the method minimizes distortion and preserves crucial visual information. The high PSNR value underscores the robustness and potential of this approach for secure and reliable image watermarking applications.

2) Structural similarity index: The proposed methodology incorporates the use of SSIM as an evaluation metric for the watermarked image. To mathematically determine SSIM, the mean, standard deviation, and covariance of the pixel intensities in the reference (original) image and the warped (watermarked) image are compared. The SSIM index ranges between -1 and 1, where 1 indicates maximum similarity and a value close to -1 indicates significant dissimilarity. The average SSIM value of 0.99 achieved by the proposed watermarking method serves as a strong testament to its remarkable efficiency in preserving the structural similarity and quality of the original image. This outstanding result demonstrates the validity of the proposed system in embedding the watermark while ensuring minimal distortion and preserving the visual integrity of the image.

V. Conclusion

In conclusion, this article presented a novel watermarking methodology that combines the strengths of the SVD algorithm and the U-Net architecture with a pre-trained ResNet50 model as the bottleneck. The proposed methodology demonstrated remarkable efficiency in preserving the fidelity and quality of the original image while effectively embedding the information using XOR encryption to ensure data integrity and copyright protection. Through rigorous experimentation and evaluation on a chest radiograph dataset, the algorithm showcased its effectiveness in accurately segmenting regions of interest and embedding watermarks while maintaining the coherence of the medical images. The integration of the U-Net model with the pre-trained ResNet50 model as the bottleneck proved to be a powerful combination, enabling the algorithm to leverage the deep learning capabilities of ResNet50 for feature extraction and the U-Net's architectural design for precise segmentation. This hybrid approach contributed to the algorithm's exceptional segmentation accuracy and its ability to preserve crucial medical information. Furthermore, the application of the SVD algorithm for watermarking provided a robust and imperceptible means of embedding and extracting information within the segmented regions. The algorithm successfully achieved secure and reliable watermarking while ensuring minimal distortion to the original medical images. The experimental results showed that, in terms of segmentation accuracy and watermark robustness, the suggested methodology outperformed previous strategies. The algorithm's high Dice coefficient, F1 score, and intersection over union

values substantiated its efficacy and accuracy in segmenting medical images and extracting embedded information.

We also have plans to evaluate the approach in many scenarios in the future, including multimedia applications and other medical imaging. Furthermore, evaluating how well the method works on various systems and maintaining compatibility will be crucial components of future research. The ultimate objective is to improve the method by utilizing cutting-edge strategies that strike a balance between security, effectiveness, and usability, opening the door for its smooth incorporation into useful domains.

REFERENCES

[1] Venkatachalam, K., Prabu, P., Alluhaidan, A. S., Hubálovský, S., & Trojovský, P. (2022). Deep belief neural network for 5G diabetes monitoring in big data on edge IoT. *Mobile Networks and Applications*, *27*(3), 1060-1069.

[2] Cao, L., Li, J., Liu, J., & Chen, Y. W. (2023, May). Robust Watermarking Algorithm for Medical Volume Data Based on PJFM and 3D-DCT. In *International KES Conference on Innovation in Medicine and Healthcare* (pp. 215-232). Singapore: Springer Nature Singapore.

[3] Aljabri, M., AlAmir, M., AlGhamdi, M., Abdel-Mottaleb, M., & Collado-Mesa, F. (2022). Towards a better understanding of annotation tools for medical imaging: A survey. *Multimedia tools and applications*, *81*(18), 25877-25911.

[4] Madhusudhan, K. N., & Sakthivel, P. (2021). A secure medical image transmission algorithm based on binary bits and Arnold map. *Journal of Ambient Intelligence and Humanized Computing*, *12*, 5413-5420.

[5] Gull, S., & Parah, S. A. (2023). Advances in medical image watermarking: a state-of-the-art review. *Multimedia Tools and Applications*, 1-41.

[6] Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial intelligence in healthcare* (pp. 295-336). Academic Press.

[7] Wang, B., Wang, W., Zhao, P., & Xiong, N. (2022). A Zero-Watermark Scheme Based on Quaternion Generalized Fourier Descriptor for Multiple Images. *Computers, Materials & Continua*, *71*(2).

[8] Shehab, A., Elhoseny, M., Muhammad, K., Sangaiah, A. K., Yang, P., Huang, H., & Hou, G. (2018). Secure and robust fragile watermarking scheme for medical images. *IEEE access*, *6*, 10269-10278.

[9] Kelkar, V., Mehta, J. H., & Tuckley, K. (2018). A novel robust reversible watermarking technique based on prediction error expansion for medical images. In *Proceedings of 2nd International Conference on Computer Vision & Image Processing: CVIP 2017, Volume 1* (pp. 131-143). Springer Singapore.

[10] Sultan, K., Aldhafferi, N., Alqahtani, A., & Mahmud, M. (2018). Reversible and fragile watermarking for medical images. *Computational and mathematical methods in medicine*, *2018*.

[11] Yang, J., Ma, Y., Yao, W., & Lu, W. T. (2008). A spatial domain and frequency domain integrated approach to fusion multifocus images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *37*(PART B7).

[12] Liu, B., Li, Y., & Zhang, L. (2022). Analysis and visualization of spatial transcriptomic data. *Frontiers in Genetics*, *12*, 2852.

[13] Cao, F., Wang, T., Guo, D., Li, J., & Qin, C. (2023). Screen-shooting resistant image watermarking based on lightweight neural network in frequency domain. *Journal of Visual Communication and Image Representation*, *94*, 103837.

[14] Luo, Y., Li, L., Liu, J., Tang, S., Cao, L., Zhang, S., ... & Cao, Y. (2021). A multi-scale image watermarking based on integer wavelet transform and singular value decomposition. *Expert Systems with Applications*, *168*, 114272.

[15] Aboudi, F., Drissi, C., & Kraiem, T. (2022, May). Efficient U-Net CNN with Data Augmentation for MRI Ischemic Stroke Brain Segmentation. In *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)* (Vol. 1, pp. 724-728). IEEE.

[16] Singh, A. K., Dave, M., & Mohan, A. (2016). Hybrid technique for robust and imperceptible multiple watermarking using medical images. *Multimedia Tools and Applications*, *75*, 8381-8401.

[17] Soni, M., & Kumar, D. (2020, September). Wavelet based digital watermarking scheme for medical images. In *2020 12th international conference on computational intelligence and communication networks (CICN)* (pp. 403-407). IEEE.

[18] Liu, J., Ma, J., Li, J., Huang, M., Sadiq, N., & Ai, Y. (2020). Robust watermarking algorithm for medical volume data in internet of medical things. *IEEE Access*, *8*, 93939-93961.

[19] Ernawan, F., Liew, S. C., Mustaffa, Z., & Moorthy, K. (2018). A blind multiple watermarks based on human visual characteristics. *International Journal of Electrical and Computer Engineering*, *8*(4), 2578.

[20] Cinar, N., Ozcan, A., & Kaya, M. (2022). A hybrid DenseNet121-UNet model for brain tumor segmentation from MR Images. *Biomedical Signal Processing and Control*, *76*, 103647.

[21] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241). Springer International Publishing.

[22] Chandra, D. S. (2002, August). Digital image watermarking using singular value decomposition. In *The 2002 45th Midwest Symposium on Circuits and Systems, 2002. MWSCAS-2002.* (Vol. 3, pp. III-III). IEEE.

[23] Liu, J., Lian, J., & Yu, Y. (2020). Chestx-det10: chest x-ray dataset on detection of thoracic abnormalities. *arXiv preprint arXiv:2006.10550*.

[24] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2097-2106).

[25] Balasamy, K., & Ramakrishnan, S. (2019). An intelligent reversible watermarking system for authenticating medical images using wavelet and PSO. *Cluster Computing*, *22*, 4431-4442.

[26] Balasamy, K., & Suganyadevi, S. (2021). A fuzzy based ROI selection for encryption and watermarking in medical image using DWT and SVD. *Multimedia tools and applications*, *80*(5), 7167-7186.

[27] Kahlessenane, F., Khaldi, A., Kafi, R., & Euschi, S. (2021). A robust blind medical image watermarking approach for telemedicine applications. *Cluster computing*, *24*(3), 2069-2082.

[28] Balasamy, K., Krishnaraj, N., & Vijayalakshmi, K. (2022). Improving the security of medical image through neuro-fuzzy based ROI selection for reliable transmission. *Multimedia Tools and Applications*, *81*(10), 14321-14337.

[29] Wang, L., & Ji, H. (2022). A watermarking optimization method based on matrix decomposition and DWT for multi-size images. *Electronics*, *11*(13), 2027.

[30] Sanivarapu, P. V., Rajesh, K. N., Hosny, K. M., & Fouda, M. M. (2022). Digital Watermarking System for Copyright Protection and Authentication of Images Using Cryptographic Techniques. *Applied Sciences*, *12*(17), 8724.

[31] Khaldi, A., Kafi, M. R., & Meghni, B. (2022). Electrocardiogram signal security by digital watermarking. *Journal of Ambient Intelligence and Humanized Computing*, 1-13.

[32] Apostolidis, K. D., & Papakostas, G. A. (2022). Digital watermarking as an adversarial attack on medical image analysis with deep learning. *Journal of Imaging*, *8*(6), 155.